



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

**Non-parametric item response theory applications in the
assessment of dementia**

Sarah McGrory

Doctor of Philosophy

The University of Edinburgh

2015

Declaration

I declare that all work presented in this thesis is my own, except as specified. This work has not been submitted elsewhere for any degree or qualification. Versions of some chapters have been published in scientific journals, on which I am first author. The systematic review in Chapter 3 was published in the February 2014 issue of *BMC Psychiatry* (McGrory, Doherty, Austin, Starr & Shenkin, 2014). The study in Chapter 9 using data from the Scottish Dementia Research Interest Register was published online November 2013 and subsequently in in the July 2014 issue of *Age and Ageing* (McGrory, Shenkin, Austin & Starr, 2014). The study in Chapter 4 using data from Frontier Research Group, at Neuroscience Research Australia has been accepted for publication in *Dementia and Geriatric Cognitive Disorders EXTRA*. The study in Chapter 8 based on analysis of data from the Lothian Birth Cohort has been accepted for publication in *Psychological Assessment*. The development of the Mini-Addenbrooke's Cognitive Examination in Chapter 5 bears resemblance to a paper in *Dementia and Geriatric Cognitive Disorders* on which I am second author (Hsieh et al., 2015). My supervisors Elizabeth Austin, Susan Shenkin and John Starr gave permission to submit these studies for peer-review publication prior to submitting this thesis.

Sarah McGrory

Acknowledgment

Firstly I would like to acknowledge my supervisors, Elizabeth Austin, Susan Shenkin and John Starr, who were a constant source of encouragement, support and advice over the past three years. Thanks must also go to Mike Allerhand for his advice and expertise throughout my PhD but particularly during the initial few months when I was finding my statistical feet.

I am grateful to all of the study participants and to those who collected the data that formed the basis for most of my research. I gratefully acknowledge the financial support for the Alzheimer Scotland studentship and the access to a wide range of resources and expertise provided through the Alzheimer Scotland Dementia Research Centre and the Centre for Cognitive Ageing and Cognitive Epidemiology. I also benefitted from some time in the sun while working with the team at Neuroscience Research Australia in Sydney. It was an honour to collaborate with a number of other researchers in Sydney, all of whom were generous with their time, advice and help.

My friends and officemates made the experience of completing my PhD enjoyable and memorable. Special thanks must go to Chris-who asked for special thanks. My parents and sisters, Ruth and Emma have been a continual source of great support throughout my extended academic journey. My parents also deserve special mention for the provision of Herschel, along with considerable other countless contributions along the way. Above all I would like to thank Daniel for continuous encouragement, making me laugh when I needed a distraction and for a great deal of understanding, patience and putting up with my ways in more stressful times, particularly over the past few months.

Abstract

This thesis sought to address the application of non-parametric item response theory (NIRT) to cognitive and functional assessment in dementia. Performance on psychometric tests is key to diagnosis and monitoring of dementia. NIRT can be used to improve the psychometric properties of tests used in dementia assessment in multiple ways: confirming an underlying unidimensional structure, establishing formal item hierarchical patterns of decline, increasing insight by examining item parameters such as difficulty and discrimination, and creating shorter tests. From a NIRT approach item difficulty refers to the ease with which an item is endorsed. Discrimination is an index of how well an item can differentiate between patients of varying levels of severity.

Firstly I carried out a systematic review to identify applications of both parametric and non-parametric IRT to measures assessing global cognitive functioning in people with dementia. This review demonstrated that IRT can increase the interpretive power of cognitive assessment scales and confirmed the limited number of IRT analyses of cognitive scales in dementia populations. This thesis extended this approach by applying Mokken scaling analysis to commonly used measures of current cognitive ability (Addenbrooke's Cognitive Examination-Revised (ACE-R)) and of premorbid cognitive ability (National Adult Reading Test (NART)). Differential item functioning (DIF) by diagnosis identified slight variations in the patterns of hierarchical decline in the ACE-R. These disease-specific sequences of decline could serve as an adjunct to diagnosis, for example where learning a name and address is a more difficult task than being orientated in time, late onset Alzheimer's disease is a more probable diagnosis than mixed Alzheimer's and vascular dementia. These analyses also allowed key items to be identified which can be used to create briefer scales (mini-ACE and Mini-NART) which have good psychometric properties. These scales are clinically relevant,

comprising highly *discriminatory*, invariantly ordered items. They also allow sensitive measurement and adaptive testing and can reduce test administration time and patient stress.

Impairment of functional abilities represents a crucial component of dementia diagnosis with performance on these functional tasks predictive of overall disease. A second aspect of this thesis, therefore, was the application of Mokken scaling analyses to measures of functional decline in dementia, specifically the Lawton Instrumental Activities of Daily Living (IADL) scale and Physical Self-Maintenance Scale (PSMS). While gender DIF was observed for several items, implying the likelihood of equal responses from men and women is not equal a generally consistent pattern of impairment in functional ability was observed across different types of dementia.

Glossary

1PLM	One-parameter logistic model
2PLM	Two-parameter logistic model
AISP	Automated item selection procedure
CTT	Classical test theory
DIF	Differential item functioning
DMM	Double monotonicity model
IRT	Item response theory
ISRF	Item step response functions
LID	Local item dependence
MS	Molenaar Sijtsma statistic
MHM	Monotone homogeneity model
MIIO	Manifest invariant item ordering
MSCPM	Manifest scale - cumulative probability mode
NIRT	Non-parametric item response theory
TCC	Test Characteristic Curve

Explanation of key terms

Latent trait (θ)	Latent construct to be measured using the scale
Item response function/ Item characteristic curve	Curve of the probability of item response as a function of the latent trait
Unidimensionality	Scale items measure the same latent trait/ a single latent trait accounts for the data structure
Local stochastic independence	Response to one item is not influenced by responses to other scale items/ item responses are independent
Monotonicity	As the level of latent trait increases the probability of endorsing the item increases or remains the same/ response is a non-decreasing function of the latent trait
Invariant item ordering	Items have the same ordering by <i>difficulty</i> regardless of the level of latent trait.
Scalability coefficients	Index of the homogeneity of items (H_i), item-pairs (H_{ij}) and the scale as a whole (H). Used in the assessment of unidimensionality

Abbreviations

ACE	Addenbrookes Cognitive Examination
ACE-III	Addenbrooke's Cognitive Examination III
ACE-R	Addenbrooke's Cognitive Examination-Revised
AD	Alzheimer's disease
ADL	Activities of Daily Living
AMPS	Assessment of Motor and Process Skills
ART	Adult Reading Test
BADL	Basic Activities of Daily Living
BIMCT	Blessed Information Memory Concentration Test
BRDRS	Blessed-Roth Dementia Rating Scale
bv-FTD	Behavioural variant frontotemporal dementia
CBD	Corticobasal degeneration
CDR	Clinical Dementia Rating Scale
CFA	Confirmatory factor analysis
CFI	Comparative Fit Index
DLB	Dementia with Lewy bodies
DPUK	Dementias Platform UK
DRS	Dementia Rating Scale
fNART	French language version of the NART
FTD-MND	Frontotemporal dementia with motor neurone disease.
IADL	Instrumental Activities of Daily Living
LBC1936	Lothian Birth Cohort 1936
LPA	Logopenic progressive aphasia
MCI	Mild cognitive impairment
MHT	Moray House Test
MI	Modification index

MIDAS	Myocardial Infarction Dimensional Assessment Scale
Mini-ACE	Mini-Addenbrooke's Cognitive Examination
MMSE	Mini Mental State Examination
MND	Motor neurone disease
MoCA	Montreal Cognitive Assessment
NAART	North American Adult Reading Test
NART	National Adult Reading Test
NeuRA	Neuroscience Research Australia
NICE	National Institute for Health and Care Excellence
OARS	Older Americans Resources and Services
PCA	Principal component analysis
PDD	Parkinson's disease dementia
PNFA	Progressive nonfluent aphasia
PPA	Progressive primary aphasia
PRF	Person response function
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSMS	Physical Self-Maintenance Scale
RMSEA	Root mean square error of approximation
SD	Semantic dementia
SDCRN	Scottish Dementia Clinical Research Network
SDRIR	Scottish Dementia Research Interest Register
SMS1947	Scottish Mental Survey 1947
VaD	Vascular dementia
WAIS-III	Wechsler Adult Intelligence Scale-Third Edition
WASI	Wechsler Adult Intelligence Scale
WTAR	Wechsler Test of Adult Reading

Table of Contents

Declaration.....	i
Acknowledgments.....	ii
Abstract.....	iii
Glossary	v
Explanation of key terms	vi
Abbreviations.....	vii
List of Tables	xvi
List of Figures.....	xx
Chapter 1: Introduction: Assessment in dementia -current and prior cognitive ability and functional ability.....	1
1.1 Cognitive assessment.....	4
1.2 Estimation of premorbid cognitive ability	7
1.3 Functional assessment.....	10
Chapter 2: Can item response theory methods be used to improve accurate assessment of dementia?	16
2.1 Problems with existing measurement methods.....	15
2.2 Classical test theory	17
2.3 Item response theory.....	18
2.3.1 The one-parameter logistic model	20
2.3.2 The two-parameter logistic model	21
2.3.3 Nonparametric item response theory	23
2.4. Mokken scaling-Origins in the Guttman scale.....	25
2.4.1 Mokken scaling analysis	26
2.4.2 Exploratory versus confirmatory Mokken scaling analysis.....	33
2.4.3 Mokken scaling analysis of dichotomous items versus polytomous items	34
2.4.4 Mokken scaling analysis versus Rasch analysis	35

2.5 IRT applications in assessment and quantification of dementia progression	37
2.5.1 Establishing hierarchical scales	39
2.5.2 Scale development and evaluation	42
2.6 Summary	43

Chapter 3: Item response theory analysis of cognitive tests in people with dementia: a systematic review 45

3.1 Introduction.....	45
3.2 Method	49
3.2.1 Search Strategy	49
3.2.2 Data Extraction	49
3.2.3 Inclusion/exclusion criteria.....	50
3.3 Results.....	54
3.4 Discussion	63
3.4.1 Item difficulty	63
3.4.2 Discrimination.....	64
3.4.3 Linearity and the assessment of change in severity	66
3.4.4 Information	69
3.4.5 Limitations	72
3.5 Conclusion	73

Chapter 4: Does the order of item difficulty of the Addenbrooke's Cognitive Examination add anything to sub-domain scores in the clinical assessment of dementia? 75

4.1 Introduction.....	75
4.2 Method	78
4.2.1 Participants.....	78
4.2.2 Measures	79
4.2.3 Factor analyses.....	79
4.2.4 Mokken scaling analysis	81
4.3 Results.....	85
4.3.1 PCA analysis.....	87
4.3.2 CFA analysis.....	88

4.3.3 Mokken Scaling Analysis	90
4.4 Discussion	99
4.5 Conclusion	106

Chapter 5: Mokken scaling analysis in the development of the Mini-Addenbrooke’s Cognitive Examination: a new assessment tool for dementia..... 107

5.1 Introduction.....	107
5.2 Method	110
5.2.1 Participants.....	110
5.2.2 Measures	111
5.2.3 Analyses.....	112
5.3 Results.....	113
5.3.1 Mokken scaling analysis	114
5.3.2 Sensitivity analysis.....	116
5.4 Item selection	124
5.4.1 Range of difficulty	125
5.4.2 Cognitive domains	129
5.4.3 Practical consideration	129
5.5 Validation of the Mini-ACE	132
5.6 Discussion.....	135
5.6.1 Limitations and future directions	136
5.6.2 Conclusion	140

Chapter 6: Hierarchical patterns of decline in the Addenbrooke’s Cognitive Examination-Revised..... 141

6.1 Introduction.....	141
6.2 Method	144
6.2.1 Participants.....	144
6.2.2 Statistical analysis	147
6.3 Results.....	149
6.3.1 Mokken scaling analyses of diagnostic groups.....	151
6.4 Discussion.....	168
6.4.1 Invariant item ordering across diagnostic subgroups.....	169

6.4.2 Assessment and interpretation of item parameters	173
6.4.3 Limitations and methodological considerations	176
6.5 Conclusion	178
Chapter 7: Development and validation of the Short ACE-R.....	181
7.1 Introduction.....	181
7.2 Method	182
7.2.1 Participants.....	183
7.2.2 Measures	186
7.3 Item selection.....	186
7.3.1 Short ACE-R selection 1.....	187
7.3.2 Short ACE-R selection 2.....	192
7.4 Validation of the Short ACE-R.....	194
7.5 Comparison of Short ACE-R and Mini-ACE	198
7.5.1 Assessment of mild impairment.....	203
7.5.2 Assessment of severe impairment.....	203
7.5.3 Importance of samples used.....	204
7.5.4 Practical implications.....	206
7.6 Discussion.....	209
7.7 Conclusion	214
Chapter 8: From ‘aisle’ to ‘labile’: a hierarchical NART scale revealed by Mokken scaling.....	215
8.1 Introduction.....	215
8.2 Method	219
8.2.1 Participants.....	219
8.2.2 Measures	221
8.2.3 Mokken scaling.....	222
8.2.4 Graphical analysis	222
8.2.5 Validation.....	222
8.3 Results.....	223
8.3.1 The Mini-NART	223
8.3.2 Item discrimination	230

8.4 Discussion.....	232
8.5 Conclusion	237

Chapter 9: Lawton Instrumental Activities of Daily Living scale in dementia: can item response theory make it more informative? 239

9.1 Introduction.....	239
9.2 Method	245
9.2.1 Participants.....	245
9.2.2 Measures	245
9.2.3 Statistical analysis.....	245
9.3 Results.....	246
9.4 Discussion.....	247

Chapter 10: Patterns of decline in Instrumental Activities of Daily Living across different types of dementia: Extension of Mokken scaling analysis on the Lawton IADL scale in the SDRIR 251

10.1 Introduction.....	251
10.1.1 Determinants of functional ability in dementia	252
10.1.2 Differential item functioning	255
10.1.3 The present study	256
10.2 Method.....	257
10.2.1 Participants.....	257
10.2.2 Measures	258
10.2.3 Mokken scaling analysis	259
10.3 Results.....	259
10.3.1 Differential item functioning assessment by diagnosis	262
10.3.2 Differential item functioning assessment by gender.....	268
10.4 Discussion	271
10.5 Conclusion	278

Chapter 11: Hierarchical patterns of functional loss in Activities in Daily Living and Instrumental Activities in Daily Living..... 281

11.1 Introduction.....	281
11.1.1 Combined hierarchical structure of BADL-IADL items	284
11.2 Method.....	286
11.2.1 Participants.....	286
11.2.2 Measures.....	287
11.2.3 Mokken scaling analysis	288
11.3 Results.....	288
11.3.1 Mokken scaling analyses of diagnostic groups.....	290
11.3.2 Mokken scaling analyses by gender.....	282
11.3.3 Combined BADL-IADL analysis	297
11.4 Discussion.....	298
11.5 Conclusion	302
Chapter 12: Discussion	303
12.1 Systematic Review.....	303
12.2 Hierarchical ordering by difficulty across all ACE analyses.....	304
12.3 Patterns of item difficulty between samples of patients with Alzheimer’s disease.....	307
12.4 Patterns of poor item discrimination.....	311
12.5 Using item discrimination to provide insight into the cognitive processes underlying item performance.....	319
12.6 Limitations of ACE analyses	319
12.6.1 Importance and implications of samples used	319
12.6.1.1 Different diagnoses of samples.....	319
12.6.1.2 Age difference.....	320
12.6.1.3 Testing conditions and location	320
12.6.2 Formation of clinical groups for analysis	322
12.6.3 Diagnostic circularity concerns.....	328
12.6.4 Polytomous item score equating	329
12.6.5 Testlets	331
12.6.6 Alternative method: Graded Response Model.....	333
12.7 Assessment of functional assessment scales.....	334
12.7.1 Hierarchical ordering by difficulty	334
12.8 Limitations of analyses of functional scales	335

12.8.1 Scoring bias of scales.....	336
12.8.2 Influence of scales analysed.....	338
12.9 Development and practical significance of hierarchical scales using Mokken scaling analysis.....	339
12.9.1 Comparison of Mini-ACE and Short ACE-R.....	340
12.9.2 Mini-NART.....	347
12.10 General limitations.....	348
12.10.1 Sample size	348
12.10.2 Manipulation of lowerbound threshold.....	352
12.10.3 Significance of standard errors and confidence intervals	352
12.10.4 Local stochastic independence.....	355
12.10.5 Items excluded from Mokken scales.....	357
12.11 Future directions and recommendations	363
12.11.1 Access to and analyses of large databases	363
12.11.2 Longitudinal analyses	365
12.11.3 Person fit analysis	366
12.12 Conclusion	367
References.....	368
Appendices.....	411
Appendix A: Addenbrooke’s Cognitive Examination scales	411
Appendix B: Systematic review search terms used for each database	425
Appendix C: Conversion of NART, Abbreviated NART and Mini-NART scores to predict premorbid IQ using regression equations derived from analyses in Chapter 8	428
Appendix D: Functional assessment scales	429
Appendix E: Published papers.....	431

List of Tables

3.1	Articles meeting inclusion criteria applying IRT methods to cognitive measures of dementia	61
3.2	Item <i>difficulty</i> comparison across studies	62
3.3	High <i>discrimination</i> items and disease stages	66
4.1	Demographic and cognitive scores in dementia groups	85
4.2	ACE-R items group by domain means and total scores	86
4.3	Correlations between ACE-R subdomains and component extracted from PCA.....	87
4.4	Items listed in order of decreasing <i>discrimination</i> for each of the three groups.....	96
4.5	IIO hierarchies with items ordered from most to least <i>difficult</i> in each diagnostic group	97
4.6	IIO hierarchies (using method MSCPM) with items ordered from most to least difficult for each diagnostic group	98
5.1	Comparison of the scale development and validation samples	111
5.2	Raw and equated mean ACE-III item scores for the Mini-ACE development sample (N=117) by cognitive domain	115
5.3	Hierarchical subset of ACE-III items revealed by Mokken scaling analysis of the Mini-ACE development sample. Items ordered from most to least <i>difficult</i> and most to least <i>discriminatory</i>	116
5.4	Equated means of ACE-III items and subdomain scores for each group in sensitivity analysis presented in order of the ACE-III	117
5.5	IIO hierarchies from sensitivity analyses: ordered from most to least <i>difficult</i>	122
5.6	Comparison of items common to all hierarchies from sensitivity analysis: ordered from most to least <i>difficult</i>	123
5.7	Mini-ACE items ordered by <i>difficulty</i> and <i>discrimination</i> for Mini-ACE validation sample	133
5.8	Mini-ACE items ordered by <i>difficulty</i> and <i>discrimination</i> (from most to least) for the Mini-ACE development and Mini-ACE validation samples	134
6.1	Demographic and cognitive information for diagnostic SDRIR groups	147

6.2	Mean equated ACE-R item scores for each SDRIR sample. Items presented in order of test administration.....	150
6.3	SDRIR mixed diagnosis sample: IIO hierarchy items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	153
6.4	SDRIR mixed diagnostic sample excluding possible stochastically dependent items: IIO hierarchy items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	155
6.5	SDRIR late onset AD: IIO hierarchy items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	158
6.6	SDRIR combined late and early onset AD: IIO hierarchy items from most to least <i>difficult</i> and most to least <i>discriminatory</i>	161
6.7	SDRIR mixed AD VaD: IIO hierarchy items ordered from most to least <i>difficult</i> and most to least <i>discriminatory</i>	164
6.8	ACE-R IIO hierarchies from most to least <i>difficult</i> : comparison across SDRIR groups.....	165
7.1	Comparison of the different scale development and validation samples.....	183
7.2	IIO hierarchy of ACE-R items (from analysis of SDRIR data ($N=808$) in Chapter 6) listed in descending order of <i>difficulty</i> and <i>discrimination</i>	188
7.3	Short ACE-R item selection 1 based on results of Mokken scaling analysis of SDRIR data ($N=808$)	190
7.4	Short ACE-R item selection 2 based on results of Mokken scaling analysis of SDRIR data ($N=808$)	194
7.5	Mean equated ACE-R scores from the Short ACE-R validation sample ($N=350$) for both SDRIR derived Short ACE-R scales	196
7.6	Short ACE-R item selection 1: items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	197
7.7	Short ACE-R item selection 2 items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	197
7.8	Mean item scores for the Short ACE-R development sample ($N=808$) and Short ACE-R validation sample ($N=350$) for two shortened versions of the ACE derived using Mokken scaling methods in order of item <i>difficulty</i> from most to least <i>difficult</i>	200
7.9	Mean item scores for the Short ACE-R validation sample ($N=350$) for the Short ACE-R and Mini-ACE. Items are presented in test order	208
8.1	Baseline sample characteristics.....	221

8.2	NART items ordered by percentage of correct responses in LBC1936 (n=587) (from least to most difficult)	224
8.3	Partitioning of NART items by the Automated Item Selection Procedure (AISP)	225
8.4	Items of the Abbreviated NART ordered by <i>discrimination</i> (H_i)	226
8.5	Item <i>difficulty</i> and <i>discrimination</i> of the Mini-NART	229
9.1	Mokken Scaling Procedure applied to Lawton IADL scale data from a mixed dementia SDRIR sample	247
10.1	Characteristics of the SDRIR samples analysed	259
10.2	Mean IADL item scores for SDRIR sample plus four diagnostic subgroups	260
10.3	IIO hierarchy items from complete sample (N=825) listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	262
10.4	Comparison of hierarchies from complete sample from the present study (N=825) and Chapter 9 (N=202). Items are listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	262
10.5	IIO hierarchy items from IADL in late onset AD listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	263
10.6	IIO hierarchy items from IADL in patients with Mixed AD VaD (N= 38) listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	264
10.7	IIO hierarchy items from IADL Mixed AD VaD and VaD (N=237) listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	265
10.8	IIO hierarchy items from IADL non-AD listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	266
10.9	IADL IIO hierarchies: comparison across diagnostic groups	267
10.10	IIO hierarchy items from full sample-male-listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	268
10.11	IIO hierarchy items from full sample-female- listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	269
10.12	IADL IIO hierarchies: comparison across gender	270
11.1	Characteristics of the SDRIR samples analysed	287
11.2	Mean PSMS item scores for SDRIR sample plus four subgroups	289

11.3	Mean PSMS and Lawton IADL scores for combined BADL and IADL sample (N=822). Items presented from most to least difficult, with lower mean scores reflecting poor functional ability	290
11.4	IIO hierarchy items from late onset AD sample listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	291
11.5	IIO hierarchy items from Mixed AD VaD sample listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	292
11.6	IIO hierarchy items from sample of male SDRIR participants listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	293
11.7	IIO hierarchy items from sample of female SDRIR participants listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	294
11.8	PSMS IIO hierarchical comparisons across groups; items listed from most to least <i>difficult</i>	296
11.9	Combined BADL IADL IIO hierarchy items listed from most to least <i>difficult</i> and most to least <i>discriminatory</i>	298
12.1	Comparison of all ACE-R and ACE-III IIO hierarchical Mokken scales	308
12.2	ACE-R IIO hierarchies from most to least <i>difficult</i> : comparison across Alzheimer's disease samples	312
12.3	Number of ACE-R/ACE-III items removed due to low scalability coefficients across samples.....	316

List of Figures

2.1	Item response function (IRF)	19
2.2	Four IRFs with different locations (<i>difficulty</i>) on the θ axis.....	21
2.3	Four IRFs according to the 2PLM with different locations on θ and different slopes (α).....	23
3.1	Flow diagram for review selection.....	53
4.1	Item-pair plot demonstrating intersection between ‘repetition-above, beyond and below’ and ‘reading’.....	91
4.2	Example of item-pair plot for ‘naming 2’ lying at some distance from a selection of remaining item-pair plots.....	92
4.3	Example of intersecting items from predominantly frontal dementia analysis.....	94
5.1	Item Characteristic Curves providing graphical representation of <i>discriminatory</i> power for: ‘identify fragmented letters’ (a) and ‘name and address learning’ (b)....	128
5.2	Range of <i>difficulty</i> coverage of item selection for the Mini-ACE.....	131
6.1	Venn diagram illustrating the relationship between the different samples used in these analyses	146
6.2	Item pair plots showing IIO violations between ‘Syntactical comprehension’ (ACERLangComprCommand) and ‘Fluency-animals’ (ACERFluencyAnimals) and ‘Draw a clock’ (ACERVisuosClock) in late onset AD analysis.....	157
6.3	Item pair plots showing IIO violations between ‘Syntactical comprehension’ (ACERLangComprCommand) and ‘Fluency-animals’ (ACERFluencyAnimals) and ‘Fluency-animals’ and ‘Naming 1’ (ACERLangNaming 1) in combined early and late onset AD analysis.....	160
6.4	Item pair plot demonstrating intersection between ‘Semantic Comprehension’ (ACERLangComprPictures) and ‘Fluency-Animals’ (ACERFluencyAnimals) in the mixed AD VaD analysis.....	163
6.5	Item pair plots demonstrating distance between ‘Naming 2’ (solid line) and other ACE-R items (dashed line) (‘Draw intersecting pentagons’, ‘3-item recall’, ‘Draw a cube’ and ‘Write a sentence’) in mixed AD VaD analysis.....	167
7.1	Venn diagram illustrating the different samples used in this Chapter.....	184

7.2	Flowchart illustrating the methods and samples used in this Chapter	185
7.3	Range of item <i>difficulty</i> of the Short ACE-R item selection 1 for the Short ACE-R development sample.....	191
7.4	Range of item <i>difficulty</i> of Short ACE-R item selection 2 for the Short ACE-R development sample.....	195
7.5	Comparison of the range of item <i>difficulty</i> of the Short ACE-R and Mini-ACE for the Short-ACE-R development sample ($N=808$).....	201
7.6	Comparison of the range of item <i>difficulty</i> of the Short ACE-R and Mini-ACE for the Short ACE-R validation sample ($N=350$).....	202
8.1	Correlations between age 11 IQ and the NART, Mini-NART, and age 70 IQ.....	228
8.2	Item response functions illustrating <i>discriminatory</i> power for two NART items: Item 43: ‘leviathan’, and Item 19: ‘radix’	231
10.1	Mean item scores for male and female participants.....	270
11.1	Mean item scores for male and female participants.....	295

Chapter 1: Introduction: Assessment in dementia -current and prior cognitive ability and functional ability

Dementia is a clinical syndrome characterised by progressive decline in cognitive functioning and the capacity to maintain independence in daily functioning. This deterioration is caused by several different underlying pathologies, the most common including Alzheimer's disease (AD), vascular dementia (VaD) and frontotemporal dementia (FTD).

Dementia is a worldwide concern and is a high political priority and health and social care responsibility in many countries. The ageing population is playing a substantial role in the emergence of the 'dementia epidemic' and is driving governmental responses in developing national dementia strategies and programmes. By 2050 people over 60 years of age will account for 22% of the world's population, an increase of 1.25 billion people (Prince et al., 2013). Estimates project the number of people living with dementia to almost double every 20 years to 115.4 million worldwide in 2050 (Prince et al., 2013). With the dramatic increase in the prevalence of dementia and related growth of memory clinics accurate diagnosis is crucial for appropriate management and assessment of disease progression.

The early and accurate diagnosis of dementia is desirable for the most effective management of the disease (Peterson, Stevens & Ganguli, 2001). Despite this it has been estimated that only a third to half of those living with dementia receive a formal diagnosis and often when they do, this diagnosis is made in the later stages of the disease progression (Bourne, 2007; Iliffe, Manthorpe & Eden 2003; Shankle et al., 2005). In a study of acute hospital admissions it was found that only half of the 42% of over 70s with dementia had ever received a diagnosis (Sampson, Blanchard, Jones, Tookman & King, 2009). Furthermore where cases are detected the situation is further complicated by the diagnostic confusion between dementia pathologies. Due to overlapping neuropsychological features distinct

CHAPTER 1: INTROUDUCTION TO ASSESSMENT IN DEMENTIA

clinical phenotypes can be difficult to establish, particularly in those with mixed pathology. A recent retrospective analysis of patients who were clinically diagnosed with Alzheimer's disease while alive found that 63% had other pathology in addition to Alzheimer's disease (Wang et al., 2012). With reports of between 12 % and 23% of those diagnosed with Alzheimer's disease lacking Alzheimer's disease pathology upon autopsy it appears misdiagnoses is not uncommon (Lim et al., 1999; Ranginwala, Hynan, Weiner & White, 2008; Pearl, 1997; Klatka, Schiffer, Powers, & Kazee, 1996).

This diagnostic confusion has considerable significance given the development of disease specific treatments which emphasizes the importance of establishing reliable diagnoses. A study comparing cases of correctly diagnosed Alzheimer's disease with those who were misdiagnosed found that 18.2% of those misdiagnosed were treated with potentially inappropriate medication (Gaugler et al., 2013). The challenge of effective and accurate dementia detection may reflect its frequently unclear aetiology and pathophysiology, variability in symptoms or weaknesses with screening instruments and assessment measures (Iliffe, Manthorpe & Eden 2003).

While great efforts are being made to identify the physiological origins of dementia, there remains no definitive biological markers for the most commonly found forms of dementia apart from autopsy. The lack of conclusive indicators in life means that neuropsychological assessment and cognitive testing remain the most effective method of differential diagnosis in the discrimination of dementia from age-related cognitive decline, cognitive deficits due to depression or other related conditions (Mathuranath, Nestor, Berrios, Rakowicz & Hodges, 2000). Assessment scales measuring cognitive and functional ability play an important role in the detection and assessment of the changes in cognitive ability that occur in dementia by reducing subjectivity. These instruments evaluate changes in memory

CHAPTER 1: INTROUDUCTION TO ASSESSMENT IN DEMENTIA

and wider cognitive ability levels and help to discriminate expected age related decline from the first signs of pathological decline.

Scales must have the necessary psychometric features, i.e. they must be sensitive and reliable and should have normative data directly referable to appropriate populations (Force, 1998). Range of assessment of the scale is also important in ensuring that the full breadth of cognitive impairment is assessed. Crucially a scale must be sufficiently sensitive to detect small but significant changes in ability. Ideally an assessment scale in dementia should provide sensitive discrimination at levels of ability covering the spectrum of impairment associated with dementia. It should contain items relevant and sensitive to mildly impaired, severely impaired and high-functioning patients. Its range of use and application would also be increased if it can be administered in a brief period of time.

Assessment scales facilitating the reliable detection and diagnosis enable the appropriate action to be taken and can provide further insight into the disease. With the development of cholinesterase inhibitors along with non-pharmacological treatment options and support models for family and carers the need for an accurate and timely diagnosis is of considerable practical significance. Rigorous cognitive assessment is a key mechanism through which progress can be made in terms of determining the efficacy of interventions and ensuring the appropriate care and support is in place for the individual and carers. Early diagnosis is at the forefront of secondary prevention, identifying preclinical and prodromal phases during which early treatment and intervention has been found to be more effective than in later stages (Hinton, Franz & Friend, 2004) and in tertiary prevention where precise monitoring of stages and symptoms can contribute towards the provision of the best possible care and treatment for individuals with diagnosed dementia.

Diagnostic criteria for dementia include a decline in cognitive functioning along with impairment in functional ability (DSM V, American Psychiatric Association, 2013).

Establishing whether deterioration in cognitive ability has taken place relies on ascertaining a valid estimate of prior ability level (Crawford, Moore & Cameron, 1992). Preferably, this would involve a comparison of current cognitive ability with an actual measure of prior or premorbid cognitive ability. However, such premorbid measures of ability are seldom available which results in the dependence upon estimates of premorbid cognitive function. Therefore assessment in dementia must involve measurement of current and prior cognitive function as well as a functional assessment.

1.1 Cognitive assessment

In both clinical and research settings much of the focus in diagnosing and monitoring dementia is on cognition. The National Institute for Health and Care Excellence (NICE) Clinical Guidance 42 (2006) recommendations for the diagnosis and assessment of dementia include the use of formal cognitive testing using a standardized instrument as part of a clinical cognitive examination to assess the patient's attention, concentration, memory, orientation, language, praxis, and executive function. Cognitive measurement scales assist in dementia assessment by: (i) screening for cognitive dysfunction; (ii) assisting in differential diagnosis; (iii) determining the severity of disease and (iv) tracking and monitoring disease progression. Given these various applications it is important to be aware that different scales may have different sensitivities which restrict them to being most effectively applied as identifying, staging or monitoring instruments (e.g. a measure might be good at assisting in initial diagnosis but demonstrate limited use for monitoring and quantifying disease progression).

CHAPTER 1: INTROUCTION TO ASSESSMENT IN DEMENTIA

A cognitive assessment tool would ideally comprise the psychometric features necessary to be applicable in both clinical and research settings. Scales should be standardized and reliable. They should be validated in appropriate populations and have discriminant and convergent validity. If scales do not meet these criteria this could limit the applicability of the scales with the potential for valuable information regarding the efficacy of pharmacological or behavioural treatment to be lost or inaccurate. For example, ensuring that scales used in clinical trials are sensitive to change in total scores both at early and later stages of disease can assist in improving the rigor of clinical practice and research outcomes.

A comprehensive and varied range of assessment and screening tools have been developed to quantitatively assess cognition in dementia. These tools range from very brief screening assessments to more comprehensive and lengthy formal neuropsychological assessments. The 30-point Mini Mental State Examination (MMSE) (Folstein, Folstein & McHugh, 1975) became the de facto standard for cognitive assessment due to its ease of use and coverage of cognitive functions such as memory, orientation, arithmetic, language comprehension and visuospatial ability. However, due to the emergence of a range of alternatives plus the risk of copyright infringement, as well as concerns including its limited assessment of some cognitive domains e.g. executive function, alternative scales are becoming more commonplace in clinical cognitive assessments. These other frequently used cognitive assessments in clinical practice in the UK include the Addenbrookes Cognitive Examination (ACE) (Mathuranath, Nestor, Berrios, Rakowicz & Hodges, 2000) and the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005).

While the MMSE has been widely used in clinical evaluation and research applications it has specific limitations. The insensitivity of the MMSE to the initial stages of AD, particularly in younger patients with high intellectual ability (Welsh, Butters, Hughes, Mohs & Heyman, 1991; Welsh, Butters, Hughes & Mohs, 1992), and to deficits associated

CHAPTER 1: INTROUDUCTION TO ASSESSMENT IN DEMENTIA

with early FTD; isolated frontal or linguistic impairment (Gregory, Orrell, Sahakian & Hodges, 1997) along with difficulties in differentiating dementia syndromes prompted the development of the ACE. The ACE was developed as a theoretically motivated clinical tool sensitive to the symptoms of early dementia and to address a need for better differentiation of dementia subtypes including Alzheimer's disease, frontotemporal dementia and Parkinsonian syndromes (Mathuranath et al., 2000). An instrument capable of reliable differentiation of different forms of dementia has clinical relevance for the treatment and management of patients can guide the advice and information given to carers regarding prognosis.

The ACE encompasses the items of the MMSE while expanding on the memory, language and visuospatial domains with the addition of items assessing verbal fluency. The ACE comprises items assessing cognitive functions across six domains; orientation, attention, memory, fluency, language and visuospatial skills. Scores from these domains can be calculated separately and summed to provide a composite score out of a maximum score of 100. An MMSE score can also be calculated. The ACE at a cut-off of 88 had high reliability, construct validity, and sensitivity (93%) in the detection of patients with early dementia (Mathuranath et al., 2000). The original ACE underwent a revision to enhance sensitivity and specificity for detecting dementia. Its successor, the Addenbrookes Cognitive Examination-Revised (ACE-R) (Mioshi, Dawson, Mitchell, Arnold & Hodges, 2006) divides the 26 items into five distinct cognitive domains; attention and orientation, memory, fluency, language and visuospatial. The modifications saw the magnitude of the memory domain diluted to provide a more balanced contribution across the five domains to the overall score. Several modifications were made, for example two further items assessing visuospatial abilities were added and the pictures in the naming task were changed to reduce the ceiling effects this item suffered in the original. The items allowing an MMSE score to be extracted were retained.

Further refinements have been made culminating in the development and validation of the Addenbrooke's Cognitive Examination III (ACE III) (Hsieh, Schubert, Hoon, Mioshi & Hodges, 2013), which remains scored out of 100 by summing the score from each of the five cognitive domains. The ACE-III compares favourably to the ACE-R retaining the high levels of sensitivity and specificity of its predecessor but does not include the MMSE items. Both the ACE-R and ACE-III can be viewed in Appendix A.

These cognitive scales have proven reliable and valid measures of current cognitive ability. However to establish a diagnosis of dementia, it must be established whether this current ability reflects a decrement from a previously higher level of ability.

1.2 Estimation of premorbid cognitive ability

Establishing a diagnosis of dementia requires a decline in cognitive ability relative to the individual's prior level of functioning. Therefore in the measurement of cognitive decline it is important to establish a reference point against which current cognitive performance can be compared. Unfortunately this kind of baseline test data is often unavailable which necessitates the need to develop a proxy for premorbid ability. An estimated prior level of ability can be used to detect and measure the rate and extent of cognitive deterioration. Generally two approaches have been applied in the estimation of prior ability; (i) a demographic-based regression approach and (ii) current or "hold" ability approaches.

The first of these approaches saw demographic variables such as socioeconomic status, educational attainment and occupational history converted into formulas designed to estimate premorbid cognitive functioning. Wilson et al. (1978) developed multiple regression equations to estimate premorbid ability from demographic variables (age, sex, race, education and occupation). The variance explained between all five variables and Wechsler Adult Intelligence Scale (WASI; Wechsler, 1955) Verbal (VIQ), Performance (PIQ) and Full Scale

CHAPTER 1: INTROUDDCTION TO ASSESSMENT IN DEMENTIA

IQ (FSIQ) was 53%, 42% and 54% respectively. These equations were applied in a UK sample by Crawford et al., (1989) where 50%, 30% and 50% of the variance in VIQ, PIQ and FSIQ respectively.

There has been also been considerable interest in the identification of measures of crystallised intelligence or “hold” abilities that are not susceptible to the effects of dementia that can be used to estimate premorbid cognitive ability with any discrepancy between current and prior level of ability serving as an estimate of the extent of decline. One of the most commonly used measures is the National Adult Reading Test (NART; Nelson, 1982). In clinical assessments it was observed that the ability to read aloud appears largely impervious to the effects of dementia in comparison to other cognitive abilities. Nelson and McKenna (1975) hypothesised that as reading ability was highly correlated with general intelligence in healthy populations the ability to read could be used to estimate premorbid levels of intelligence in dementia.

The NART is a word reading list comprised mostly of irregular words which cannot be pronounced correctly through the application of common phonetic rules. For example, the correct pronunciation of the word “ache” cannot be deduced by following the standard grapheme phoneme rules. Whereas reading regular words could largely rely on the individual’s ability to apply the spelling-to-sound conversion rules, the reading of irregular words is dependent of the individual’s familiarity with the words prior to disease onset, therefore serving as a more reliable estimation of premorbid cognitive ability.

The NART requires the reading of 50 words aloud with the response to each scored individually as correct or incorrect according to the pronunciation. This total score is used to provide an estimate of premorbid intelligence. The NART has impressive reliability. Internal reliability is reported as 0.93 (Nelson & Willison, 1991), with inter-rater reliability reports ranging between 0.96 and 0.98 (O’Carroll, 1987; Crawford et al., 1989) and a test/re-test

CHAPTER 1: INTROUDCTION TO ASSESSMENT IN DEMENTIA

reliability coefficient of 0.98 (Crawford et al., 1989). Importantly studies investigating retrospective validity rather than relying on concurrent validity have found the scale to account for over 50% of the variance in childhood intelligence scores (Crawford, Deary, Starr & Whalley, 2001).

O'Carroll (1992) concluded:

“It is, of course, highly unlikely that any cognitive measure will prove to be entirely dementia-resistant. However, it would appear that the ability to correctly pronounce irregular words ‘hold’ better than other ‘current’ ability measures, at least in the earlier stages of dementia where diagnostic problems typically occur” (p.114).

Instruments assessing premorbid ability, such as the NART, have greater predictive accuracy than the demographic approach (Crawford et al., 1989). Furthermore a combination of the two approaches - the NART and demographic variables - did not significantly increase the amount of variance in premorbid cognitive ability explained by the NART alone (Bright, Jadow & Kopelman, 2002). Given the superiority of the NART over other methods with regards to correlations with current intelligence in both controls and patients and high levels of inter-rater and test/retest reliability (Crawford et al., 1989; O'Carroll, 1987), it appears to be relatively resistant to dementia providing justification and support for its continued application in the estimation of prior cognitive ability in dementia.

Current and premorbid cognitive assessment plays a valuable role in facilitating the diagnosis and understanding of disease progress. It also permits a better estimation of related functional impairment to be made.

1.3 Functional assessment

Establishing a dementia diagnosis involves a decline in cognitive ability sufficient to have significant and detrimental effect on the individual's work, relationships with others and typical social activities (DSM V, American Psychiatric Association, 2013). This loss of functional independence is widely acknowledged as a source of significant social, health and economic cost. The progressive deterioration of functional ability increases the burden of dementia for patients themselves along with their carers and the wider society. Progressive functional impairment, proportional to the severity of dementia and associated loss of independence, can be the most obvious manifestation of dementia (Potkin, 2002).

The assessment of non-cognitive variables forms an important part in the overall assessment of an individual with dementia and is important in establishing a diagnosis and evaluating and quantifying change. Several instruments have been developed for the evaluation of functional decline by assessing performance on various tasks known as Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL). ADL scales assess abilities such as mobility, toileting, feeding, bathing and dressing. These skills are highly overlearned and are inclined to be reliant on motor-learning and praxis. IADL activities include more complex tasks such as doing housework, handling finances and shopping. These more involved activities tend to place significant demands on memory, attention, judgment and language processes. Generally IADLs, especially those reliant on memory, are lost before the more basic ADLs and both IADLs and ADLs are lost in a hierarchical fashion with skills contingent on short-term memory processes lost before overlearned skills (Galasko et al., 1995). Several IADL scales demonstrate a floor effect with patients with dementia losing the ability to perform these complex tasks early in the course of the disease. The opposite effect is found in ADL scales where ceiling effects are often found due to patients retaining the ability to perform these more basic and fundamental activities

CHAPTER 1: INTROUDDCTION TO ASSESSMENT IN DEMENTIA

until late the course of the disease as the disease becomes more pervasive (Spector, Katz, Murphy & Fulton, 1978). The observation of impairments in ADL or IADL at an earlier stage of disease than expected should prompt the implementation of physical or environmental interventions.

The Physical Self Maintenance Scale (PSMS) and Lawton-Brody Instrumental Activities of Daily Living (IADL) scale (Lawton & Brody, 1969) were among the earliest scales developed to measure an individual's capacity to perform tasks related to the maintenance of self and lifestyle. The PSMS assesses six activities: toileting, feeding, dressing, grooming, locomotion and bathing. Each ADL is rated on a five point scale of responses ranging from complete independence to total dependence. Each task is scored dichotomously as either 1 (can perform task) or 0 (cannot perform or requires some assistance) yielding a total score ranging from 0 to 6. Each of the six items have several different response options offering further grading of functional ability with the item description most closely resembling the patient's highest functional level selected. The Lawton IADL scale is a self-report scale measuring the ability to perform eight functional tasks; the ability to use the telephone, to shop, prepare food, handle finances, do housework, take medications, do laundry and to travel. Self-reported performance on each of these tasks provides information about functional abilities required to live independently in the community. Again each task is scored dichotomously as either 1 (can perform task) or 0 (cannot perform or requires some assistance) providing a total IADL score ranging from 0 to 8. For both scales in the assessment of some skills only the highest level of ability receives a score of 1 whereas for other items more than one level of ability receive a score of 1 as they indicate some minimal level of function.

CHAPTER 1: INTROUDUCTION TO ASSESSMENT IN DEMENTIA

There are several reasons why the assessment of functional ability in dementia is important. Functional scales can help to clarify the link between cognition and functional ability (Patterson, et al., 1992). While the measurement of functional performance will also capture non-cognitive processes such as normal aspects of ageing; deterioration of hearing and general physical functions including gait, mobility and strength cognitive ability is the most significant determinant of functional ability. Correlations between performance on cognitive and functional tests have been found ranging between 0.5 and 0.8 reflecting how the level of functional ability can be predictive of overall dementia severity (Galasko, 1998). Cognitive deterioration in dementia will manifest in functional performance and vice versa. Analysis of functional assessment scores and items can contribute to a better understanding of how cognitive processes underlie functional performance and yield valuable insight into the staging of dementia.

The ability to accurately measure and quantify functional ability is necessary to elucidate the relationship between cognitive impairment and functional decline. To understand how cognitive processes may cause functional difficulties the ADL/IADL items must be broken down into their constituent parts to more precisely observe the link between cognition and functional ability. Cognitive impairment in dementia can cause difficulties with many such activities for various reasons; attention, memory or concentration problems, motor skills deficits, lack of motivation, impaired executive functions, failure to initiate or maintain tasks or failure to perform activities without direction. Examining performance on ADL/IADL items can aid in deciphering the link between the functional outcome and neuropsychological cause. As IADL tasks are more complex and involve more complex neuropsychological organisation they are heavily reliant on cognitive abilities making them very susceptible to the initial effects of cognitive impairment (Njegovan, Man-Son-Hing,

CHAPTER 1: INTROUDCTION TO ASSESSMENT IN DEMENTIA

Mitchell & Molnar, 2001). Therefore the assessment of IADL can prove valuable in the detection and diagnosis of early dementia (Desai, Grossberg & Sheth, 2004).

In addition to aiding and corroborating a diagnosis of dementia functional ability scales can be used as secondary outcome measures in the assessment of cognitive ability with the expectation that any treatment that improves cognitive ability will bring about improvement in functional ability. Evidence of a transfer of the effects of cognitive training to functional ability suggests that cognitive interventions can prevent or delay functional decline (Willis et al., 2006). In this way any change in functional behaviour may serve as important markers of the effectiveness of such treatments or interventions. The measurement of responses to treatment and interventions requires effective scales to assess functional decline in dementia.

Functional assessment instruments are also valuable for clarifying the relationship between functional performance and non-cognitive emotional and behavioural changes in dementia such as frustration, wandering, depression and apathy. The progressive deterioration of ADL/IADL performance and associated loss of independence is associated with loss of self-esteem and poorer quality of life (Andersen, Wittrup-Jensen, Lolk, Andersen, & Kragh-Sørensen, 2004). Continued involvement in normal functional activities and tasks is important to maintain self-esteem (Patterson et al., 1992). Where possible activities should be simplified to enable the patient to participate in certain steps where the entire task is beyond their capability. Adequate estimation of functional abilities can help in the assessment of care-giver burden and to develop and implement appropriate personalised interventions which contribute to ensuring the appropriate level of assistance is in place, avoiding unnecessary interventions which could lower the patient's self-esteem. Whereas overly restrictive support could result in frustration, depression or aggressiveness an

CHAPTER 1: INTROUDCTION TO ASSESSMENT IN DEMENTIA

insufficient level or lack of support presents a safety concern and could cause stress and worry for the patient and carers. Therefore it is important that functional assessment scales are accurate and reliable.

The prognostic relationship between cognitive and functional impairments and the manner in which each relate to dementia progression can be further elucidated by the simultaneous analyses of both forms of assessment. This concurrent analysis could enable the discovery of more exact connections between cognitive and functional processes (McGough et al., 2011; Tschanz et al., 2011).

Dementia is clinically characterized by a progressive decline of cognitive ability sufficient to impair functional ability and cause behavioural disturbances. Cognitive and functional assessment scales developed to specifically tap and assess these functions are applied clinically to facilitate dementia detection, diagnosis and monitoring. These scales, aside from being simple and cost effective must be reliable and valid (Reid, Lachs, Feinstein 1995; Sackett, 1992). While there is no definitive ‘dementia test’ the concurrent application of measures of current and premorbid cognitive and functional impairment contributes significantly to the identification and understanding of dementia.

Chapter 2: Can item response theory methods be used to improve accurate assessment of dementia?

2.1. Problems with existing measurement methods

Accurate assessment of cognition in dementia is necessary in order to develop procedures for prevention and treatment of dementia. Precise measurement can help to further our understanding of the manifestation and progression of the disease. It is necessary in the assessment of patients' responses to experimental pharmacotherapy and therapeutic interventions and in its application to helping delineate various forms of dementia by identifying different patterns of cognitive decline for differential diagnosis.

In practice however measuring cognitive and associated functional impairments in dementia is not without its challenges. Typically a patient's cognitive functioning is examined by summing their responses to test items to reach a total score. While this is a quick and easily interpreted method it may yield a relatively inaccurate estimate of underlying cognitive impairment. Total scores can provide rather imprecise estimates of cognitive impairment as for any given total score there is a range of associated latent scores. Summed total scores do not take into account important differences in item characteristics and the different possible patterns of responses. It is possible for two individuals to respond differently to the items within a scale yet reach the same summed score. A patient may get a total score via numerous different possible combinations of response (Balsis, Lowe & Bengel, 2012). With regards to the ACE-R, which is scored from 0 to 100, with 0 reflecting severe cognitive dysfunction and 100 reflecting preserved cognitive functioning, there are 101 different possible total scores. A respondent may respond incorrectly to any combination of

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

items across the five domains assessed to yield a given total summed score. There are over 24 quadrillion possible patterns of response across the items that lead to any one of the 101 possible total scores on the ACE-R. This can be calculated according to the rule of permutations. There are 26 different items and across these 26 items there are different possible raw scores (there are 5 items with 6 possible scores (0-5), 3 items with 4 possible scores, 4 items with 8 possible scores, 4 items with 5 possible scores, 6 items with 2 possible scores, 3 items with 3 possible scores, and 1 item with 11 possible scores). The total number of possible response patterns across the items is: $6^5 \times 4^3 \times 8^4 \times 5^4 \times 2^6 \times 3^3 \times 11 = 2.421657 \times 10^{16}$. Out of these possible patterns some of course will be much more likely to occur than others. However the vast array of possible patterns helps to illustrate the importance of looking beyond the total score. Significant information may be missed by disregarding the pattern of scores.

Items within cognitive assessment scales such as the ACE-R differ from each other, not just in how difficult they are but also in how strongly related they are to the latent construct. Two respondents can achieve the same score on a scale by responding incorrectly to very different items; respondents may make errors on items less strongly associated with cognitive impairment, or the more difficult items, while others could make errors on items very strongly associated with cognitive impairment, less difficult items or a combination of easy and difficult items. Classifying patients with the same total score as having the same degree of cognitive impairment could be inaccurate.

The use of summed total scores to reflect the level of cognitive or functional impairment has significant drawbacks particularly if used to assess longitudinal change. Total scores can conceal important changes in degree of impairment. Even if a patient receives the same total score on two occasions it is not until the individual pattern of item response is examined can we determine whether any change has occurred. Equally, it is possible that a

patient who receives a different score at two time points has undergone no significant change in their degree of impairment. Alternatively a patient achieving the same total score following treatment could be considered a nonresponder even if their response pattern was different before and after treatment. Therefore a more sophisticated measurement system is needed to move beyond these existing limitations.

2.2 Classical test theory

Traditionally analysis involved an application of classical test theory (CTT).

CTT is a relatively simple psychometric model for testing with extensive application in scale construction and assessment of tests. Spearman laid the foundations with the introduction of the notion of an observed score arising from an element of true score and a random error score. In its basic form the equation of CTT assuming the raw score (X) is comprised of a true score component (T) and an error term (E) is:

$$X=T+E \tag{2.1}$$

While CTT permits the prediction of test outcomes such as the ability of the respondent and item difficulty, it cannot guarantee that the method of scoring will provide equally distributed measurement precision across the latent trait being measured (Fraley, Waller & Brennan, 2000). Moreover, the features and properties of the aggregated total score are dependent on the sample properties as well as the number of items within (Hambleton, Swaminathan & Rogers, 1991). In order to circumvent these issues a model capable of relating latent traits to responses to individual items is required (Fraley et al., 2000).

To obtain a more accurate assessment of any possible change between and within patients over time and in response to treatment a more sophisticated statistical framework should be considered. Item response theory (IRT) (Hambleton & Swaminathan, 1985; Lord, 1980) models which have become very popular in modern psychometric test development can be used to determine whether cognitive impairment can be better measured and understood with item properties and patterns taken into account. CTT and IRT represent two very distinct statistical measurement frameworks. An important advantage of IRT models over CTT models is that the former involve item parameters allowing for the explicit assessment of item properties.

2.3 Item response theory

Item response theory (IRT) is model-based measurement in which the trait level estimate is dependent on both the individual's response and the properties of the items within the administered test (Embretson & Reise, 2000). IRT describes the relationship between an individual's trait level and the probability of a given response to an item using a nonlinear monotonic function (Reise, Widaman & Pugh, 1993). Within the IRT framework ability or level of latent trait is represented by theta (θ), which determines each respondent's item and test performance. Ability or θ will be used from here on to denote latent trait level. Θ , on which both respondents and items have a position, is an unknown parameter, which cannot be explicitly quantified but is estimated through IRT analysis.

Modern IRT models are stochastic in nature where the probability of responding correctly to any given test item is based on the individual's ability or trait level and item parameters. This probability is referred to as the item response function (IRF) or item

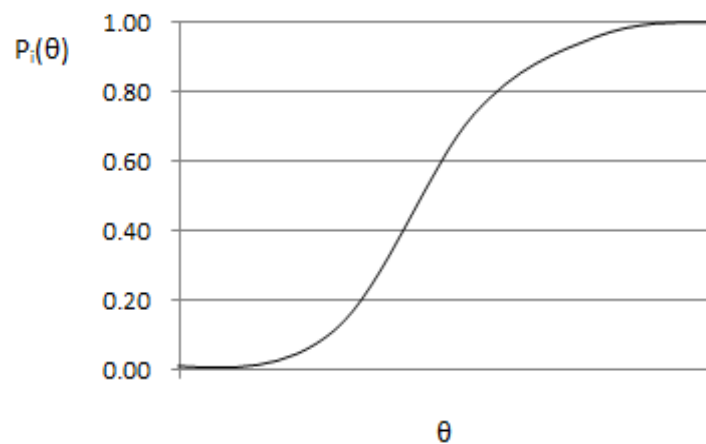
CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

characteristic curve (ICC). The IRF models the behaviour of items relative to the latent trait.

With θ representing the underlying latent trait the IRF represents $P(\theta)$ for each item.

Fundamentally the IRF is a non-linear regression on ability of the probability of a correct response to an item (Mungas & Reed, 2000). For each item within a scale the IRF describes the probability of a respondent's score on the item for a given degree of latent trait, with the probability increasing with increasing levels of latent trait in a non-linear fashion. Many IRT models assume that the IRF of each item conforms to the specific shape of a logistic curve. Generally speaking, the IRF illustrates that the higher the latent trait value θ , the higher the probability of responding correctly on an item measuring θ (see Figure 2.1).

Figure 2.1 Item response function (IRF)



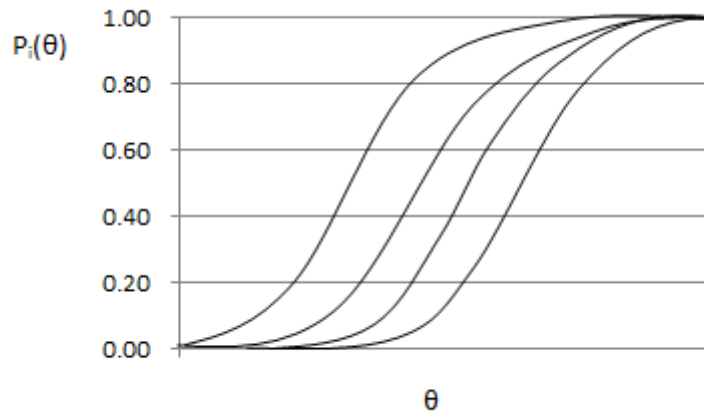
Where X_i equals the score on a dichotomously scored item (0 for incorrect, 1 for correct) the IRF is defined:

$$P_i(\theta) = P(X_i=1|\theta) \quad (2.2)$$

However this function is solely dependent on θ and does not explicitly show the effects of item properties. More detailed models of IRFs expand this function to relate to response probabilities of specific items.

2.3.1 The one-parameter logistic model

The most frequently used IRT model is the one-parameter logistic (1PLM) or Rasch (Rasch, 1960; 1966) model (Jenkinson, Fitzpatrick, Garratt, Peto & Stewart-Brown, 2001). Items within a measure will vary in terms of difficulty and these differences affect the expected response probability. The 1PLM or Rasch model accounts for this with a logistic function that depends on the difference between θ and the item's location parameter, δ_i , often interpreted as item *difficulty*; see Figure 2.2. In IRT an item has high *difficulty* if a high level of ability or latent trait is necessary to respond correctly. In the context of parametric IRT this *difficulty* parameter is defined as the θ value required for a respondent to have a .50 probability of correctly responding to the item. For example, if an item has a *difficulty* level of 2.0 then a respondent with a corresponding trait level of 2.0 would have a 50% chance of correctly responding to the item.

Figure 2.2 Four IRFs with different locations (*difficulty*) on the θ axis

The IRF for item i is defined as:

$$P_i(\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} \quad (2.3)$$

The 1PLM transforms the sum of the number of items correct to the scale of θ and in doing so considers all items equally important for measuring. The 1PLM assumes all slopes of IRFs are the same. This restrictive assumption is akin to assuming the relationship between the item score and the latent trait in regression analyses is the same across all items.

2.3.2 The two-parameter logistic model

To allow for different degrees of association between the items and the latent trait the 2-parameter logistic model (2PLM) takes the slope parameter for item i (α_i) into account.

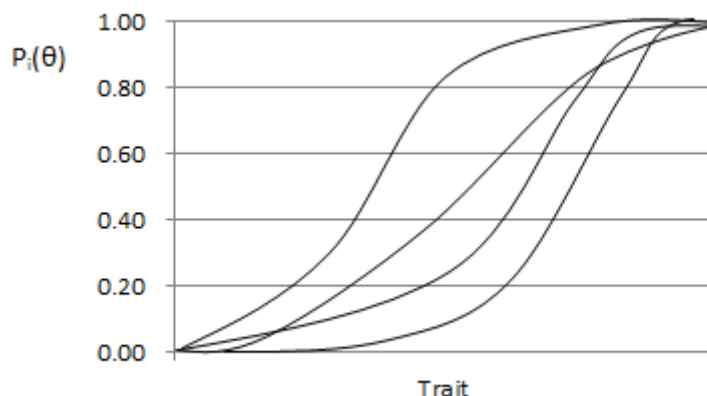
The 2PLM IRF for item i is defined as:

$$P_i(\theta) = \frac{\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]} \quad (2.4)$$

This slope parameter is often interpreted as item *discrimination*. Whereas the 1PLM estimates a single slope for all items the 2PLM defines a specific *discrimination* value for each item within the scale. The steeper the slope the more accurately and precisely the item can differentiate between respondents with low levels of ability or latent trait (low θ) and those with high levels of ability or latent trait level (high θ). The 2PLM weights each item score by the slope, or *discrimination* of the item, giving more weight to more *discriminatory* items, before transforming the sum of the weighted scores to θ . Figure 2.3 shows four IRFs with different locations, or *difficulty* levels, and different slopes or *discrimination*.

Both the 1PLM and 2PLM are parametric models as they estimate the relationship between $P_i(\theta)$ and θ by means of a logistic function with scalar parameters, δ for the 1PLM and both δ and α for the 2PLM.

Figure 2.3 Four IRFs according to the 2PLM with different locations on θ and different slopes (α)



2.3.3 Nonparametric item response theory

While the Rasch is a popular model, its assumptions make it restrictive and it is best applied when the number of items is relatively high (e.g. over 20). Analysing a high number of items is desirable as it increases the likelihood of a reasonable number of items fitting the restrictive model (van Schuur, 2003). Also fewer items are likely to meet the assumptions of parametric models in comparison to nonparametric IRT (NIRT). NIRT is a family of statistical models which use the minimum number of assumptions necessary to allow assessment of both items and respondents. In this way they provide a more flexible framework for the analysis of item responses.

The IRFs in NIRT models are not parametrically defined and may take on any shape and no assumptions are made regarding the distribution of the latent trait. Whereas parametric IRT may be unduly restrictive with regards to form, IRFs under the NIRT framework may be described by linear, partly linear or exponential equations, they need not be symmetric and

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

can take any functional form so long as the relationship between $P_i(\theta)$ and θ is monotonically nondecreasing.

The paucity of assumptions of NIRT models means they can fit many data sets, yet NIRT models remain sufficiently powerful to provide many valuable measurement properties. The main advantage of NIRT is the loosening of some of the stringent assumptions concerning the form of the response probabilities in parametric IRT models (Sijtsma & Molenaar, 2002).

Although parametric models may be favoured for their familiarity and for the interpretability of their item parameters and results they are also associated with a smaller number of items conforming to the final scale. This is a consequence of potentially valuable items being excluded due to the shape of their IRFs. Under a parametric model more items are likely to be rejected as poorly fitting. Rejecting too many items can result in low reliability values for the remaining scale due to a limited number of items.

While parametric models allow numerical estimates of person and item parameters in many measurement applications the ordering of respondents on a latent trait is sufficient. With this in mind Robert J. Mokken (Mokken, 1971; Mokken & Lewis, 1982; Mokken, 1997) developed two models, jointly referred to as Mokken scaling, which use the unweighted total score for rank ordering of respondents on θ . Mokken scaling analysis is a probabilistic reworking of the deterministic Guttman scaling.

2.4.Mokken scaling-Origins in the Guttman scale

Developed by Louis Guttman (Stouffer et al., 1950) Guttman scaling analysis was designed to accurately predict item responses from the total score. This method of analysis arose in an effort to make up for the limitations of summing ordinal rated items, which may hinder interpretation of the results. Guttman scaling does not permit a probabilistic relationship between the item responses and the underlying construct (Croon, 1991). Being deterministic it views the association as a clear dichotomy between the presence or absence of the trait solely determined from endorsement or lack of endorsement of the item.

In a perfectly unidimensional Guttman scale a respondent who endorses a more *difficult* (or more unpopular) item will have endorsed all less *difficult* (more popular) items, likewise a respondent who fails to correctly respond to a *difficult* item will also fail to correctly respond to any of the more *difficult* items.

In this way a Guttman scale is cumulative, i.e. all respondents accumulate responses to the items in the same consistent order, from the least *difficult* to the most *difficult* items. Therefore all respondents follow the same pattern of item endorsement and those with the same total score will have identical responses to each item within the scale. This allows us to predict all item responses from a respondent's total score. For example, a participant with a total score of 6 will have endorsed or correctly responded to the six least *difficult* items. In this way a score on a Guttman scale identifies which items have been endorsed by that individual.

Guttman scaling has been criticised for its deterministic nature. A deterministic model based on the belief that respondents will correctly endorse all items below their level of ability and incorrectly endorse all items beyond their ability level, is unrealistic as there is always some element of error in measuring latent traits (Bond, Ughrin & Fox, 2001).

Guttman scaling does not provide any insight into sampling error and does not take into account that, practically speaking, response patterns involve more than just the underlying trait. The mood and motivation of the participant or their interpretation of the item can influence the responses (Kempen, Myers & Powell, 1995). For this reason the deterministic nature of Guttman is rather limiting and the relationship between the items and trait is better conceptualized as probabilistic.

It is very unusual to obtain data perfectly conforming to a Guttman scale as this scaling unrealistically assumes the data are error free. As a consequence of these deterministic and unrealistic properties of Guttman scaling in addition to the creation of more sophisticated stochastic models (i.e. item response theory analysis), Guttman scaling methods have seen a decline in applications in health and psychological research (Vittengl, White, McGovern & Morton, 2006).

2.4.1 Mokken scaling analysis

Mokken scaling is a nonparametric model as the IRFs are not parametrically defined and no assumptions regarding the distribution of the latent trait are made. Mokken scaling comprises two nested scaling models; the monotone homogeneity model (MHM) and double monotonicity model (DMM). Mokken scaling's first model; the MHM, effectively enforces no other assumption than nondecreasing IRFs enabling respondents to be ordered with respect to the latent trait. The second Mokken model; the DMM, enables the ordering of the items with the additional restriction that not only do the IRFs increase but they also cannot intersect. Like the 2PLM Mokken scaling allows the estimation of both item *difficulty* and *discrimination* parameters.

The first and least restrictive model of Mokken scaling, the monotone homogeneity model (MHM) comprises three assumptions common across IRT models: unidimensionality,

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

local independence and monotonicity. The assumption of unidimensionality means that all of the items within the scale measure the same underlying latent trait, denoted by θ . There are two interpretations of this assumption; the psychological interpretation and the mathematical interpretation (Sijtsma & Molenaar, 2002). Unidimensionality from a psychological perspective means that all the test items measure one construct, for example quality of life of respondents or their ability to solve linear equations. From a mathematical perspective unidimensionality means that only one latent trait (e.g. quality of life) is necessary to account for the inter-item correlations in the data. Unidimensional measures simplify the interpretation of the test scores by assessing only one latent trait at a time.

Local independence is a strong assumption meaning that a respondent's response to one item in the test is not affected by his or her response to any other item in the scale. Local independence implies that all systematic variation in responses to the items is exclusively caused by the variation of respondents over θ (Mokken, 1997). Local independence implies that all items are uncorrelated with each other when the latent trait is controlled for (McDonald, 1981). In other words, item correlations are completely accounted for by the latent trait. This assumption can easily be violated. For example a respondent's score in a test of mathematical ability may change, improving as they gain familiarity with the type of calculations involved or from knowledge gained from previous items. In this way local independence is violated by learning through practice. Equally it is also possible that scores decrease as a result of fatigue and loss of focus causing the latent trait to be thought of as decreasing. Similar effects can be observed in personality or attitude measures. The questions asked of respondents in some questionnaires may cause the respondent to modify their responses in response to the apparent agenda of the questionnaire. These examples can cause latent trait levels (θ) to change during the test administration where local independence implies θ remains constant and unaffected by the test itself. While unidimensionality and

local independence are related concepts neither alone is sufficient to imply the other (Sijtsma & Molenaar, 2002).

The third and final assumption of the MHM states that for each item the probability of a response to the item $P_i(\theta)$ is a monotonically non-decreasing function of the underlying latent trait θ . This is expressed by Equation 2.4.

For any pair of randomly selected values θ_a and θ_b , with $\theta_a < \theta_b$,

$$P_i(\theta_a) \leq P_i(\theta_b) \quad (2.4)$$

Equation 2.4 implies that the IRF is a nondecreasing function of θ and implies that the greater the respondent's ability the greater the probability of correctly responding to an item measuring that ability. The IRF can take any form; they can be logistic functions, partly linear or exponential. Any form is permitted so long as the IRFs are nondecreasing functions of the latent trait.

Once the assumptions of MHM are met the scores on all items should increase as the score on the underlying latent trait increases. This enables the ordering of respondents on the latent construct by the sum of their item responses. This is often assumed of scales but it is important to test this explicitly. As Mokken scaling does not permit a numerical estimation of θ the ordering on the scale of θ is instead ordered by the true score (T) from classical test theory.

The second Mokken model, the double monotonicity model (DMM) is characterised by the same assumptions of the MHM and adds the assumption of non-intersection of IRFs

across the latent trait. This non-intersection concerns the IRFs and maintains that IRFs can be ranked and numbered such that:

$$P_1(\theta) \leq P_2(\theta) \leq P_3(\theta) \leq \dots \leq P_k(\theta), \text{ across all } \theta \quad (2.5)$$

Equation 2.5 demonstrates that the IRFs do not intersect and that the first item is the most *difficult* followed by the second item and so on for all values of the latent trait allowing for ties in the ranking by *difficulty*. Practically speaking, the non-intersection of IRFs means that the ordering of items by mean scores is the same for all values of the latent trait, with the exception of the possibility of ties. This feature is known as invariant item ordering (IIO). This means the item ordering by means is the same, with the exception of potential tied scores, for every value of θ (Sijtsma & Molenaar, 2002). The DMM allows for an IIO in terms of *difficulty* and can be considered as the ordinal form of the Rasch model (van Schuur, 2003).

Establishing IIO is imperative for confirming hierarchical scales. It is often desirable that items within a scale follow the same order of *difficulty* for all respondents. These hierarchical scales simplify the interpretation and comparability of responses and respondents as the *difficulty* of items is the same for all respondents irrespective of ability level (Sijtsma & Junker, 1996). A set of k dichotomously scored items meet criteria for invariant item ordering if the items can be ordered and numbered such that:

$$P_1(\theta) \leq P_2(\theta) \leq \dots \leq P_k(\theta), \text{ for each } \theta$$

IIO concerns the *ordering* of the items, not the mean values themselves; these values may differ across subgroups, as they are dependent on the distribution of θ . The exact value of any P_i may not be the same across subgroups. For example, P_1 may be 0.4 for the men and 0.6 for women but under IIO assumptions Item 1 is the consistently the most *difficult* item in relation to the rest of the items in the test; $P_2 \geq 0.4$ for men and $P_2 \geq 0.6$ for women. This illustrates the important concept of IIO; that the *ordering* is consistent across all subgroups of the population. It should be noted that as Mokken scaling is a nonparametric model the definition of item *difficulty* for logistic models such as the 1PLM is not appropriate. Rather in the context of Mokken scaling analysis item *difficulty* is based on ordering typically by mean item scores.

IIO means that the ordering is invariant across all subgroups from the population of interest. This is an important property which allows the population of interest to be divided into separate groups, for example male and female or high and low ability groups, and the item ordering of the population as a whole is the same as the ordering in each of the subgroups of the population.

As a probabilistic model the expectation of ordering by item *difficulty* is based on the likelihood of response. For example, the probability of a correct response to a low *difficulty* item is close to, but not equal to 1 for an individual with a high level of ability, likewise for an individual with a low level of ability the probability of a correct response to a *difficult* item is close to but not equal to 0. From this example it is clear that the probability of response to each item is based on both the *difficulty* of the item and the respondent's level of ability.

The Mokken scaling procedure often starts with an assessment of the scalability of items and of the scale itself. These scalability coefficients of both the items within and also of the scale as a whole define Mokken scales (Mokken, 1971). These coefficients play an

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

important role in the assessment of and construction of Mokken scales. These coefficients were first applied by Loevinger (Loevinger, 1947) to determine the homogeneity of a group of items. There are scalability coefficients for each item pair in the scale (H_{ij}), for each individual item (H_i) and a coefficient for the scale as a whole (H).

The item-pair scalability coefficient (H_{ij}) describes the ratio of the covariance between the item pair, and the maximum covariance between both items, given the marginal distributions of the items. The item scalability coefficients of all items in the same Mokken scale must be positive. Given that the variance of scores on item i and item j are both positive then H_{ij} is the normed covariance between the scores on the items:

$$H_{ij} = \frac{\text{COV}(X_i, X_j)}{\text{COV max}(X_i, X_j)} \quad (2.6)$$

Each item has its own scalability coefficient (H_i). This coefficient is a measure of the item's *discrimination* (i.e. how well it can differentiate between high and low attribute respondents). H_i reflects the strength of the association between an item and the other items within the scale making it comparable to a regression coefficient in a regression model (van der Ark, Croon & Sijtsma, 2008). A high H_i value implies the item fits well with the rest of the items within the scale and can discriminate between respondents with low levels of latent trait and those with high levels of latent trait as assessed by the whole scale (Embretson & Reise, 2000). Items with low H_i values have less sensitive response probabilities; these items will not detect change or differences in trait level as effectively as high H_i items (Embretson & Reise, 2000).

$$H_i = \frac{\sum_{j \neq i} \text{COV}(X_i, X_j)}{\sum_{j \neq i} \text{COV max}(X_i, X_j)} \quad (2.7)$$

The scalability coefficient (H), the weighted mean of item coefficients, for the scale as a whole expresses the accuracy by which the items in the scale are able to order the participants (Mokken, Lewis & Sijtsma, 1986). This coefficient is derived from the aggregated item scalability coefficients of the items in the scale.

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{COV}(X_i, X_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{COV}_{\max}(X_i, X_j)} \quad (2.8)$$

Given the assumptions of the MHM all scalability coefficients must take values of between 0 and 1 with higher values indicating greater scalability. If $H=1$ this signifies that the items within the scale are perfectly ordered with no disordering of item responses, whereas H of 0 indicates the absence of any linear relationship between the items. The following general rule of thumb has been proposed for the interpretation of H : Scales with H values less than 0.3 are not considered as unidimensional. Scales with H values between 0.3 and 0.4 are considered as meeting the assumption of unidimensionality but of weak strength in regards to item scaling. Scales with H values between 0.4 and 0.5 are considered to be of medium strength and scales with H greater than 0.5 are considered strong (Sijtsma & Molenaar, 2002). Generally speaking the higher the H value the steeper the slopes of the IRFs, which reflects the high *discrimination* values of the items within the scale.

Mokken scaling partitions items into scales on the basis of their scalability coefficients. An item can be included in a Mokken scale once two conditions are met: (1) that the scalability coefficient for each item pair (H_{ij}) is positive, as negative covariances violate the assumption of the MHM (Sijtsma & Molenaar, 2002), and (2) the item's scalability coefficient (H_i) is greater than some a priori determined criterion c (typically, and default in most Mokken scaling packages, 0.3). This criterion level arises from the need to form scales

comprising items with sufficient *discrimination*. Prior to removing items based on their scalability coefficients IRFs should be inspected. IRFs can help to determine the potential cause of a low H_i value. The IRF may be almost flat, irregular with unexpected peaks across θ or an IRF with a single peak (Sijtsma & Molenarr, 2002).

In Guttman scaling the differences in the item *difficulties* must be sufficiently large to enable a clear and distinct correct/incorrect point to be established for each participant (Fisher & Fisher, 1993). This results in a scale with poor sensitivity to small differences in ability between respondents or changes within respondents (Finch, Kane and Philip, 1994). In Mokken scaling however this likely cause for reduced measurement accuracy is diminished, as good item *discrimination* is necessary for inclusion in a Mokken scale. Also Mokken scaling's stochastic nature makes it less restrictive therefore it is more likely to include a greater number of items in the Mokken scale.

2.4.2 Exploratory versus confirmatory Mokken scaling analysis

Mokken scaling analysis can be applied to both test and explore the dimensionality and scalability of scales. Exploratory Mokken scaling can be used to reveal unidimensional clusters of items from a larger collection of items without establishing a particular dimensional structure a priori. Confirmatory Mokken scaling can be applied to determine whether a set of predefined items are unidimensional. Both variations use the scalability coefficients to determine whether the items satisfy the criteria for inclusion in a Mokken scale.

The exploratory method analyses a given set of items to determine whether they conform to one or more scales. In exploratory Mokken scaling an automated item selection procedure (AISP) is used to partition items into scales, or groups of related items measuring a common latent trait, using a hierarchical clustering algorithm. The AISP is a bottom-up

sequential item selection method which is based on inter-item covariances and the strength of the association between the items and the latent trait. The process begins with the selection of the pair of items with the highest positive item-pair scalability coefficient (H_{ij}). This selection procedure proceeds until no additional items meet this criterion. From any items remaining unselected a new scale can be formed in the same way. Any items remaining outwith a scale are deemed unscalable (Sijtsma & Molenaar, 2002). The scalability coefficients of each scale and the items and item pairs within are calculated and assessed with regards to the criteria for inclusion in Mokken scales.

Confirmatory Mokken analysis begins with evaluating a set of J items with regards to their scalability coefficients. The scale is considered an a priori scale with Mokken scaling analysis applied to assess the scale properties as it stands without the objective of removing items. Scalability coefficients are calculated for the scale, each of its items and all item pairs to determine whether they conform to the requirements of a Mokken scale.

2.4.3 Mokken scaling analysis of dichotomous items versus polytomous items

Although originally conceived to analyse dichotomous items Mokken scaling analysis has been developed to handle polytomous items (Sijtsma, Debets & Molenaar, 1990; Hemker & Sijtsma, 1995). The analysis of non-intersection changes from considering the IRF for the dichotomous item to the responses to each embedded level within the item. The relationship between the responses to these levels and the score on the latent construct is described by the item step response function (ISRF). The IRSF models the relationship between the responses of each step in the scale and the latent trait.

Although the level of analysis is different for polytomous items the principles and procedure for determining whether items conform to a Mokken scale is similar to that

concerning dichotomous items. Establishing IIO however marks a fundamental difference between the treatment of dichotomous items and polytomous items.

The DMM assumption of non-intersection is not sufficient to confirm IIO of polytomous items (Sijtsma, Meijer & van der Ark, 2011). Unlike in the case of dichotomous items meeting the assumptions of the DMM does not guarantee polytomous items are invariantly ordered by *difficulty*.

IIO can be determined using a method developed by Ligtvoet, van der Ark, te Marvelde and Sijtsma (2010). Method manifest IIO and coefficient H^T can be applied to the investigation of IIO. H^T is derived from H computed on the transposed data matrix (van der Ark, 2012). Ligtvoet et al., (2010) propose the generalisability of the rule of thumb concerning the interpretation of H to the interpretation of H^T . Once IIO has been established the following heuristic rule can be applied; H^T values below 0.3 means the item ordering is too inaccurate to be of practical use, H^T between 0.3 and 0.4 means the item ordering is of low accuracy, H^T between 0.4 and 0.5 indicates medium accuracy, and values of H^T greater than 0.5 mean high accuracy of item ordering.

2.4.4 Mokken scaling analysis versus Rasch analysis

Much empirical IRT research over the last few years has focused on Mokken scaling and the Rasch model. Although IRT analyses typically just apply one model which limits the degree of comparison between Rasch and Mokken analyses which is possible, theoretical and empirical differences between the two have been studied (Meijer, Sijtsma & Smid, 1990). Both models are unidimensional and cumulative in nature; both assume the presence of a single latent trait underlying the item responses and that the probability of item response is a non-decreasing function of that latent trait value. Where the models differ is in regard to the

assumptions made regarding the form of the response functions of both respondents and items.

The Rasch model assumes the slopes of IRFs are all the same. This strong assumption, in a regression analysis framework, would imply that the correlation between the score on an item and the latent trait is the same across all items. Mokken scaling analysis allows for items to have different slopes, or *discrimination*. The models also differ in terms of the estimates of IRFs: Rasch modelling allows for the numerical, directly observable estimation of θ whereas Mokken scaling relies on the total summed score of all scale items minus item i under assessment ($k-1$ items) in its estimation of θ . The score for $k-1$ items is referred to as the restscore ($R(i)$). Under the Mokken scaling framework IRFs are estimated by firstly defining respondents by their restscore and secondly by calculating the proportion of respondents with a given restscore who obtain a correct score on item i (Sijtsma & Molenaar, 2002).

Rasch modelling rests upon the same assumptions of the MHM with the additional assumption of minimal sufficiency of the unweighted person and item summed score for the approximation of θ and δ parameters (Meijer, Sijtsma & Smid, 1990). For these four assumptions to be met there must be no intersecting IRFs. Scales meeting IIO assumptions can be interpreted as nonparametric equivalents to scales derived using parametric Rasch analysis (Stochl, Jones & Croudace, 2012). Therefore the Rasch model can be considered a special case of the Mokken double monotonicity model (DMM), albeit more restrictive. That fewer items conform to the Rasch model than the MHM and DMM reflects the restrictive assumptions underlying the model. Mokken's greater degree of flexibility compared to Rasch marks a significant difference between the two models as it is more likely to represent a framework in which the data fits.

2.5 IRT applications in assessment and quantification of dementia progression

Balsis et al. (2012) suggest IRT can provide the level of measurement necessary to improve the current standards of measurement in dementia and have found that precision is gained from this level of analysis (Balsis, Unger, Bengt, Geraci & Doody, 2012; Bengt, Balsis, Geraci, Massman & Doody, 2009). IRT comprises various psychometric models allowing for the development and improvement of psychometric measures by calculating the parameters of a mathematical function responsible for the association between the underlying latent trait and the responses to the items. In this way IRT provides the statistical framework necessary to move away from relying on the summed scale score to considering the unique contributions of specific items to the measurement of cognitive impairment in dementia.

Aggregate scoring methods using traditional measurement models, although common, involve two large theoretical limitations: firstly the absence of an explicit and ordered continuum of items that represent a unidimensional trait; and also the absence of additivity of ordinal raw scores (Merbitz, Morris & Grip, 1989; Fisher, 1993). Supporters of IRT methods stress the significance of an ordered continuum along which to represent the trait being measured (Hambleton et al., 1991). Contrary to summative scoring techniques IRT models meet the conceptual requirements of order and additivity (McHorney, Hayley & Ware, 1997).

IRT methods enable us to assess dementia along a continuous spectrum rather than the categorical approach of classical test theory. Dementia can take years to develop, which allows the opportunity to implement intervention or prevention strategies. The success of early intervention relies on the early detection and identification of the pre-clinical

symptoms. Considering this prodromal stage on a continuum of dementia would assist the development of such strategies.

One practical application of IRT methods is in the evaluation of existing measures of cognitive and functional abilities assessment for their suitability in accurately measuring dementia progression. IRT models can provide information on the specific contribution of each item within a scale shedding light on what they reveal about the underlying construct. Importantly items within a scale are not equal, neither in terms of their perceived *difficulty* nor in the strength of their association with the latent trait. IRT techniques allow researchers and clinicians to more reliably determine each patient's actual degree of cognitive or functional impairment by examining item properties such as *discrimination* and *difficulty*. These properties are graphically represented in the IRF (Hambleton & Swaminathan 1985).

Items within a scale will differ in how well they can differentiate between patients with relatively intact cognitive abilities and those whose cognitive ability has deteriorated to moderate or severe levels of impairment. This is indexed by the item's *discrimination* value. *Discrimination* is reflected by the slope of the IRF with IRFs with steeper slopes being more discriminating than those with flatter curves. This information can reveal which items in a scale are contributing well to accurately detecting small changes within and differences between patients and which items are not. Such items are likely to be responded to in a similar manner by patients of different levels of cognitive dysfunction making their function largely redundant. Determining item *discrimination* also reveals the degree of association between items and the construct of cognitive impairment.

2.5.1 Establishing hierarchical scales

Item *difficulty* refers to the ease with which the item is correctly responded to or endorsed by the respondent. This information can be applied to form hierarchical scales based on item *difficulty*. That items will vary in the ease with which they are endorsed forms the basis for forming hierarchical scales. From an array of items differing in *difficulty* level more respondents will endorse the less *difficult* items than the more *difficult* items. Analyses of item *difficulty* can be applied to establish hierarchies of decline. Hierarchies of items are established based on the relative *difficulty* of items and the degree to which item pairs are consistently ordered by this *difficulty*.

An analogy of climbing a staircase can be used to illustrate the properties of a hierarchical scale, with each step representing an item in a scale and the staircase representing the level of latent trait, it follows that you cannot reach the ninth step without having previously climbed the eight steps below; and by having climbed to the ninth step you will not have reached any step above this level.

The ability to quantify dementia progression would provide an objective record of the course and sequence of decline. Confirming a hierarchy of items with regards to *difficulty* adds another facet to a scale's application other than just using summed total scores. Once a hierarchy has been established the items can be ordered relative to each other with all items ordered along the underlying latent trait in question. This means that all changes in response have direct and meaningful significance-each of the steps in the hierarchy represents a specific performance pattern. Hierarchical scales allow the accurate specification of distinct levels of impairment in a way that cannot be achieved with summed scores (Morris, Fries & Morris 1999).

A further dimension is added if it can be confirmed that the item ordering is consistent across all subgroups of the population on interest. Invariant item ordering (IIO) is an

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

important characteristic of hierarchical scales and greatly enhances the interpretive power and applicability of results (Sijtsma & Hemker, 1998). Confirming that the item ordering by *difficulty* holds for individual respondents, as well as for the population as a whole, lends considerable credibility to a measure. This property confers practical significance in the assessment of cognitive and functional ability in dementia. If we consider items within an IIO scale as indicators or symptoms of cognitive dysfunction then a patient with a higher total score on this scale has the same symptoms plus additional symptoms of a greater degree of impairment.

IIO is important when comparing different respondents or groups and has many significant practical applications. A formal hierarchy of item *difficulty* has the power to improve construct validity, for example, by supporting or contradicting the belief that division is a more *difficult* calculation than addition (Chiu, Fritz, Light & Velozo, 2006). Hierarchical scales can be used to test theories about the construct under assessment. For example, it may be hypothesised that cognitive abilities decline along a specific trajectory in a sequence of stages as dementia progresses. Hierarchical scales can identify which cognitive tasks are associated with different degrees of impairment or stage of dementia. This insight would make it possible to develop a test comprised of items designed to assess the specific stages of decline. With such a test we would expect the ordering of items by *difficulty* to parallel the stages or path of decline and be the same for all patients.

Hierarchically arranging scales provides interesting implications for both researchers and clinicians. Responses to any given item in a hierarchical scale in isolation can provide information on the respondent's level of latent trait. Hierarchical scales can provide more focused information on the trait in question than the usual summation of trait level afforded by total scores. Hierarchies can provide prognostic value to clinicians assessing patients. A respondent who responds correctly to a high *difficulty* item on the hierarchical scale is likely

to respond correctly to all of the previous less *difficult* items and similarly a respondent unable to correctly respond to a moderately *difficult* item is unlikely to be able to respond correctly to any more *difficult* item in the scale. In cases such as this where the individual's prior responses predict that they would be unable to successfully continue beyond a certain level of *difficulty* proceeding with the test administration may cause undue anxiety for the participant. Continuing testing would be unlikely to provide any valuable information and therefore test administration could be tailored to particular levels of ability with the use of hierarchical scales. In this way information is provided from the individual responses and not just from the total score enabling quicker estimations of a participant's level of ability and the power to adapt the test administration on an individual level. This kind of adaptive testing is valuable in cognitive assessment as it reduces testing time and stress and burden on patients. Hierarchical scales facilitate adaptive testing whereby only a selection of items, either from the more *difficult* or the less *difficult* range of the scale depending on the ability of the specific patient, is required for testing (van der Lee, Roorda, Beckerman, Lankhorst & Bouter, 2002).

Hierarchies of *difficulty* can also help to identify specific patterns of cognitive decline for different forms of dementia and can detect abnormal patterns or sequences of decline, which should be investigated further. Deviations from the typical hierarchical trajectory can be detected and investigated if necessary (Daltroy, Logigian, Iverson & Lian, 1992). The identification of distinct patterns of *difficulty* may help to delineate different pathological causes and manifestations of dementia.

2.5.2 Scale development and evaluation

Examining the *difficulty* levels of items can help to determine whether there are any gaps in a scale's coverage of the trait continuum (i.e. a scale with very low *difficulty* items may be

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

efficient in the measurement of patients with severe dementia but may not assess or detect any impairment in patients with mild or prodromal dementia). Likewise scales can be created to be specific to the assessment of any given particular range of impairment. In this way IRT can be used to specifically tailor tests to patients (van der Lee et al., 2002).

A direct application of IRT is in the development and refinement of existing scales of cognitive and functional impairment for use in evaluating dementia progression (Mungas & Reed, 2000). Optimal measures of cognitive and functional ability in dementia require inclusive coverage of the breadth of ability. Characterising items in terms of *difficulty* and *discrimination* can help in the refinement or development of new scales, which ideally would consist of items of high *discrimination* with different levels of *difficulty* to ensure changes and differences at all levels of ability or latent trait will be detected (Mungas & Reed, 2000).

This added information would provide more insight into a patient's degree of cognitive impairment than could be interpreted from the sum of their responses. Examining these item differences may offer a more accurate estimation of cognitive dysfunction. IRT methods afford us the opportunity to investigate and take these item differences into account. This can have significant clinical implications and may contribute to the ability to accurately anticipate and identify differences between and changes within respondents across the natural trajectory of the disease or over the course of a treatment intervention or clinical trial. With outcomes of clinical drug trials assessed by participants' responses to cognitive measures it is essential that any changes in cognition are reliably quantified.

Scales developed using parametric models can be used to compare the results from different tests where the items have been drawn from the same item bank. This practice is referred to as equating. Equating is applied in adaptive testing where a measure is tailored to a specific individual's ability level by selecting items from the item bank in successive

stages. Nonparametric IRT models have also been used for equating and adaptive testing (Tellegen & Laros, 1993; Huisman & Molenaar, 2001).

2.6 Summary

In summary, the most commonly used practice of scoring cognitive and functional assessment tools in dementia is summing raw scores to obtain a total score. This method's popularity is most likely related to its simplicity and ease of understanding. However research suggests that solely relying on this aggregated method may yield misleading information about the underlying degree of cognitive or functional impairment, which has significant impact on the ability to accurately interpret differences between individuals in cross-sectional studies and changes within individuals in longitudinal research and clinical trials (Balsis et al., 2012).

IRT methods have been widely applied to overcome the limitations of classical test theory measurement (Reise & Waller, 2009). Moving beyond merely examining the number of items scored correctly or incorrectly IRT methods can examine which items a respondent gets correct or incorrect and what the response to specific items and response patterns can tell us about a respondent's cognitive or functional impairment. From an IRT perspective cognitive and functional decline in dementia can be conceptualized along a spectrum of cognitive and related functional tasks declining at different rates. Cognitive and functional scales as interpreted using IRT methods could become a powerful tool in the diagnosis, assessment of progression and barometer of the outcomes of dementia.

Mokken scaling analysis uses a modelling approach which is less restrictive than that of parametric IRT. It can be used to enhance the interpretive power of cognitive and

CHAPTER 2: IRT METHODS IN DEMENTIA ASSESSMENT

functional scales by examining item properties of individual items such as *difficulty* and *discrimination*, reducing test burden on patients through the elimination of unnecessary or redundant items. In addition it can be used to establish hierarchical scales and to test whether the hierarchical structure is invariant across all subpopulations.

Chapter 3: Item response theory analysis of cognitive tests in people with dementia: a systematic review

Work presented in the following chapter is taken from the following paper:

McGrory, S., Doherty, J. M., Austin, E. J., Starr, J. M., & Shenkin, S. D. (2014). Item response theory analysis of cognitive tests in people with dementia: a systematic review. *BMC Psychiatry, 14*(1), 47.

3.1 Introduction

Global cognitive functioning measures are the mainstay diagnostic tool for dementia, in conjunction with determination of functional decline, and are also used to track and measure disease course. Measures of cognition in dementia should be able to both reliably detect the disease in its early stages and to evaluate the severity of the disease (Mungas & Reed, 2000). The most common method of scoring a cognitive test is to sum the raw score. The total score is used to aid diagnosis and to assess and monitor disease severity. This method is quick and simple to apply and is based on the premise of all test items reflecting a common unobservable trait or ability range along which cognitive impairment can be measured (Wouters, van Gool, Schmand & Lindeboom, 2008).

However the simple summation of raw scores overlooks any differences between the items and information the pattern of response can provide. It may therefore lead to an inaccurate estimation of cognitive impairment (Wouters et al., 2008).

CHAPTER 3: SYTEMATIC REVIEW

Items within a cognitive assessment scale will differ in several ways. Firstly some items may be more *difficult* than others, for example, for most people, repeating a noun would be less *difficult* than remembering a phrase or list of words. Secondly, some items may be more sensitive to the early stages of cognitive decline and others to the later stages of the disease. Thirdly, items may differ in how sensitive they are to clinical change. Finally, some items may be redundant and provide no meaningful variability to the measure. These items could be removed to ease the burden on patients and clinicians.

The same total score can be achieved via many different patterns of response. For example, two individuals scoring 20 on the MMSE may have correctly and incorrectly answered completely different items. Likewise an individual obtaining the same total score before and after treatment would be considered as having experienced no change in cognitive impairment even if the pattern of response across the items had changed. Therefore, there is a need to look beyond the total score and to investigate the pattern of response to the individual items. This level of analysis is permitted using IRT methods.

In addition to providing item parameters of *discrimination* and *difficulty* IRT also permits the examination of the performance of the overall scale using the Test Characteristic Curve (TCC). The TCC is a valuable tool for assessing the range of measurement and the degree of *discrimination* at various points along the ability continuum. Also the extent to which the TCC is linear illustrates the degree to which the scale provides interval scale or linear measurement.

IRT can calculate item *information* for all trait levels which can be used to plot an item information curve (Fraley, Waller & Brennan, 2000). *Information* is the equivalent of variance explained, showing how effectively a measure captures the latent trait. *Information*

CHAPTER 3: SYTEMATIC REVIEW

can be calculated for each ability level. The greater the amount of *information*, the more precision with which the ability can be estimated.

With regards to the suitability of IRT in the assessment of cognitive assessment measure in dementia this level of analysis could potentially improve the tests used for diagnosing and monitoring people with dementia. By determining the *difficulty* of items within a scale it is possible to develop a hierarchy of item *difficulty* i.e. a list of questions from those with lowest *difficulty* (where the expected probability of a correct answer of 50% is reached at a low overall score) to those with highest *difficulty* (where the expected probability of a correct response of 50% is reached at a high score). This confirms the sequence of cognitive decline. Establishing a hierarchy of *difficulty* confirming the sequence of decline will allow clinicians and researchers to identify any deviations in the rate or sequence of cognitive decline from the usual trajectory of loss. Hierarchies of item *difficulty* may differ according to diagnosis or by country/region or by different translations of measures. Identifying unique sequences of cognitive decline for different forms of dementia could aid in diagnoses. Additionally being aware of the ordering of *difficulty* makes it possible for clinicians to tailor their assessments according to severity level, e.g., selecting less *difficult* items for patients with established dementia and the more *difficult* items for healthy elderly or those with mild or early stages of cognitive impairment (Wouters, Zwiderman, van Gool, Schmand & Lindboom, 2009).

IRT can also examine the sensitivities of the items within a measure. By examining the slope of the ICC the items *discrimination* can be assessed. The range of cognitive impairment at which the slope is the steepest is where that item will be maximally *discriminative*, differentiating well between various gradations of impairment and providing increased sensitivity to change. Determining the *discrimination* of items can reveal which items are most likely to expose changes in cognition and those with weaker *discriminatory*

CHAPTER 3: SYTEMATIC REVIEW

power that are unresponsive to such changes (Weiss, Fried & Brandeen-Roche, 2007; Sijtsma, Emons, Bouwmesster, Nyklicek & Roorda, 2008). Looking at the item curves in relation to each other provides useful information on the breadth of measurement of an instrument. IRT can also identify key items which provide valuable information or whether any items within the scale are redundant, i.e. items with similar ICCs.

Applying IRT techniques to measures of cognitive functioning in dementia could have far reaching implications for clinicians and researchers leading to advancements in screening assessments and diagnosis, the charting of disease course and the measurement of change with disease progression and in response to treatment. In addition, IRT methodology will be useful to industry in the design of psychometric tests. IRT has been used to analyse clinical measures in several different fields: schizophrenia (Santor, Ascher-Svanum, Lindenmayer, Obenchain, 2007), depression (Aggen, Neale & Kendler, 2005), attachment (Fraleay, Waller, Brennan, 2000), social inhibition (Emons, Meijer & Denollet, 2007) and quality of life (Hill et al., 2007). IRT has also been used to examine ADL and Instrumental Activities of Daily Living (IADL) scales (Fieo, Watson, Deary & Starr, 2010; Chan, Kasper, Brandt & Pezzon, 2012). IRT methods have been successful in improving functional scales by establishing interval level measurement (Spector & Fleishman, 1998); hierarchies of item *difficulty* (Fieo et al., 2010; Jette et al., 1998; Sheehan, DeChello, Garcia, Fifield, Rothfield & Reisine, 2002); *discrimination* of items (Fieo et al., 2010; McHorney & Cohen, 2000); as well as identifying ways of increasing measurement precision (Spector & Fleishman, 1998). IRT analyses of measures of cognitive functioning in the general population have been described (Kecukdeveci, Kutlay, Elhan & Tannant, 2005; Zheng et al., 2012), including several papers with samples including some participants with dementia (Lindeboom, Schmand, Holman, de Haan & Vermeulen, 2004; Prieto, Delgado, Perea & Ladera, 2011; Ideno, Takayama, Hayashi, Takagi & Sugai, 2012; Teresi, Golden, Cross, Gurland, Kleinman

& Wilder, 1995; Wouters, van Gool, Schmand, Zwinderman & Lindeboom, 2010). However, despite the strong theoretical basis outlined above for using IRT in people with dementia, there is limited published data. Therefore we performed a systematic review of the published studies that use IRT to revise or develop instruments assessing cognitive ability in people with dementia.

3.2 Method

3.2.1 Search Strategy

Published studies were identified through searches of Medline (including work in progress from 1946 until 5th September 2013), Embase (1980 until 5th September 2013), PsychInfo (1806 until 5th September 2013) and CINAHL (1981 until 5th September 2013). Search filters included were keyword, title and abstract information. Search terms relating to IRT and dementia were combined. Articles with any combination of any of the IRT terms and any dementia term were reviewed. For full search strategy see Appendix B. References of included studies were hand-searched and a forward citation search was performed on all included studies to establish all articles which cited them.

3.2.2 Data Extraction

A total of 384 articles were identified from this search. After duplicates were removed the titles and abstracts of 203 articles were screened by two independent researchers. One hundred and sixty articles were excluded on review of title and/or abstract (for example, non IRT methods, IRT analyses of functional or other non-cognitive assessments). Forty three articles considered to be relevant were retrieved and assessed for agreement with the following inclusion and exclusion criteria. Data were extracted from original studies onto forms which were refined following piloting. Figure 3.1 shows the flow chart for this review.

3.2.3 Inclusion/ exclusion criteria

This review aimed to include all published studies that applied item response theory methods to instruments with face validity for measuring global cognitive impairment in dementia. The initial search did not restrict results to those published in the English language.

Exclusion criteria were as follows: (i) unpublished studies, dissertations, theses, journal conference abstracts and poster presentations; (ii) studies using proxy reports as there is evidence of discrepancy between self-report and informant measures of cognitive functioning (DeBettignies, Mahurin & Pirozzolo, 1990); (iii) studies with participants without diagnosed dementia; (iv) studies without details of dementia diagnosis criteria or percentages of participants with dementia; (v) studies reporting IRT applications to domain specific measures of cognition rather than global cognitive functioning, for example the Boston Naming Test (Kaplan, Goodglass & Weintraub, 1983) used to measure confrontational word retrieval; (vi) studies that did not provide information on item level performance or overall test performance; (vii) studies examining non-cognitive scales, although studies which reviewed a range of outcomes had the results from the cognitive scales included; (viii) no language restrictions were made in the search, but non-English language articles were not included in the final review as they used non-English scales; (ix) use of Guttman scaling procedures (Guttman, 1950).

While studies have found increased sensitivity of domain specific neuropsychological tests to early impairment than test of global cognition (Harrison, 2007) this review chose to restrict its focus to IRT analyses of global cognitive instruments to increase clinical relevance as these are the most commonly used for testing in routine practice.

The decision to exclude Guttman scaling was based on the considerable evidence stating the inferiority of these methods in comparison to the more advanced item response

CHAPTER 3: SYTEMATIC REVIEW

methods (Kempen, Myers & Powell, 1995). The method was included in the search strategy; however, as some studies may have applied another method of analysis without indexing it and the exclusion of this term may have led to some relevant studies being overlooked.

Non-English language versions of cognitive measures were excluded. While several measures, most notably the MMSE (Folstein, Folstein & McHugh, 1975), have been translated into many languages for use in different countries and cultures there are concerns over the cross-cultural validity. The language in which a test is administered can affect performance leading to a potential overestimation of cognitive impairment in individuals who do not speak English (Salmon, Riekkinen, Katzman, Zhang, Jin & Yu, 1989; Escobar, Burnam, Karno, Forsythe, Landsverk & Golding, 1986; Hohl, Grundman, Salmon, Thomas, Thal, 1999). Differential Item Functioning (DIF) (Holland & Wainer, 1993) can be applied to examine the effect of language bias of items and tests administered in different languages. For example, if patients of equal cognitive ability tested in English and Spanish have unequal probabilities of responding correctly to a particular item on a cognitive assessment, then the item functions differently with respect to language. The effect of different test languages of cognitive assessments has been examined in this way (Teresi et al., 1995; Edelen, Thissen, Teresi, Kleinman & Ocepek-Welikson, 2006; Teresi et al., 2000; Morales, Flowers, Gutierrez, Kleinman & Teresi, 2006; Crane, Gibbons, Jolley & van Belle, 2006; Marshall, Mungas, Weldon, Reed & Hann, 1997). However these studies did not examine DIF in dementia populations and were therefore not included. Also the non-English language versions administered makes comparison with scales in English problematic because the semantic range of items cannot be assumed in translation (van de Vijer & Hambleton, 1996), for example, repeating “No ifs, ands, or buts” corresponds to repeating “We put ones’ efforts all together and pull the rope” in the Japanese version of the MMSE (Ideno, Takayama, Hayashi, Takagi & Sugai, 2012) and to a tongue-twisting phrase “en un trigal habia tres

CHAPTER 3: SYSTEMATIC REVIEW

tigres” (“there were three tigers in a wheat field”) in the Spanish version (Prieto, Contador, Tapias-Merino, Mitchell & Bermejo-Parejo, 2012). To avoid any potential confounding these articles were not included for full review (Marshall, Mungas, Weldon, Reed & Haan, 1997). The decision to exclude articles using non-English language assessments has no implications for the validity of cognitive testing in other languages.

CHAPTER 3: SYTEMATIC REVIEW

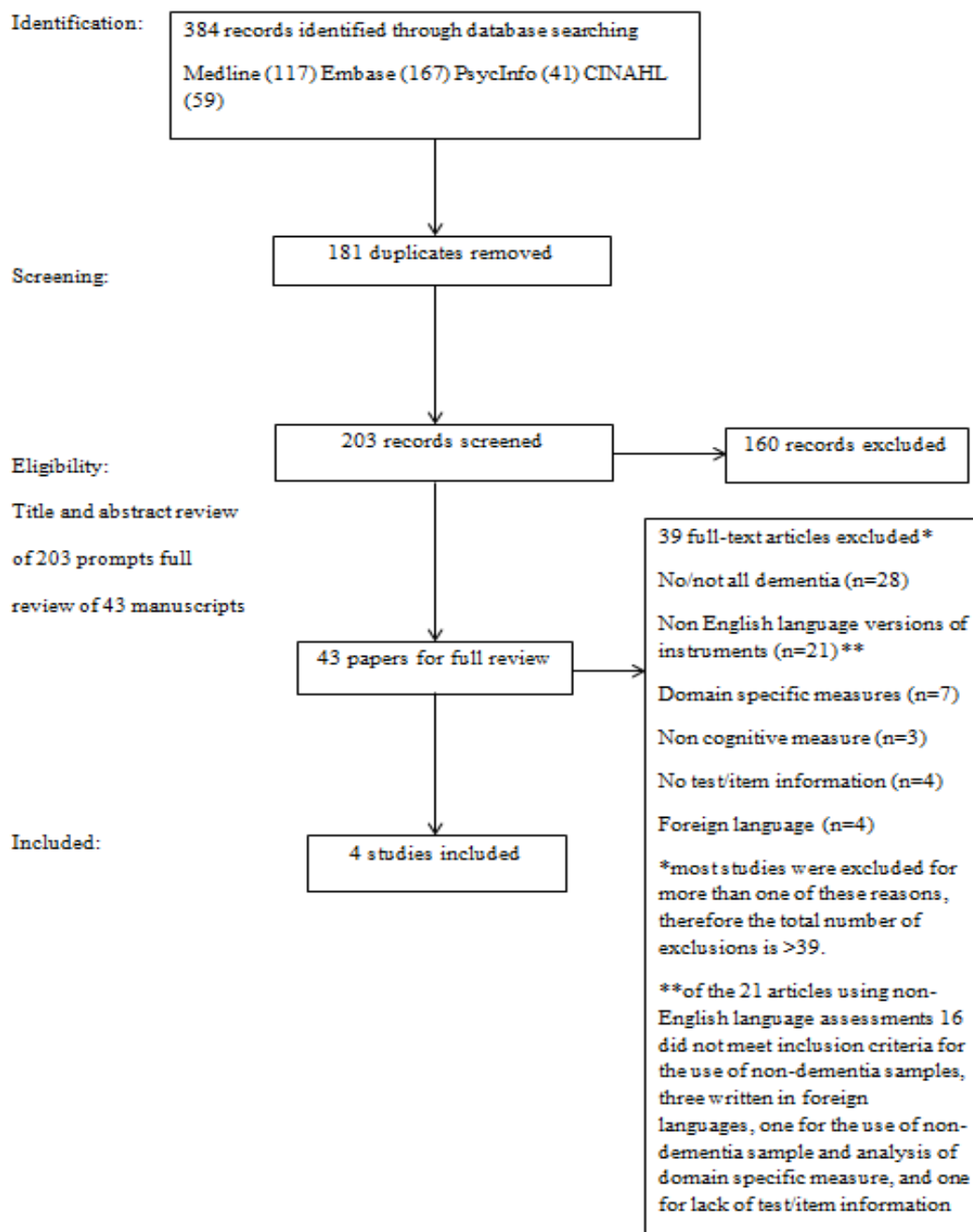


Figure 3.1 Flow diagram for review selection

3.3 Results

Four cross-sectional studies met inclusion criteria, including 2,920 patients from six centers in two countries: Table 3.1 describes the characteristics of the studies reviewed. In total dementia aetiologies comprise 74.1% (2165) probable Alzheimer's disease (AD), 9.3% (273) possible AD, 2% (60) vascular dementia, 11.1% (325) mixed and other dementia. For individual studies see Table 3.1. Most patients fall within the moderate range of severity of dementia. Three cognitive tests (MMSE, Alzheimer's Disease Assessment Scale-Cognition (ADAS-cog, Rosen, Mohs & Davis, 1984), Blessed Information Memory Concentration Test (BIMCT, Blessed, Roth & Tomlinson, 1968)) and three different IRT methods (Item Characteristic Curve analysis, Samejima's graded model, Two-parameter model) were used.

Ashford, Kolm, Colliver, Bekian and Hsu (1989) applied IRT techniques to identify the degree of AD severity at which individual items of the MMSE are lost and the rate at which they are lost at that level of severity. MMSE scores from 86 AD patients were analysed. Most people had moderately severe AD (mean MMSE score=18).

A hierarchy of item *difficulty* was formed (see Table 3.2). Most *difficult* items were the three memory items and 'orientation to date' (which also tests recent memory), and 'serial sevens'. These findings suggest that the mental functions assumed to underlie performance of these items- memory and attention and calculation- are lost earliest in the progression of AD. Least *difficult* items, i.e. late loss, were 'verbal directions', 'name pencil' and 'repeat nouns'. This pattern is consistent with the typical clinical course of AD starting with memory problems ultimately leading to problems with over-learned associations and early-learned verbal mimicking.

CHAPTER 3: SYTEMATIC REVIEW

For one of the least *difficult* items ‘name pencil’ participants with a score of 6.6 had a 50% probability of getting this item correct. At a score of 10 participants had an almost 100% chance of correctly identifying the pencil. This is in sharp contrast to the most *difficult* items ‘recall nouns’. A participant with a score of 20 had approximately 25% chance of getting ‘recall: tree’ correct. These recall items were answered incorrectly by approximately 83% of the participants.

Item *discrimination* was used as an index of the rate of loss. The most *discriminatory* items were: ‘name pencil’, ‘write sentence’, ‘orientation to month’, ‘name watch’, ‘orientation to date’, ‘orientation to year’, ‘close eyes’. For these items there is a sharp cut-off of ability level at which the item was passed or failed. The items with the lowest *discriminative* power are those items lost earliest; ‘recall: tree’ and recall: flag’, and latest in disease course; ‘verbal directions’. Due to these items assessing abilities which are either lost almost immediately or not until very late stages the rate of loss is not meaningful but the items do serve a useful purpose as they measure ability at either extreme of the MMSE scale. Some limitations of this study include the fact that participants with possible AD were not excluded for sensitivity analysis. Also there was no explicit investigation of unidimensionality of the MMSE. However the item-by item analysis of the variability in AD implies that there is a strong unidimensional component in the course of AD. There was no report of who administered the MMSE to the participants and whether they were blind to diagnoses. This introduces potential for bias.

Mungas and Reed (2000) analysed MMSE and BIMCT scores from 1207 participants. A very broad range of cognitive impairment across the full range of MMSE and BIMCT scores was represented. Here IRT methods were employed to evaluate existing measures and to develop a new global functioning measure by selecting items from the

CHAPTER 3: SYTEMATIC REVIEW

existing scales with *difficulty* ranges spanning the breadth of ability levels to increase *discrimination* at all ability levels.

Items were recoded as dichotomized variables for analysis. Ordinal scale items such as ‘world backwards’ in the MMSE were converted to a number of dichotomous items equal to the maximum score on this item, leading to total scores of 30 for the MMSE, 33 for the BIMCT. Cognitive tests were administered by a neuropsychologist, neuropsychology trainee or a trained psychometrist. The authors did not mention if these individuals were blind to diagnoses.

Test characteristic curves (TCCs) for both scales were generated. TCCs of the MMSE and BIMCT were distinctly non-linear, showing decreased *discrimination* at both ends of the ability continuum with linear measurement for moderate levels of impairment. This indicates relative insensitivity to changes in ability at each end of the ability spectrum.

A more linear brief composite instrument; ‘Global Function’ was created. Items were selected from the MMSE, BIMCT and a functional measure; Blessed-Roth Dementia Rating Scale (BRDRS). Items fitting uniform distribution of *difficulty* across the spectrum of ability measured by the three measures were selected. The new scale showed improved *discrimination* at low ability levels but due to the relative absence of high *difficulty* items in the MMSE, BIMCT and BRDRS the scale showed decreased *discrimination* at high ability levels. This illustrates the need to develop and add more *difficult* items to existing and new measures to decrease ceiling effects. The hierarchy of item *difficulty* of the cognitive items from this measure is provided in Table 3.2. While this measure included functional items which is beyond the scope of this review the most *difficult* items were memory items which is in line with previous findings.

CHAPTER 3: SYTEMATIC REVIEW

Again there was no assessment of whether the items in the tests are sufficiently unidimensional for the use of IRT. It was not reported whether those who tested the participants were involved in the analysis.

Gibbons et al. (2002) used IRT to compare the relative *difficulties* of MMSE items between people with AD living in the US and UK. The 401 US participants were comparatively less impaired (mean MMSE 19.6) than the 139 UK participants (mean MMSE 16.5). There were some differences between items used for the two samples. Orientation to state and county items in US sample were substituted for orientation to county and 2 streets nearby for the UK cohort and the nouns to repeat and remember were also different for the two cohorts. Although these differences limit the direct comparison of *difficulty* between these items as the differences are limited to these items they are unlikely to explain the entire difference observed between the two samples. Reports indicate the interview structures did not differ between samples in any substantial way. For analysis all items which could have a score greater than one were dichotomized. All three nouns must be repeated and all stages of following the verbal directions must be carried out for these items to be scored as correct. 'Recall nouns' was scored correctly if any one of the three nouns were recalled. Two points for 'serial sevens' were sufficient to be scored as correct. Therefore ability level was represented by the score of the 19 dichotomized items, excluding the score of the item under assessment resulting in score ranges from 0-18. Gibbons et al., (2002) established the relative *difficulties* of items for both cohorts, adjusted to an education level of high school or less.

UK results:

The most *difficult* items were 'no ifs, ands or buts' and 'recall nouns'. At the uppermost score of 18 only an estimated 29% of participants could repeat the phrase 'no ifs, ands or buts'.

The easiest items were 'close eyes' and 'name objects'. Here at an estimate of less than zero

CHAPTER 3: SYTEMATIC REVIEW

most participants could still answer correctly so again these estimates were truncated at 0.

This reflects the relative simplicity of these items.

US results:

The most *difficult* items were ‘orientation to date’ and ‘no ifs, ands or buts’. At ability scores of 17.5 and 15.3 half of the participants could correctly identify the date and repeat ‘no ifs, ands or buts’ respectively. The easiest item was ‘repeat nouns’. The ability score was again truncated at 0 indicating that even at this low level of ability most participants were able to answer correctly. ‘Name objects’ and ‘close eyes’ were also relatively easy items.

Hierarchies of item *difficulty* for both UK and US samples are presented in Table 3.2. Five items; ‘no ifs, ands or buts’, ‘recall nouns’, ‘orientation to state/county’, ‘repeat nouns’ and ‘verbal directions’ were significantly more *difficult* for the UK sample. While some items were more *difficult* for the US cohort the differences were not significant. A score of 15.6 was necessary for a UK participant to have a 50% chance of correctly responding to “Verbal directions” in comparison to a US participant having the same probability at a score of seven.

Additional analyses excluding ‘possible’ AD, MMSE items which differed between samples, and accounting for international differences in educational standards did not affect the results.

Attempting to control for the differing levels of severity between the samples, dementia severity (as assessed by the Dementia Rating Scale; DRS) along with age, education and gender were assessed as possible confounders of the relative *difficulty* of items. The relative *difficulty* of the items was not affected by the DRS. It is possible however that controlling for the DRS may not have been enough to compensate for the differences between the two groups.

CHAPTER 3: SYTEMATIC REVIEW

The methodology applied here was rather robust given the additional analyses performed. However the researchers did not explicitly investigate unidimensionality of the instruments. The MMSE was administered at home by trained research interviewers for both cohorts. The scores used were taken from interviews preceding diagnosis which eliminated risk of bias. The diagnoses were not made by the researchers doing the analysis again limiting any potential bias.

Benge, Balsis, Geraci, Massman and Doody (2009) used IRT analyses to examine the measurement properties of the ADAS-cog across the spectrum of cognitive decline in AD. To determine the relationship between the level of impairment and the probability of achieving observed scores on the test as a whole and the test's subscales scores from 1087 AD participants were analysed. 43 patients with mild cognitive impairment (MCI), diagnosed using Petersen (2004) criteria, were included. This is the only study to include MCI participants and although they account for only 4% of the sample it is worth keeping this difference in mind when interpreting the results. The mean ADAS-cog score was 31.2 indicative of moderate to severe dementia.

Benge et al., (2009) assessed the unidimensionality of the ADAS-cog. Results from an exploratory factor analysis and confirmatory factor analysis confirmed the ADAS-cog as a one-factor scale.

The measure's subscales were grouped into three domains: memory, praxis and language for analysis. Curves permitting the comparison of the domain performance across the spectrum of cognitive decline were created. These curves indicate that memory has most *discriminative* power at the relatively milder stages of decline in comparison to language and praxis which were maximally *discriminative* at the same stages later in the disease course.

Analysis of the 11 subscales showed 'word recall' to be the most *discriminative* at mild stages of disease making it the best indicator of mild cognitive decline. 'Recall of

CHAPTER 3: SYTEMATIC REVIEW

instructions' remained relatively unaffected until the later stages of disease. Praxis and language subscale curves indicate that as with the domains, these subscales maximally *discriminate* at moderate levels of decline. The curves for 'ideational praxis', 'construction' and 'word finding', 'speech comprehension', 'commands', 'speech content' and 'naming' overlap considerably implying that they yield more or less the same information about patient's stage of cognitive decline. All items *discriminate* well at moderate levels of severity.

Information analysis found perhaps not surprisingly the highest level of *information* is found at moderate levels of cognitive dysfunction. At this level a unit change in cognitive dysfunction represents a greater change in performance than the same change at either ends of the range. This indicates that the ADAS-cog as a whole has relatively high levels of *discrimination* and can differentiate between various degrees of ability at this moderate stage. This study was the only one to report an assessment of unidimensionality prior to IRT analyses. This is an important assumptions underlying IRT theory and it is therefore important to have established that the ADAS-cog meets this assumption.

Analyses were carried out using the most recent of the patients' ADAS-cog scores. It was not reported whether the researchers who carried out the analysis also scored and diagnosed the patients. This introduces some possibility of bias.

Table 3.1 Articles meeting inclusion criteria applying IRT methods to cognitive measures of dementia

Study	Ashford et al. (1989)	Mungas & Reed (2000)	Gibbons et al. (2002)	Benge et al. (2009)
Country	USA	USA	USA and UK	USA
Setting	Geriatric Psychiatry Outpatient clinic	Two clinical sites of Alzheimer's Disease Centre	Two community based samples from USA and UK	Alzheimer's Disease and Memory Disorders clinic
N	86	1207	540 (US: 401, UK: 139)	1087
Sex	73.2% female	64.7% female	(US) 64% female (UK) 75% female	66.6% female
Age			(US)	(UK)
Mean	74	76	82	84.7
SD	8	8.9	4.7	5.3
Range	53-91	39-100	> 75	>75
Etiology; n (%)	Probable AD: 52 (60) Possible AD: 34 (40)	Probable AD: 592 (49.0) Possible AD: 176 (14.6) Vascular: 60 (5.0) Mixed and other dementia: 325 (26.9) No cognitive impairment: 27 (2.2) Diagnosis deferred: 27 (2.2)	UK: AD: 139 (100) US: Probable AD: 338 (84.2) Possible AD: 63 (15.7)	AD: 1044 (96) MCI: 43 (4)
Dementia Severity	Mean MMSE=18 SD=7.1 Range=1-29	Mean MMSE= 17.7 SD=7.3 Range=0-30 Mean BIMCT= 16.9 SD=8.3 Range=0-33	US: Mean MMSE=19.6 SD=4.9 Range=1-29 UK: Mean MMSE=16.5 SD=5.5 Range=0-25	Mean ADAS cog=31.2 SD=16.5 Range= Not reported
Cognitive Measure	MMSE	MMSE, BIMCT	MMSE	ADAS-cog
IRT method	Item Characteristic Curve analysis	Two-Parameter model	Item Characteristic Curve Analysis	Samejima's graded model
Outcome	Hierarchy of item difficulty and discrimination	Hierarchy of item difficulty of Global Function scale. Investigation of linearity of MMSE, BIMCT and Global Function.	Hierarchy of item difficulty from 2 samples	Discrimination and information statistics on ADAS-cog test as whole, plus domains and subscales

Note. AD=Alzheimer's disease, MCI=mild cognitive impairment, MMSE=Mini Mental State Examination, ADAS-cog=Alzheimer's disease Assessment Scale-Cognitive subscale, BIMCT=Blessed Information Memory Concentration Test.

Table 3.2 Item *difficulty* comparison across studies

	Ashford et al. (1989) (MMSE)	Gibbons et al. (2002) UK (MMSE)	Gibbons et al. (2002) US (MMSE)	Mungas and Reed (2000) (BIMCT/MMSE)
Truncated above upper limit	Recall: Tree Recall: Flag	No ifs ands or buts Recall nouns		
1 st Quartile (Most difficult)	Serial sevens: Subtraction 5 Serial sevens: Subtraction 3 Orientation to date Recall: Ball	Orientation to date Verbal directions Intersecting Pentagons Serial sevens	Orientation to date No ifs ands or buts Intersecting Pentagons Serial sevens	Recall '42' (BIMCT) Recall 'Market Street' (BIMCT) Recall 'John' (BIMCT) Recall 'Chicago' (BIMCT) Recall 'Brown' (BIMCT)
2 nd Quartile	Serial sevens: Subtraction 4 Serial sevens: Subtraction 2 Orientation to day Orientation to county Orientation to month Serial sevens: Subtraction 1 Orientation to year Orientation to season Orientation to place Orientation to floor	Orientation to year Orientation to county/streets Orientation to day Orientation to month	Recall nouns Orientation to day Orientation to year Orientation to season Orientation to month Orientation to county/streets	Orientation to year (BIMCT/MMSE) Orientation to month(BIMCT/MMSE) Age (BIMCT)
3 rd Quartile	Orientation to city Intersecting Pentagons Orientation to state Write sentence No ifs ands or buts Name watch Verbal directions: Paper-on floor	Orientation to state/county Write sentence Orientation to Season Orientation to Address	Orientation to address Verbal directions Write sentence Orientation to place Orientation to city	Orientation to state (MMSE) Type of work (BIMCT) Count forward(BIMCT) Name watch (MMSE)
4 th Quartile (Least difficult)	Close eyes Repeat: Flag Name pencil Repeat: Ball Repeat: Tree Verbal directions: Paper-take in right hand Verbal directions: Paper-fold in half	Repeat nouns Orientation to city/town/village Orientation to room	Orientation to state Close eyes Name objects	Place of birth (BIMCT) Name pencil (MMSE) Name (BIMCT)
Truncated below 0		Close eyes Name objects	Repeat nouns	

Note. MMSE=Mini Mental State Examination, BIMCT=Blessed Information Memory Concentration Test. Ashford et al. (1989) and Gibbons et al. (2002) test items divided into quartiles based on range of scores. Mungas and Reed (2000) items divided into quartiles based on *difficulty* parameters. Most *difficult* items were truncated above upper limit as *difficulty* estimates were above the upper limit. Easiest items were truncated below 0 as even this low level of ability most participants were able to answer correctly. Some differences between MMSE versions between studies led to some discrepancies between items, e.g.: state/county

3.4 Discussion

This is the first systematic review of studies applying IRT methods to the assessment of cognitive decline in dementia. This review employed a comprehensive search strategy and included a detailed narrative review of the studies meeting the inclusion criteria.

This review appraised four published studies of IRT analyses of the cognitive decline of 2,920 participants with dementia. The four studies reviewed provided demonstrations of the applicability of IRT to assessment of cognitive functioning in dementia.

3.4.1 Item difficulty

Three of the four studies established a hierarchy of item *difficulty* (Mungas & Reed, 2000; Ashford et al., 1989; Gibbons et al., 2002). Two of these hierarchies were of the MMSE items (Ashford et al., 1989; Gibbons et al., 2002) and the third was of the Mungas and Reed ‘Global Function’ scale (Mungas & Reed, 2000). The dichotomization of MMSE items in Gibbons et al. (2002) decreased the ease at which direct comparisons of item *difficulties* between different studies could be made. In an attempt to equate the different range of MMSE scores across the studies items were divided into quartiles based on score ranges and *difficulty* parameters.

Table 3.2 shows that ‘orientation to date’, ‘recall nouns’ and ‘serial sevens’ are consistently the most *difficult* items across studies. A clinician identifying problems with these tasks could expect the patient to develop further cognitive *difficulties* in the progression suggested by the hierarchies in Table 3.2. Generally the least *difficult* items were; ‘name objects’, ‘repeat nouns’ and ‘close eyes’. Problems with these items can help identify severe dementia. From a clinical perspective this information is very useful. It provides a clearer

insight into decline than the traditional scoring method. *Difficult* items are very informative as it is likely that a patient with no *difficulties* here will not have limitations with other less *difficult* items. The items most consistently found the least *difficult* could be used in a similar fashion. It is likely that a patient unable to correctly respond to these items would have problems with most of the other items in the scale. In this way IRT analyses can identify key items from a scale that can quickly inform clinicians of a patient's level of functioning, for example, a clinician could select from the most *difficult* items such as 'recall nouns' to identify potential early cognitive difficulties in the healthy elderly.

None of the studies attempted to determine whether the hierarchies of *difficulty* held at the individual level (ordering items in terms of *difficulty* does not necessarily mean the ordering is the same for every person; those with higher levels of ability may find one item more *difficult* than the other yet the ordering may be reversed for those with lower ability levels (Ligtvoet, 2010; Ashford et al., 1989) by considering IIO. As invariantly ordered hierarchies are of great clinical value this should be included in future studies.

3.4.2 Discrimination

Two studies determined item *discrimination* (Benge et al., 2009; Ashford et al., 1989). Table 3.3 summarises the findings from these papers, showing the most *discriminatory* items at the various stages of disease. High *discrimination* for low *difficulty* items indicates that the abilities assessed by these items are lost at an advanced stage and that these losses are rapid once this stage has been reached. For more *difficult* items high *discrimination* means that these abilities are lost in the early stages and quickly at this stage.

Items with low *discrimination*; 'repeat nouns', 'no ifs, ands or buts', 'orientation to day and season', 'orientation to country, floor and city', 'copy pentagons' also reveal valuable insights. For these items the range of scores in which participants respond either

CHAPTER 3: SYTEMATIC REVIEW

correctly or incorrectly is wider than high *discriminating* items. Either the abilities being measured by these items are lost with more variability or more gradually or the functions measured here are assessed less concisely by these items.

Including more items like ‘word recall’ and ‘orientation to date’ may help to detect changes in milder stages of the disease as these abilities are lost quickly at an early stage. For severe dementia the inclusion of simple repetition tasks or non-cognitive functioning tasks could help to introduce greater *discrimination* in this stage. Items such as recalling or recognizing one’s name, from the Severe Cognitive Impairment Rating Scale, measuring the ability of overlearned autobiographic memory, could be applied to broaden the range of assessment in cognitive instruments.

From a large battery of items those demonstrating the best *discrimination* across the disease course could be used to create an instrument to accurately measure patients in early and late stages. More precise assessment would lead to enhanced measurement of the rate of decline and improve predication of impending deterioration.

While these studies demonstrate the use of IRT to examine item *difficulty* and *discrimination* the investigation of item differences has also been addressed using classical test theory. Chapman and Chapman (1973) identified the need to study these item parameters in their analyses of specific and differential deficits in psychopathology research, for example, specific deficits in schizophrenia or the analysis of domains or abilities which remain relatively intact in dementia. Chapman and Chapman’s analyses of differential deficits is rooted in classical test theory and IRT, as a newer statistical model, offers alternative means of exploring the differential deficit problem. When examining differential deficits between different groups IRT, unlike CCT, can offer estimates of measurement error for different levels of cognitive ability, without having to conduct separate studies, and can establish whether different items or measures are equally *difficult*.

Table 3.3 High *discrimination* items and disease stages

	Early disease/ High difficulty	Moderate stages	Late disease/ Low difficulty
High discrimination	Orientation to date (MMSE) Word recall (ADAS-cog)	ADAS-cog Ideational praxis (ADAS-cog) Construction (ADAS-cog) Word finding (ADAS-cog) Speech comprehension (ADAS-cog) Commands (ADAS-cog) Speech content (ADAS-cog) Naming (ADAS-cog)	Name pencil (MMSE) Close eyes (MMSE) Name watch (MMSE)

Note. MMSE=Mini-Mental State Examination, ADAS-cog=Alzheimer's Disease Assessment Scale-Cognition

3.4.3 Linearity and the assessment of change in severity

Two studies investigated whether the magnitude of cognitive dysfunction represented by each item on the cognitive scale was equal across the scale (Mungas & Reed, 2000; Bengtson et al., 2009). In a recent paper Balsis, Unger, Bengtson, Geraci and Doody (2012) also drew attention to the limitations associated with the traditional method of measuring cognitive dysfunction with the ADAS-cog. This study was not included in the review as it did not provide information on the individual items or subscales however its analysis of IRT scoring of the ADAS-cog is worth noting. Balsis et al. (2012) found that individuals with the same total score can have different degrees of cognitive impairment and conversely those with different total scores can have the same amount of cognitive impairment. These findings are supported by a similar study also failing to meet inclusion criteria due to some use of non-English language measures and a lack of information on test/item information (Wouters et al., 2008). Results indicate that participants with equal ADAS-cog scores had distinctly different levels

CHAPTER 3: SYTEMATIC REVIEW

of cognitive impairment. Equally, participants with the same estimated level of impairment had wide ranging ADAS-cog scores. The same differences in scores did not reflect the same differences in level of cognitive impairment along the continuum of test score range. Without equal intervals between adjacent test items change scores may reflect different amounts of change for subjects with differing levels of severity, or may fail to identify change at all (de Morton, Keating & Davidson (2008). Wouters et al. (2008) revised the ADAS-cog scoring based on the results of this IRT analysis by weighting the items in accordance with their measurement precision and by collapsing their categories until each category was hierarchically ordered, ensuring the number of errors increase with a decline along the continuum of cognitive ability. Examining *difficulty* hierarchies of the error categories within the items revealed some disordered item categories. As the categories are only useful if they have a meaningful hierarchy of *difficulty* these disordered categories were collapsed until all categories were correctly ordered in hierarchies of *difficulty*. This revision resulted in a valid one to one correspondence between the summed ADAS-cog scores and estimated levels of impairment.

These studies demonstrate the potential to misinterpret test scores due to a lack of measurement precision. This is illustrated the examination of linearity of the MMSE, BIMCT and the ‘Global Function’ scale (Mungas & Reed, 2000). The findings of non-linearity of the MMSE and BIMCT indicate that a change in total score is less for a given specified change in ability at the two ends of ability distribution than it is in the middle of the ability distribution. For example, a two standard deviation change in ability from 3.0 to 1.0 reflects an approximate five point MMSE score loss, whereas the same degree of change from 1.0 to -1.0 represents a 15-point MMSE score loss. A similar pattern was found for the BIMCT. IRT methods can be used to create a scale with greater linearity by establishing item *difficulties*, as illustrated by the ‘Global Function’ scale (Mungas & Reed, 2000). The ‘Global Function’

CHAPTER 3: SYTEMATIC REVIEW

scale shows promise of linear measurement throughout the majority of the continuum of ability. This new measure, along with any new IRT measure, would need to be cross-validated and directly compared to existing clinical instruments to ensure this test development technique is truly beneficial. It is worth noting that this measure also incorporates items assessing independent functioning. The inclusion of tasks such as these with meaningful variability even in the late stages of dementia could afford the test more *discriminatory* power increasing the information at this stage. While this review did not aim to include functional scales this study suggests that scales that combine cognitive and functional items, or concomitant use of both types, may provide added value. A limitation of this and many other cognitive functioning scales is the lack of items sensitive to very mild early stage of dementia. The inclusion of items capable of *discriminating* mild dementia could improve measurement properties in much the same way.

The measurement properties of a scale can impact the interpretation of clinical trials as change scores are used to determine the efficacy of interventions and treatments. A Cochrane review of AD pharmaceutical trials methods included ADAS-cog change scores to help ascertain the effectiveness of cholinesterase inhibitors (Birks, 2006). Benge et al. (2009) confirmed that the degree of cognitive ability symbolized by each point on the ADAS-cog was not uniform across the scale. A three point change in raw scores can represent a change in cognitive abilities ranging from 0.85 standard deviations of cognitive functioning (representing a change from a score of 4 to 1) to 0.14 standard deviations of cognitive functioning (from a score of 37 to 34).

The observation of differences between and within people may be greatly aided using an IRT approach. In clinical trials it is possible that these analyses will lead to an increased ability to correctly identify group treatment differences and to recognize responders and nonresponders to treatment.

3.4.4 Information

Another advantage of IRT is the increased reliability it provides however, only Bengtson et al. (2009) estimated the *information* parameter. The ADAS-cog has the highest level of *information* at moderate levels of cognitive impairment. At milder levels of impairment the *information* function remains low which indicates that the test domains; language, memory and praxis, and the measure as a whole do a relatively poor job discriminating among the different levels of impairment in the mild severity range. The same can be said about the severe levels of impairment. That moderate levels have the highest *information* function is unsurprising as the ADAS-cog was originally designed to measure moderate AD. Decreased *information* at mild and severe levels could affect the interpretation of the significance of the change scores at these levels of impairment.

This review excluded 28 studies using general populations, some of which included some dementia subgroups. In an effort to widen the scope of the review studies using general populations including some participants with dementia were looked at to determine if these dementia subgroups could be analysed separately. However it was determined that these papers failed to meet inclusion criteria for reasons beyond the sample characteristics, mostly for the use of non-English language measures, and therefore the authors of the papers were not contacted for further details. One such study analysed a Japanese version of the MMSE within a general population (Ideno et al., 2012). However the ordering of items was examined for the AD subgroup in isolation illustrating the sequence of cognitive decline. IRT analysis found the scale could be simplified with the removal of items showing similar ICCs and factor loadings, reflecting potential redundancy. ‘Naming’ was deemed to be similar to ‘three-step command’ and was deleted along with ‘read and follow instruction’ showing similarity to ‘repeat a sentence’ and ‘orientation to time’ as its function was comparable to ‘orientation to place’. The ordering from least to most *difficult* was ‘three-step command’,

CHAPTER 3: SYTEMATIC REVIEW

‘registration’, ‘repeat a sentence’, ‘write a complete sentence’, ‘copy drawings of two polygons’, ‘delayed recall’, ‘orientation to place’ and ‘serial sevens’.

Twenty one studies were excluded for administering non-English measures. However, all except one were excluded for other reasons also (16 did not meet inclusion criteria for the use of non-dementia samples, three written in foreign languages, one for the use of non-dementia sample and analysis of domain specific measure, and one for lack of test/item information). The results of the single study (Korner, Brogaard, Wissum & Petersen, 2012) which was only excluded due to use of a Dutch version of the Baylor Profound Mental State Examination are discussed. Korner et al. (2012) applied Mokken analysis and the one-parameter Rasch analysis in a validation study of the cognitive part of the Danish version Baylor Profound Mental State Examination. In doing so the relative *difficulty* of the test items were estimated. The *difficulties* of the 25 items were evenly distributed along the ability range with no redundant items. The least *difficult* items in this measure were; “What is your name?” and the repetition of the first word (one syllable). The most *difficult* item was the drawing of ‘intersecting pentagons’. While the other studies administering such measures would not have been included for various other reasons there are data that may be informative (Lindeboom et al., 2004; Ideno et al., 2012; Wouters et al., 2010; Dodge, Meguro, Ishii et al., 2009).

While global cognitive instruments such as the MMSE are probably the most commonly used measure of cognitive functioning domain specific neuropsychological tests have been demonstrated to show increased sensitivity to early stages of cognitive impairment than measures of global cognition (Harrison, 2007). However of the seven studies applying IRT methods to domain specific measures identified (Teresi et al., 2000; Bengtson, Miller, Bengtson & Doody, 2011; Crane et al., 2008; del Toro, 2011; Fernandez-Blazquez et al., 2012; Graces, Bezeau, Fogarty & Blair, 2004; Diesfeldt, 2004) only one; Bengtson et al. (2011)

CHAPTER 3: SYTEMATIC REVIEW

otherwise met inclusion criteria. This study's findings were briefly discussed here. Temporal ('day of month', 'year', 'month', 'day of week' and 'season'), and spatial ('name of hospital', 'floor', 'town', 'country' and 'state') Orientation items of the MMSE, were analysed to determine their *difficulty* and *discrimination* parameters. The most *difficult* item was 'floor of hospital' and the least *difficult* item was 'state'. The full order of item *difficulty* was; 'floor', 'name of Hospital', 'date', 'day of week', 'year', 'month', 'season', 'country', 'town' and 'state'. A relatively high level of ability (2.81SD) is required to have a 95% chance of correctly identifying the floor of the building which illustrates that knowing which floor of the hospital reflects a relatively high level of cognitive ability. Clinicians can use this sort of knowledge to help interpret the information they get from their assessments.

The spatial orientation items *discriminate* best at varying levels of cognitive ability with a wider range of *difficulties* assessed than the temporal items. Spatial items could be used to create a short scale sensitive to a relatively broad range of abilities. The temporal items assess a narrower breadth of abilities at a relatively modest degree of impairment and therefore would be best suited to identifying change within this range of cognition.

The value contributed by each item was examined to reveal key items and those whose function was largely redundant. 'Year' and 'month' provide roughly the same information as they have similar levels of *discrimination* and *difficulty*, as do 'State' and 'town'. Both item pairs provide no meaningful variability to the set of items. One item from each pair would be sufficient to capture the same information as both. 'Date', 'name of Hospital' and 'State' together sample the range of cognitive abilities assessed by the orientation items and could together provide key information about a wide range of abilities.

3.4.5 Limitations

While the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed insofar as they were applicable for methodological studies the studies identified in this review did not allow a statistical summary or to perform a meta-analysis due to the variety of subjects, sites, diagnostic criteria and the varied statistical item response theory methods applied. The four studies cross a 20 year span with the earliest data collection and diagnoses in 1984 (Ashford et al., 1989) with the most recent in 2002 (Benge et al., 2009). This will affect criteria for diagnosing dementia. With mostly moderate ranges of dementia the studies also represented a rather restricted range of severity limiting the scope of the analysis, as the findings cannot be extrapolated to mild or severe dementia. IRT analyses assume unidimensionality which limits its application to measures assessing a single latent construct. However only one study reviewed here explicitly assessed unidimensionality prior to IRT analyses (Benge et al., 2009).

Three of the four studies failed to report who administered the test to participants and whether these individuals were blind to the diagnoses (Mungas & Reed, 2000; Ashford et al., 1989; Benge et al., 2009). This introduces some potential bias in these studies.

This review was limited to analyses of only three global cognitive function; MMSE, BIMCT and ADAS-cog. This was a consequence of the articles meeting inclusion criteria. However, an analysis of the Baylor Profound Mental State Examination, while not reviewed due to use of a Dutch version, was mentioned in the discussion (Korner et al., 2012).

With the exception of Mungas and Reed (2000) all studies solely included patients with Alzheimer's disease. This could have an impact on findings, as there should be a different pattern of decline between different aetiologies. Of the excluded articles one included patients with amyotrophic lateral sclerosis and behavioural variant frontotemporal dementia which would have expanded the scope of this review (Lillo, Savage, Mioshi,

Kiernan & Hodges, 2012). However this study failed to provide data on the measure of cognition in isolation from the other outcomes studied and for this reason was excluded.

3.5 Conclusion

This systematic review of IRT use in cognitive tests in people with dementia found only four relevant published papers. These include heterogeneous populations, with widely varying sample sizes, different methods of dementia diagnosis (and inclusion of possible dementia or MCI), and samples are mostly derived from specialist clinical populations, with a risk of inclusion bias. Most participants had Alzheimer's dementia of moderate severity, and were resident in the United States, so the relevance of this method to other subtypes of dementia, and other countries, cannot be determined. Different cognitive tests, and IRT methods, were used, and different statistics were reported. However, the studies show that IRT can demonstrate which items within scales are most *difficult*, and *discriminatory*, at different severities of dementia. IRT analyses can also be used to reveal non-uniform distances between scale scores and facilitate the creation of scales with enhanced measurement properties allowing more accurate assessment of change across the ability spectrum.

There is a need for more IRT analyses of cognitive scales used to assess dementia. These should include standard methodologies, and report item *difficulty* and *discriminatory* statistics along with a measure of *information* and an assessment of linearity of measurement. They should include large numbers, from a variety of countries (both English speaking and non-English-speaking), different dementia subtypes, the full range of severity of dementia, and a wider range of cognitive tests, focusing on those that are widely used in clinical practice. This will allow refinement of these tools to improve the information provided to

CHAPTER 3: SYTEMATIC REVIEW

clinicians on how performance on items within the scale is informative at different stages in dementia.

Chapter four, the first empirical chapter of the thesis, will address this need for item level analyses of cognitive scales in dementia assessment.

Chapter 4: Does the order of item difficulty of the Addenbrooke's Cognitive Examination add anything to sub-domain scores in the clinical assessment of dementia?

Work presented in the following chapter is taken from the following paper:

McGrory, S., Starr, J.M., Shenkin, S.D., Austin, E.A. and Hodges, J.R. (*in press*). Does the order of item difficulty of the Addenbrooke's Cognitive Examination add anything to sub-domain scores in the clinical assessment of dementia? *Dementia and Geriatric Cognitive Disorderd EXTRA*, doi: 10.1159/000375364

4.1 Introduction

Cognitive measures are commonly used to screen for dementia as well as assessing severity and monitoring disease progression. Often underlying these tests is the assumption that cognition deteriorates along a fixed course of decline on a single cognitive trait (i.e. total test scores are considered meaningful in themselves) and that the impairment and severity can be measured when a patient is unable to respond correctly to certain cognitive challenges (Wouters, van Gool, Schmand, Zwinderman & Lindeboom, 2010). Looking at total and sub-domain scores may lead to important information being neglected. For example, two different individuals achieving the same score on a cognitive measure may have reached this score by missing different combinations of items. Using the summed score as a measure of cognition fails to take into account the information embedded in the specific pattern of scores. Items may differ in several ways. Different items on a scale may be unequally related to the

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCLTY

construct of cognitive impairment. Additionally test items are likely to differ in terms of *difficulty* - how *difficult* an individual finds it to respond correctly to an item-(Reise, Ainsworth & Haviland, 2005).

The ACE (Mathuranath et al., 2000) was originally developed to provide a brief test that would be both sensitive to the initial symptoms of dementia, and to be capable of discriminating different types of dementia including Alzheimer's disease and frontotemporal dementia (Mathuranath et al., 2000). The ACE and the revised version ACE-R encompass tests of attention/orientation, memory, language, visuospatial abilities and executive function. They also incorporate the Mini Mental State Examination (MMSE) (Folstein, Folstein & McHugh, 1975) so this score may also be produced. The ACE is relatively quick to administer (approximately 15 minutes) and has good sensitivity and specificity for identifying dementia (Mathuranath et al., 2000). While modifications to the ACE have been made to address the original scale's weaknesses there have been no examination of the item properties or hierarchical structure of either the original ACE or its successor the ACE-R using item response theory methods.

Factor analysis can be used to investigate the relationship between ACE-R items and the total score. While this method offers some insight into the dimensionality of the ACE-R, item response theory (IRT) (Hambleton & Swaminathan, 1985) can provide further insight into the item properties and how they function in relation to the other items within the scale. This item level analysis can be applied to determine whether the items form a hierarchy of item *difficulty*.

The interpretation of the ACE-R and other cognitive measures would be greatly improved if the ordering of the *difficulty* of the cognitive tasks (items) is similar for patients at different stages of dementia. When the ordering of the items by mean scores is the same across different values of the latent construct it can be said that items conform to a

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFICULTY

hierarchical scale with IIO. IIO can facilitate diagnosing dementia (Meijer, Egberink, 2011). For example, an IIO hierarchy detailing the expected trajectory of decline in Alzheimer's disease may differ from an IIO hierarchy of decline in semantic dementia. In this way IIO hierarchies can be used to identify distinctive profiles of cognitive dysfunction which can serve as an adjunct to diagnosis and help with escalation of care planning. IIO can also facilitate the comparison of patients with respect to their degree of cognitive decline, for example a patient experiencing problems with one of the least *difficult* items in the hierarchy would be considered more severely impaired than a patient only experiencing a problem with one of the most *difficult* items in the hierarchy. IIO hierarchies can also be useful in the detection of unexpected score patterns (Sijtsma & Molenaar, 2002) and in characterising differences among subgroups and different forms of dementia.

Mokken scaling analysis (Sijtsma & Molenaar, 2002; Mokken, 1971) described in detail in Chapter 2 is based on IRT principles and is commonly applied to determine whether hierarchical scales meeting IIO criteria exist within data. This method has been more frequently applied to dichotomous items within scales. However examining polytomous scales (i.e. scales with more than two response options, for example “strongly disagree”, “disagree”, “agree”, “strongly agree” or an item with a score range of 0-3) for IIO has recently become possible (Ligtvoet, van der Ark, Bergsma & Sijtsma, 2011; Stochl, Jones, Croudace, 2012).

The aim of this chapter is to determine whether the ACE-R has hierarchical properties with IIO and to compare these findings with factor analysis using structural equation modelling to determine whether this hierarchy can add to the information provided by the sub-domain scores.

4.2 Method

4.2.1 Participants

A sample of 350 was sourced from the specialist multidisciplinary tertiary referral centre, the Frontier Research Group, at Neuroscience Research Australia (NeuRA), Sydney. Patients meeting current clinical diagnostic criteria for behavioural variant frontotemporal dementia (bv-FTD) (Rascovsky et al., 2011), Alzheimer's disease (McKhann et al., 2011), logopenic progressive aphasia (LPA) (Gorno-Tempini et al., 2011), motor neurone disease (MND) (Brooks, Miller, Swash & Munsat, 2000), progressive nonfluent aphasia (PNFA), or semantic dementia (SD) (Neary et al., 1998), were recruited through the Frontier Research Group. Diagnosis was established by consensus among neurologist, neuropsychologist and occupational therapist, based on extensive clinical assessments, cognitive assessment, and evidence of atrophy on structural magnetic resonance imaging (MRI) brain scans. All patients provided informed consent for the study and dual consent was obtained from the carer in some cases. Patients underwent clinical, neuropsychological, behavioural and imaging assessment between 2007 and 2011. Data from patients with complete itemised ACE-R data ($N=350$) were included in the analysis.

The sample was very diagnostically heterogenous and in an attempt to limit the effects of this heterogeneity the sample was divided into three groups: Alzheimer's type: AD and LPA ($n=131$), predominantly frontal dementia; bv-FTD and FTD-MND ($n=119$), other frontotemporal lobe degenerative disorders; other frontotemporal lobe degenerative disorders temporal: SD and PNFA ($n=100$).

4.2.2 Measures

The ACE-R comprises 26 items and is scored out of 100 and includes items assessing 5 cognitive domains: attention/orientation (18 points), memory (26 points), fluency (14 points), language (26 points) and Visuospatial (16 points). The total ACE-R score is created by the addition of all the domains.

The mean for each ACE-R item score was divided by the maximum number of points available for that item to equate scores for comparison (i.e. equal weighting of items even though items can contribute different weighted values to the summed total score), giving a score with minimum 0 and maximum 1. For example, the mean score of 2.5 for ‘memory retrograde’ for the predominantly frontal group was divided by 4 (the maximum number of points available on this item) to give a new ‘overall’ mean score of 0.625. These equated mean item scores were used for the analyses.

Although the rescoring of ‘naming (10 items)’ potentially removes some important variation in response, this was minimized by collapsing the item responses at the bottom end of the range since the prevalence of responses in the lowest category is very low ($n=34$, 9.7%).

4.2.3 Factor analyses

To identify the underlying factor structure, an exploratory principal component analysis (PCA) was performed on the subdomain scores for each of the diagnostic groups using the IBM SPSS, version 19. Inspection of scree plots and the Kaiser criterion of eigenvalues > 1 were used to decide on the number of components to extract.

The final factor solution derived from the PCA was entered into AMOS and converted to a simple structure confirmatory factor analysis (CFA) model, in which one latent

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

variable explained the covariance in the five subdomains. CFA was performed on the emergent factor structure to evaluate whether the PCA model fit the data well. Subsequent analyses were conducted on the best-fitting model to determine whether the model exhibited invariance across different diagnostic groups.

The Comparative Fit Index (CFI; Bentler, 1980) and the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993) were used to estimate the model fit. The following rules of thumb with regards to model fit were used: CFI < 0.90 indicates poor fit, $0.90 < \text{CFI} < 0.95$ indicates a reasonable model fit, and CFI > 0.95 indicates good model fit; RMSEA > 0.10 indicates poor fit, $0.05 < \text{RMSEA} < 0.10$ indicates reasonable model fit, and RMSEA < 0.05 indicates good fit (Kline, 2005). Modification index (MI) values were inspected to determine whether alterations to the model were required to ensure the model fit indices were within the acceptable ranges for good model fit. All confirmatory and invariance analyses were conducted with AMOS 19.0 (Arbuckle, 2009).

Invariance analyses are one method to explicitly test whether there are qualitative differences in a model across different diagnostic groups (i.e. patients diagnosed with Alzheimer's disease, patients diagnosed with predominantly frontal dementia and patients diagnosed with other frontotemporal lobe degenerative disorders).

Following recommendations, between-group invariance was tested through a series of increasingly restrictive models (Byrne, van De Vijver, 2010). Firstly, configural invariance was tested by assessing the fit of an unconstrained model. This model serves as the baseline model against which to compare more constrained models. In the first constrained model metric invariance was assessed by constraining item loadings to be the same across groups. Structural invariance was tested by constraining structural variances fixing factor variance across groups. At each step, whether the further constraints reduced model fit in comparison to the baseline model was tested.

4.2.4 Mokken scaling analysis

To determine whether the ACE-R conforms to a hierarchical scale and if so, how this hierarchy relates to the factor structure Mokken scaling analysis was carried out. Data were analysed using the Mokken scaling analysis package in the public domain software ‘R’ in which software is available to test the assumptions of both Mokken models; the monotone homogeneity model (MHM) and the double monotonicity model (DMM) (van der Ark, 2007).

The 26 items of the ACE-R were analysed. Some of these items are a composite of several embedded questions such as ‘orientation in time’ on which a patient receives a score from 0-5 based on their ability to correctly identify the correct day, date, month, year and season. Mokken scaling does permit polytomous data but as only the total score from these items (i.e. a score out of five for ‘orientation in time’) was reported the embedded items (e.g. ‘what is the month?’) could not be isolated for analysis. Instead Mokken scaling was performed on the polytomous composite item score.

This chapter firstly assesses the fit of the ACE-R data to the MHM. The fit of this model implies that all respondents within the sample can be invariantly ordered along the latent trait, i.e. cognitive impairment (Junker & Sijtsma, 2001). The MHM is based on the assumptions of; unidimensionality, local stochastic independence and monotonicity. With regards to the current analysis the assumptions of the MHM were assessed to determine whether the ACE-R singularly measures cognitive impairment and that all respondents can be rank ordered by their level of cognitive functioning by means of the summed score of their responses to the ACE-R items. The assessment of monotonicity is important as it enables the respondents to be ordered on the latent trait with respect to the summed score of the scale

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

(Stochl et al., 2012). Practically the monotonicity of an item (k) is determined by replacing the latent trait value with a restscore (the sum of scores for all items except for item k). If monotonicity holds, it can be assumed that a higher proportion of respondents with restscore i will respond correctly to the item than respondents with restscore j , for any pair of number $j < i$ (Stochl et al., 2012). This is an important psychometric property of any scale as it implies that respondents with a summed total score on a scale of 10 for example, have a level of latent trait that is at least as high as those with a total score of 9 and that these respondents in turn will have a trait level that is at least as high as those with a score of 8 and so on.

Mokken scaling procedures provides several parameters for determining whether the data meet the assumptions of the MHM and conform to Mokken scales. While Mokken scaling is considered as a probabilistic reworking of the deterministic Guttman scaling (Guttman, 1944), the strength of Mokken scales is determined based on the number of Guttman errors (Niemoller & van Schurr, 1983). A Guttman error can be observed when the relative response to a pair of items is not in the expected direction (Watson, Wang & Thompson, 2014). The fewer the Guttman errors the stronger the Mokken scale is considered (Sijtsma & Molenaar, 2002). Loevinger's H a measure of the strength and quality of a Mokken scale is used to indicate the extent of Guttman errors and as such is an expression of the degree to which the items consistently appear in the same relative order and justifies their use in forming a unidimensional latent variable (Sijtsma & Molenaar, 2002). The overall H for the scale as a whole along with a coefficient (H_i) for each of the individual items within the scale is calculated. Generally, $H=0.3$ is the minimum value for a Mokken scale with higher values reflecting greater strength of ordering and fewer violations (DeJong & Molenaar, 1987). Violations, or Guttman errors, are defined as any deviations of the data from the expected ordering. For example, if an item (i) with a lower mean score than another

item (j), indicating that item i is a more “*difficult*” item, then any time a item i has a higher score than item j an error has occurred (Watson, 1996).

Monotonicity can be examined by calling function *check.monotonicity* in R. This function calculates the number of scaling violations; where the predicted order of an item pair is reversed, and summarises these in the output for inspection.

Local stochastic independence (LSI) of items implies that that a respondent’s response to one item in the test is not affected by his or her response to any other item in the scale. Local independence implies that all systematic variation in responses to the items is exclusively caused by the variation of respondents over θ (Mokken, 1997). By nature items belonging to the same scale have to covary to some degree (Nader, Tran, Baranyai & Voracek, 2012) but LSI implies that this item covariance is due to the latent trait they all measure (Sijtsma & Molenaar, 2002). This means that an individual’s response to one item within a scale is independent of their response to any other item within the scale. Methods for estimating stochastic dependence in polytomous items within Mokken scaling procedure are in development but are currently unavailable (Straat, 2012).

These diagnostics and parameters are used to establish whether the assumptions of the MHM hold and are used to assess the fit of the data to Mokken’s first level of analysis; the scalability of the items. Scalability measures the extent to which respondents can be reliably ordered on the level of latent trait by means of their summed total score (Roorda, Scholtes, Van der Lee, Becher & Dallmeijer, 2010). However, it is important to note that while H can inform on whether a set of items form a unidimensional and monotonic scale it is not sufficient to determine whether the items form a hierarchical scale (Meijer, 2010).

Only Mokken’s second level of analysis; the assessment of IIO, can confirm whether a set of items form a hierarchical scale (Meijer, 2010). Determining IIO of polytomous items involves an analysis of the responses to each of the levels in the items (e.g. score of 0-5 for

one item) as opposed the item response functions (IRFs). For dichotomous items IIO is established by examining non-intersection of IRFs. Determining IIO of polytomous items involves an analysis of the responses to each of the levels in the items-item step response functions (ISRFs) (e.g. score of 0-5 for one item) as opposed the item response functions (IRFs). Non-intersection in the case of polytomous items is established where there is no intersection in each of the steps between response categories. The relationship between these responses and the score on the latent trait is symbolised using item step response functions (ISRFs). In the case of an item scored from 0-5 where there are four steps between the five possible responses there are four IRSFs. Non-intersection in the case of polytomous items is established where there is no intersection in each of the steps between these response categories. However the non-intersection of IRSFs does not imply IIO for polytomous items (Meijer, 2010; Sijtsma, Meijer & van der Ark, 2011). Most software cannot be used to assess IIO for polytomous items as only the ISRFs within each item are analysed and not the items themselves. This key element of Mokken scaling is only currently possible using the “mokken” package of R software using the function *check.iio* (Stochl et al., 2012). This function uses a backward selection procedure that starts with the items with the highest number of IIO violations and iteratively removes items until there are no significant violations of IIO remaining. Using this method a diagnostic H_{trans} or H^T is used to establish the strength of IIO, similar to the heuristics of H , with H^T values >0.3 indicating a scale with IIO (Ligtvoet, 2010). Various methods are available within the function *check.iio* including manifest invariant item ordering (MIIO), which is R’s default method, and manifest scale - cumulative probability mode (MSCPM). MSCPM investigates the manifest item step response functions for all pairs of items. However this stronger form of IIO has several practical disadvantages. This method results in an extremely large number of comparisons and as this method is particularly sensitive it tends to suggest the removal of all items.

Visual inspection of item-pair plots was also used to assess IIO. Item rest-score regression plots were visually inspected to identify item overlap or ‘outlying’ items: items located far away from the cluster of the other scale items. These items can cause artificially exaggerated IIO and can result in the misleading appearance of IIO (Meijer & Egberink, 2012).

4.3 Results

Three hundred and fifty participants (232 male, 118 female) with a mean age of 65.38 (SD=8.5) years, diagnosed with dementia were included in the analysis (see Table 4.1). The sample was diagnostically heterogeneous; bv-FTD ($n=96$), AD ($n=88$), SD ($n=61$), LPA ($n=43$), PNFA ($n=39$), and FTD-MND ($n=23$).

Table 4.1 Demographic and cognitive scores in dementia groups.

	AD type	Predominantly frontal	Other Frontotemporal lobe degenerative disorders
N (% male)	131 (55%)	119 (77%)	100 (68%)
Age (SD)	66.5 (8.3)	63.7 (8.9)	65.8 (8.9)
Education (years) (SD)	12.9 (3.4)	12.3 (3.3)	12.0 (3.6)
MMSE (SD)	22.0 (5.7)	23.7 (5.8)	21.3 (5.8)
ACE-R (SD)	64.2 (19.0)	70.3 (18.9)	54.9 (17.3)

Note. AD=Alzheimer’s disease; MMSE=Mini Mental State Examination; ACE-R=Addenbrooke’s Cognitive Examination-Revised, SD=standard deviation
Equated ACE-R item scores were used to designate item *difficulty* in Mokken scaling. These scores are presented for each of the three clinical groups in Table 4.2.

Table 4.2 ACE-R items group by domain, means and total scores.

Item	Domain	Label	AD type	Predominantly frontal	Mean	Max	
					Other frontotemporal lobe degenerative disorders		
1	Attention	Orientation in time	3.5	3.8	4.1	5	
2		Orientation in geography	3.9	4.1	3.6	5	
3		Three item registration	2.7	2.9	2.6	3	
4	Memory	Serial sevens	3.9	4.0	3.9	5	
5		Three item recall	1.1	1.5	1.1	3	
6		Name and address learning	4.6	5.8	4.6	7	
7		Memory retrograde	2.0	2.5	1.2	4	
25		Name and address recall	1.6	3.1	1.6	7	
26	Fluency	Recognition	3.5	3.7	3.3	5	
8		Verbal fluency-Letters	3.4	2.5	2.0	7	
9		Verbal fluency-Animals	2.5	2.5	1.3	7	
10		Language	Follow written command-close eyes	0.8	0.9	0.8	1
11			Syntactical comprehension	2.3	2.5	2.1	3
12			Write a sentence	0.8	0.8	0.7	1
13			Repetition of single multi-syllabic words	1.4	1.6	1.1	2
14		Repetition-above, beyond and below	0.7	0.8	0.6	1	
15	Repetition-no ifs, ands or buts	0.4	0.4	0.3	1		
16	Visuospatial	Naming (pencil and watch)	1.7	1.9	1.3	2	
17		Naming (10 items)	6.2	6.8	2.5	9*	
18		Semantic comprehension	3.1	2.8	1.8	4	
19		Reading	0.6	0.7	0.2	1	
20		Draw overlapping pentagons	0.7	0.8	0.9	1	
21		Draw a cube	1.2	1.5	1.7	2	
22		Draw a clock	3.4	3.9	3.5	5	
23		Count dot arrays	3.4	3.4	3.6	4	
24	Identify fragmented letters	3.7	3.9	3.7	4		

Note. Item numbers indicate the item locations in ACE-R test order. AD=Alzheimer's disease, Max=maximum score for each ACE-R item. *Maximum score for Naming (10 items) =10. All scores of 0 ($n=34$, 9.7%) were recoded as 1 to provide a range of 0-9 as Mokken Scaling analysis is unable to analyse items with scores >9.

4.3.1 PCA analysis

Visual inspection of scree plots and Kaiser's criterion (eigenvalue >1) were used to determine the number of factors to extract. Both methods suggested a single factor structure with the extraction of one component with an eigenvalue greater than one for the Alzheimer's type, predominantly frontal and patients diagnosed with other frontotemporal lobe degenerative disorders, explaining 65%, 68% and 61% of the variance in the groups respectively. The correlations between the extracted component and the ACE-R subdomains are similar across the three diagnostic groups as shown in Table 4.3.

Table 4.3 Correlations between ACE-R subdomains and component extracted from PCA

	Alzheimer's type	Predominantly frontal dementia	Other frontotemporal lobe degenerative disorders
	Component 1	Component 1	Component 1
Attention	0.855	0.865	0.827
Memory	0.857	0.861	0.815
Fluency	0.790	0.775	0.775
Language	0.838	0.842	0.842
Visuospatial	0.666	0.781	0.629

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised, PCA=Principal Components Analysis. Subdomain value derived from addition of mean item scores within each domain.

4.3.2 CFA analysis

This one-factor model derived from PCA was converted to a CFA model. CFA was performed to evaluate whether the PCA model fit the data well. Whereas PCA examined all

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

variance the CFA model examined the shared variance. This model fits the data well in the predominantly frontal group ($\chi^2 = 6.754$, $df = 5$; CFI = .994; RMSEA = .054) but less successfully in the AD type groups ($\chi^2 = 18.405$, $df = 5$; CFI = .957; RMSEA = .143) and other frontotemporal lobe degenerative disorders group ($\chi^2 = 40.327$, $df = 5$; CFI = .841; RMSEA = .269).

Modification index (MI) values prompted the addition of an error covariance (memory and visuospatial) to improve fit. This model fitted the data adequately in the AD type group ($\chi^2 = 5.365$, $df = 4$; CFI = .996; RMSEA = .051) and predominantly frontal group ($\chi^2 = 1.359$, $df = 4$; CFI = 1.000; RMSEA = .000) but less well for the other frontotemporal lobe degenerative disorders group ($\chi^2 = 23.216$, $df = 4$; CFI = 0.913; RMSEA = .221). Subsequent invariance analyses were conducted on this model to determine whether the model exhibited invariance across the groups. As the model did not fit the data well in the other frontotemporal lobe degenerative disorders group invariance analyses were performed on the AD type and predominantly frontal groups only.

The unconstrained model showed a good fit ($\chi^2 = 6.723$, $df = 8$; CFI = 1.000; RMSEA = 0.000) suggesting that the model was an appropriate representation of the data across groups with a common factor structure across both groups. Constraining the factor loadings to be the same in the two groups resulted in continued good fit (χ^2 of 3 ($df = 4$), $p = ns$) suggesting that factor loadings are invariant across both groups. Further constraining factor variances to be the same in the two groups produced non-significant changes in model fit (χ^2 of 3 ($df = 5$), $p = ns$), suggesting structural variances do not differ across the two groups. Therefore PCA shows a single factor in all three dementia types, but an invariant factor structure in only two. While the addition of an error covariance in order to improve model fit is a common practice in CFA there are concerns regarding the legitimacy of this practice (MacCallum, Roznowski & Necowitz, 1992). Where a model fails to fit the data well it is common practice to modify

the model to improve its fit to the sample data. Modifying the model with the addition of error covariances for example, is a data-driven approach which raises concerns about the generalizability of the modified model. Modifications to the initial model such as the addition of covariances among error terms have been described as “wastebasket” parameters (Browne, 1982) that offer no valuable contribution to the model and can conceal the lack of model fit. However the addition of these parameters can influence the size, significance and general interpretation of model fit (Brannick, 1995).

With these methodological issues in mind the results of the initial model without the addition of any data-driven modifications was considered. Based on the CFI and RMSEA without the addition of the error covariance between memory and visuospatial domains the one-factor model fitted the AD type and other frontotemporal lobe degenerative disorders groups poorly (AD type: $X^2 = 18.405$, $df = 5$; CFI = .957; RMSEA = .143, other frontotemporal lobe degenerative disorders: $X^2 = 40.327$, $df = 5$; CFI = .841; RMSEA = .269).

In the assessment of invariance only two groups were compared. The small sample sizes of these groups ($n=131$, $n=119$) raise the possibility of implying insufficient power to detect any differential domain functioning. Therefore the results of (i) analysis including the addition of an error covariance and subsequent assessment of invariance and (ii) the more conservative level of analysis without the addition of the error covariance or assessment of invariance will be discussed and interpreted.

4.3.3 Mokken Scaling Analysis

Alzheimer's type

Mokken's scalability coefficients were examined to assess the unidimensionality of the items. The H_i values of six items were below the recommended threshold level (0.3) for retaining items. These items; 'draw overlapping pentagons', 'draw a cube', 'count dot arrays', 'follow

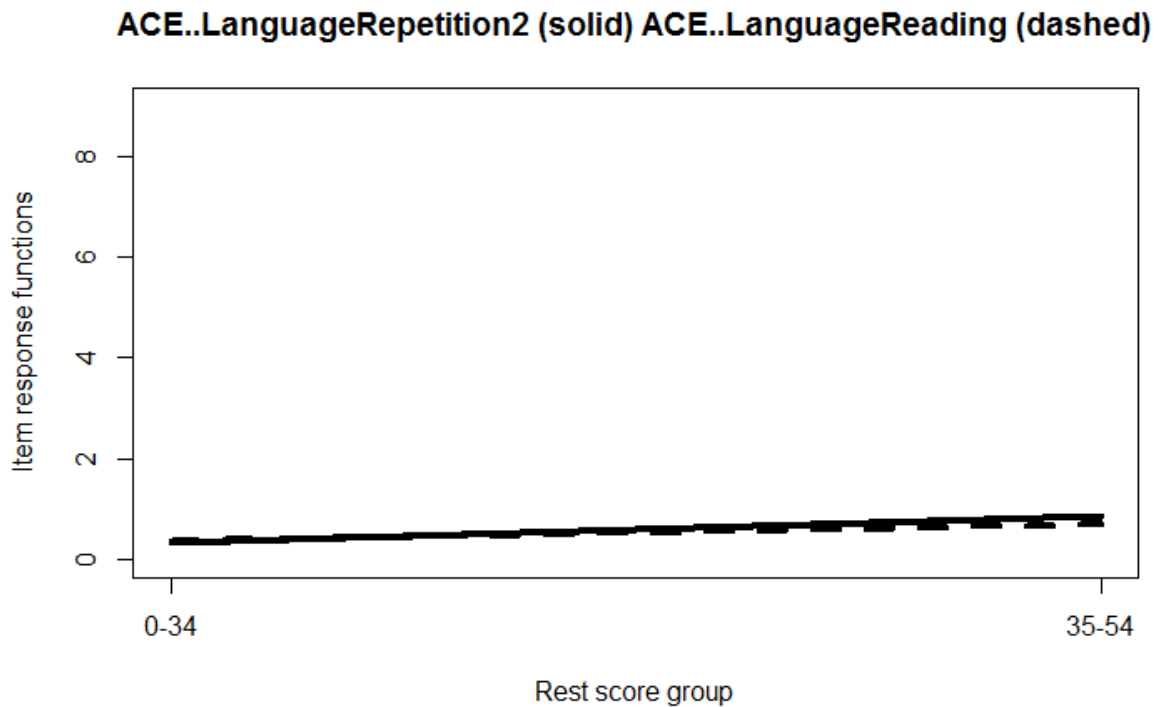
CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

written command-close eyes', 'three item recall', and 'repetition-no ifs, ands or buts' were removed. These low values suggest that the items have weak *discriminatory* power. There were no violations of monotonicity. Therefore the remaining 20 items were deemed sufficiently homogenous to be unidimensional on the basis of the item scalability coefficients and H of 0.45.

Assessment of IIO using method MIIO resulted in 32 violations, 16 of which were significant. Starting with the item with the greatest violation items were removed iteratively until no further violations remained. This process prompted the removal of a further six items ('identify fragmented letters', 'verbal fluency-animal', 'name and address learning', 'semantic comprehension', 'verbal fluency-letter', 'repetition of single multi-syllabic words'). The removal of these items resulted in 14 out of the original 26 items being retained in a moderately strong hierarchical Mokken scale ($H=0.44$, $SE=0.04$) with IIO ($H^T=0.69$). Inspection of item pair plots resulted in the further exclusion of three items; 'repetition-above, beyond and below' and 'reading' were shown to intersect (as shown in Figure 4.1) and 'naming 2' was identified as being located at some distance from the other items which could be driving the high H^T value (see Figure 4.2). The removal of these additional items left 11 items conforming to a moderate Mokken scale ($H=0.43$, $SE=0.04$) and lowered the strength of IIO ($H^T=0.52$).

Discriminatory values of items are presented in Table 4.4 in order of decreasing item scalability coefficients. Standard errors (SE) of scalability coefficients are also provided. These items, presented in Table 4.5 ordered according to their *difficulty* level (by mean score) have the same *difficulty* ordering irrespective of the value of the respondent's cognitive ability.

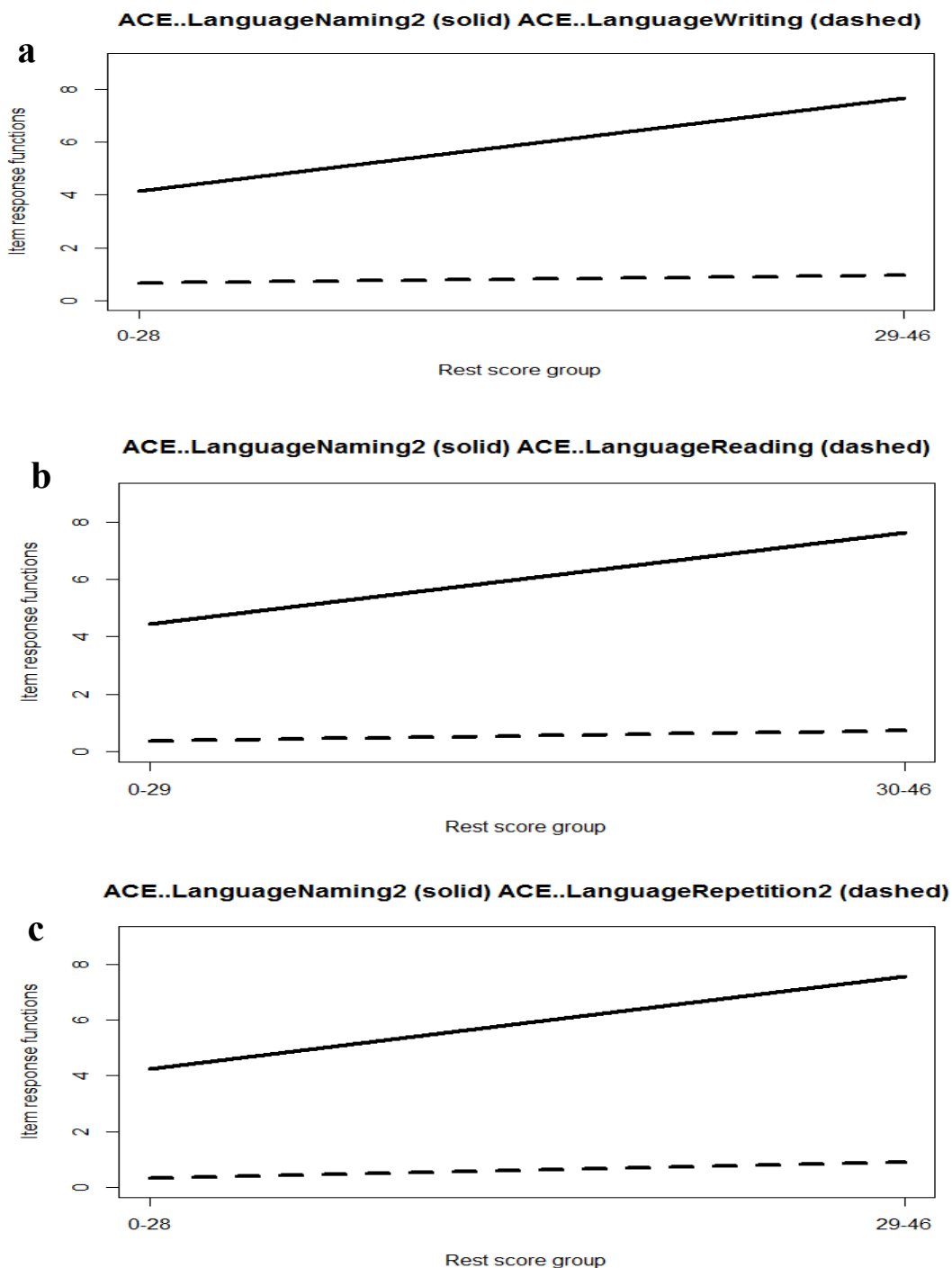
Figure 4.1 Item-pair plot demonstrating intersection between 'repetition-above, beyond and below' and 'reading'.



Note. X-axis reflecting the 'rest score group' \approx latent trait

Figure 4.2 Example of item-pair plot for 'naming 2' lying at some distance from a selection of remaining item-pair plots.

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCLTY



Note.

(a) Item-pair plots for 'naming 2' and 'write a sentence', (b) item-pair plots for 'naming 2' and 'reading', (c) item-pair plots for 'naming 2' and 'repetition-above, beyond and below'. X-axis reflecting the 'rest score group' \approx latent trait

Predominantly frontal dementia

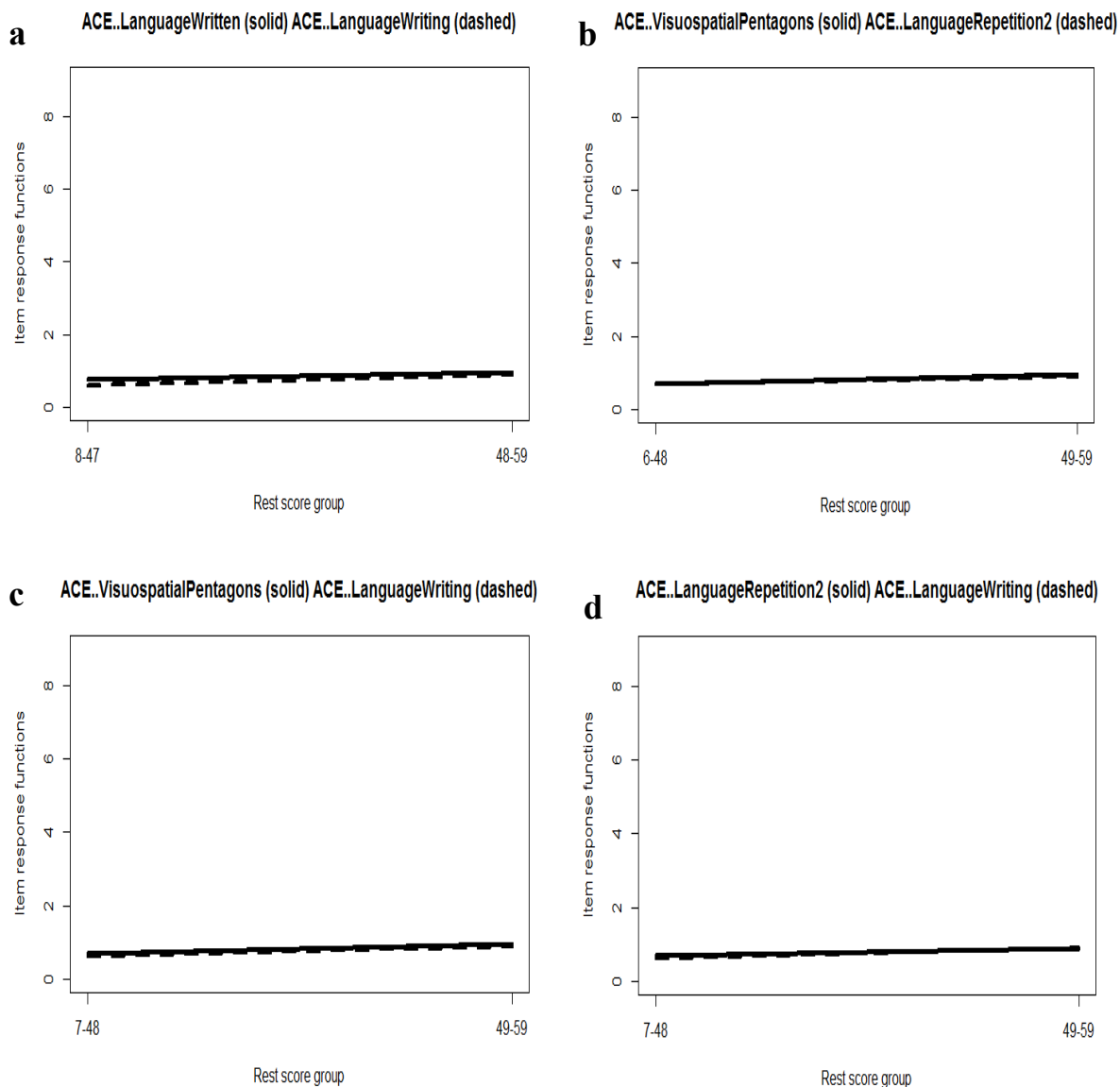
Four items were removed due to low H_i values; ‘draw a cube’, ‘repetition-no ifs, ands or buts’, ‘repetition of single multi-syllabic words’, and ‘reading’. There were no violations of monotonicity. The remaining 22 items were sufficiently homogenous to be considered unidimensional ($H=0.52$).

There were 42 violations of IIO using method MIIO, 32 of which were significant. This process resulted in the removal of six items (‘identify fragmented letters’, ‘three item registration’, ‘syntactical comprehension’, ‘verbal fluency-animal’, ‘verbal fluency-letter’, ‘name and address recall’). Following the removal of these items 16 items were retained in a strong Mokken scale ($H=0.52$, $SE=0.05$) with IIO ($H^T=0.82$).

Some intersection was observed from visual inspection of the item pair plots. This warranted the further exclusion of four items: ‘write a sentence’, ‘draw intersecting pentagons’, ‘repetition-above beyond and below’ and ‘follow written command-close eyes’ (see Figure 4.3). This left 12 items which conformed to a strong Mokken scale ($H=0.54$, $SE=0.05$) with a lowered strength of IIO ($H^T=0.72$) (see Table 4.4 for item ordering by *discrimination* and Table 4.5 for item ordering by *difficulty*).

Figure 4.3 Example of intersecting items from predominantly frontal dementia analysis.

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCTY



Note. Item pair-plots for (a) ‘follow written command-close eyes’ and ‘write a sentence’, (b) item pair-plots for ‘draw intersecting pentagons’ and ‘repetition-above, beyond and below’, (c) item pair-plot for ‘draw intersecting pentagons’ and ‘write a sentence’, (d) item-pair plots for ‘repetition-above, beyond and below’ and ‘write a sentence’. X-axis reflecting the ‘rest score group’ \approx latent trait

Other frontotemporal lobe degenerative disorders

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

Ten items were removed due to low H_i values ('follow written command-close eyes', 'repetition of single multi-syllabic words', 'repetition-above, beyond and below', 'repetition-no ifs, ands or buts', 'naming (10 items)', 'draw overlapping pentagons', 'draw a cube', 'count dot arrays, 'syntactical comprehension', 'semantic comprehension'). Again, there were no violations of monotonicity. The remaining 16 items were sufficiently homogenous to be considered unidimensional ($H=0.44$).

There were only two violations of IIO. This resulted in the exclusion of a further two items ('identify fragmented letters, 'naming (pencil and watch)'). Following the removal of these items 14 items were retained in a moderate Mokken scale ($H=0.45$, $SE=0.04$) with IIO, $H^T=0.65$. No further items were excluded following inspection of item plots. Items of this IIO subset are presented in order of *discrimination* (Table 4.4) and *difficulty* (Table 4.5).

IIO was also assessed using method MSCPM. This stronger method resulted in a greater number of items being removed. The IIO hierarchies revealed for each of the three groups using this more stringent method are listed in Table 4.6.

Table 4.4 Items listed in order of decreasing *discrimination* for each of the three groups

AD type			Predominantly frontal dementia			Other frontotemporal lobe degenerative disorders		
Domain: Item	H_i	SE	Domain: Item	H_i	SE	Domain: Item	H_i	SE
<i>Memory</i> : Name and address recall	0.51	0.05	<i>Visuospatial</i> : Draw a clock	0.60	0.05	<i>Memory</i> : Name and address recall	0.55	0.05
<i>Language</i> : Naming (pencil and watch)	0.49	0.06	<i>Memory</i> : Name and address learning	0.58	0.05	<i>Attention</i> : Orientation in time	0.54	0.05
<i>Attention</i> : Orientation in geography	0.47	0.05	<i>Memory</i> : Recognition	0.58	0.05	<i>Language</i> : Reading	0.53	0.09
<i>Memory</i> : Retrograde	0.46	0.05	<i>Language</i> : Naming (pencil and watch)	0.57	0.07	<i>Fluency</i> : Animal	0.52	0.05
<i>Language</i> : Write a sentence	0.46	0.07	<i>Attention</i> : Orientation in geography	0.57	0.06	<i>Memory</i> : Recognition	0.51	0.05
<i>Attention</i> : Serial sevens	0.44	0.05	<i>Attention</i> : Serial sevens	0.56	0.06	<i>Memory</i> : Name and address learning	0.48	0.05
<i>Attention</i> : 3 item registration	0.43	0.08	<i>Language</i> : Semantic comprehension	0.55	0.05	<i>Attention</i> : 3 item registration	0.45	0.08
<i>Visuospatial</i> : Draw a clock	0.42	0.05	<i>Attention</i> : Orientation in time	0.51	0.06	<i>Attention</i> : Orientation in geography	0.42	0.06
<i>Memory</i> : Recognition	0.41	0.05	<i>Language</i> : Naming (10 items)	0.50	0.06	<i>Visuospatial</i> : Draw a clock	0.41	0.05
<i>Language</i> : Syntactical comprehension	0.37	0.06	<i>Memory</i> : Retrograde	0.49	0.06	<i>Language</i> : Write a sentence	0.40	0.08
<i>Attention</i> : Orientation in time	0.35	0.06	<i>Memory</i> : 3 item recall	0.48	0.06	<i>Memory</i> : Retrograde	0.40	0.07
			<i>Visuospatial</i> : Count dot arrays	0.38	0.07	<i>Attention</i> : Serial sevens	0.39	0.07
						<i>Memory</i> : 3 item recall	0.37	0.07
						<i>Fluency</i> : Letter	0.34	0.06

Note. AD=Alzheimer’s disease, H_i =item scalability coefficient with higher values reflecting greater item *discrimination*. SE=Standard error.

Table 4.5 IIO hierarchies with items ordered from most to least *difficult* in each diagnostic group

Alzheimer's type		Predominantly frontal dementia		Other frontotemporal lobe degenerative disorders	
<i>Domain: Item</i>	Mean	<i>Domain: Item</i>	Mean	<i>Domain: Item</i>	Mean
<i>Memory: Name and address recall</i>	0.22	<i>Memory: 3 item recall</i>	0.50	<i>Language: Reading</i>	0.17
<i>Memory: Retrograde</i>	0.50	<i>Memory: Retrograde</i>	0.61	<i>Fluency: Animal</i>	0.19
<i>Visuospatial: Draw a clock</i>	0.68	<i>Language: Semantic comprehension</i>	0.70	<i>Memory: Name and address recall</i>	0.23
<i>Memory: Recognition</i>	0.69	<i>Memory: Recognition</i>	0.74	<i>Fluency: Letter</i>	0.29
<i>Attention: Orientation in time</i>	0.70	<i>Language: Naming (10 items)</i>	0.75	<i>Memory: Retrograde</i>	0.31
<i>Attention: Serial sevens</i>	0.77	<i>Attention: Orientation in time</i>	0.77	<i>Memory: 3 item recall</i>	0.36
<i>Language: Syntactical comprehension</i>	0.78	<i>Visuospatial: Draw a clock</i>	0.79	<i>Memory: Recognition</i>	0.65
<i>Attention: Orientation in geography</i>	0.79	<i>Attention: Serial sevens</i>	0.81	<i>Memory: Name and address learning</i>	0.66
<i>Language: Write a sentence</i>	0.84	<i>Attention: Orientation in geography</i>	0.82	<i>Language: Write a sentence</i>	0.69
<i>Language: Naming (pencil and watch)</i>	0.87	<i>Memory: Name and address learning</i>	0.83	<i>Visuospatial: Draw a clock</i>	0.71
<i>Attention: 3 item registration</i>	0.91	<i>Visuospatial: Count dot arrays</i>	0.85	<i>Attention: Orientation in geography</i>	0.71
		<i>Language: Naming (pencil and watch)</i>	0.92	<i>Attention: Serial sevens</i>	0.78
				<i>Attention: Orientation in time</i>	0.82
				<i>Attention: 3 item registration</i>	0.88
$H^T=0.52$		$H^T=0.72$		$H^T=0.65$	

Note. Mean scores reflect item *difficulty* with lower values indicating greater *difficulty*. $H^T=H$ trans with higher values indicating greater accuracy of ordering

Table 4.6 IIO hierarchies (using method MSCPM) with items ordered from most to least *difficult* for each diagnostic group.

Alzheimer's type		Predominantly frontal dementia		Other frontotemporal lobe degenerative disorders	
Item	Mean	Item	Mean	Item	Mean
Repetition-above, beyond and below	0.70	Memory Retrograde	0.61	Reading	0.17
Syntactical comprehension	0.78	Semantic comprehension	0.70	Verbal fluency-Animal	0.19
Write a sentence	0.84	Recognition	0.74	Memory retrograde	0.31
Three item registration	0.91	Draw a clock	0.79	Recognition	0.65
Identify fragmented letters	0.92	Repetition-above, beyond and below	0.80	Orientation in time	0.82
		Draw overlapping pentagons	0.80		
		Write a sentence	0.80		
		Orientation in geography	0.82		
		Follow written command-close eyes	0.90		
$H^T=0.92$		$H^T=0.87$		$H^T=0.89$	

Note. MSCPM= manifest scale - cumulative probability mode. Mean scores reflect item *difficulty* with lower values indicating greater *difficulty*. H^T = H trans with higher values indicating greater accuracy of ordering

4.4 Discussion

This chapter aimed to determine if hierarchies of ACE-R items meeting IIO criteria were present in three different samples consisting of different dementia diagnoses and to establish whether these hierarchies add anything to the subdomain scores. Mokken scaling analyses of the full 26 items of the scale for each of the three samples resulted in 11 items being retained in an IIO hierarchy in the Alzheimer's type sample, 12 items in the predominantly frontal dementia sample and 14 items in the other frontotemporal lobe degenerative disorders sample.

Fewer items were retained in IIO hierarchies using method MSCPM. This is not unusual as method MSCPM is a very strict method and we would expect this method to suggest the removal of many items. Although this stronger method is the preferred method from a methodological perspective as it may be better able to assess IIO of items with different numbers of categories the method may be too restrictive to be of practical use as it is extremely sensitive. With the exception of 'identify fragmented letters' in the AD type group all IIO items from the MSCPM method are drawn from the larger hierarchies revealed by method MIIO prior to the removal of further items following examination of item plots. The results of PCA did not indicate a difference between groups with all groups being dominated by a large single component with similar item loadings. Initial CFA and invariance analyses indicate that the structure of the one-factor model with an error covariance and the magnitude of the relationship between the observed variables and the latent constructs are invariant across the AD type and predominantly frontal dementia groups.

However due to concerns raised earlier the findings of an additional more conservative CFA were also examined. Here, without the addition of covariance among error

terms; memory and visuospatial, the CFA analysis indicated that the structure of the one-factor model did not fit two of the three groups; AD type and other frontotemporal lobe degenerative disorders. While this difference in model fit restricts the interpretability of factor analysis the results of Mokken scaling analysis present a mixed profile of cognitive decline for different patient groups.

In comparison with factor analysis Mokken scaling has considerable theoretical and practical advantages (DeJong & Molenaar, 1987). Whereas factor analysis identifies groups of highly correlating items Mokken scaling can illustrate the systematic order relationship between the items in a scale which improves construct validity (DeJong & Molenaar, 1987). Additionally factor loadings disregard how item performance may differ across levels of the latent trait (Meijer & Baneke, 2001). These advantages of Mokken scaling offer meaningful clinical implications. Establishing a formal hierarchy of item *difficulty* within a scale adds to the possibilities of interpretation and application. Hierarchical scales are appealing for their ease of use and scoring (Kempen, Myers & Powell, 1995). Responses to individual items, not just total scores, can provide an insight into a patient's level of ability based on the item's degree of *difficulty* (Watson, Deary & Shipley, 2008). For example, across the three diagnostic groups a patient responding correctly to the 'memory retrograde' item is unlikely to have problems with any of the less *difficult* items. This insight enables quicker estimations of a patient's cognitive functioning and can facilitate adaptive testing whereby only a selection of items, either from the more *difficult* or the less *difficult* range of the scale depending on the ability of the specific patient, is required for testing (van der Lee, Roorda, Beckerman, Lankhorst & Bouter, 2002). Tailoring tests to specific levels of ability can reduce testing time and stress and burden on patients.

Mokken scaling of dementia screening instruments can be used to assess whether the cognitive abilities are lost— or retained—hierarchically. Establishing whether these

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

hierarchies differ across diagnostic groups can be useful in differential diagnosis. While the more comprehensive CFA analyses in this chapter suggests the factor structure of the ACE-R domains may be invariant across the AD type and predominantly frontal diagnostic groups the results the more cautious approach taken with the removal of the error-covariance together with results from Mokken scaling suggests that there are differences in the ordering of items across groups. Although the comparison between the hierarchies is hampered by the lack of common items between the hierarchies there are several notable differences in the ordering of item *difficulty* among the common items between the groups. Looking at ordering of these items can provide some insight into the order of progressive decline in each group. Practically, if we examine the item ordering of the following items; ‘memory retrograde’, ‘orientation in time’, ‘orientation in geography’, ‘draw a clock’ and ‘write a sentence’ some different patterns emerge across the three groups: (i) if ‘write a sentence’ is less *difficult* than ‘draw a clock’ the patterns here suggest a diagnosis of AD or LPA is more likely than SD or PNFA, (ii) where ‘write a sentence’ is less *difficult* than ‘orientation in time’ the ordering here suggests AD type rather the semantic dementia and PNFA group, (iii) if ‘memory recognition’ is less *difficult* than ‘draw a clock’ these results indicate that a diagnosis of AD or LPA is most likely, and (iv) comparing ‘orientation in time’ with ‘orientation in geography’, if the score for ‘orientation in geography’ is lower than that for ‘orientation in time’, the most likely diagnosis from those considered in this sample one of the other frontotemporal lobe degenerative disorders; either semantic dementia or PNFA. Since both ‘orientation in time’ and ‘orientation in geography’ scores contribute to the attention and orientation sub-scale of the ACE-R, conventional comparisons between sub-scales are insensitive to this difference between diagnostic groups.

Mokken analysis demonstrates that the IIO hierarchies present a more mixed profile than what can be observed from mean subdomain scores with some items within the domains

being more *difficult* than others, for example in the predominantly frontal group ‘name and address learning’ is less *difficult* than the other Memory items. There is a wide spread of the Language items in terms of *difficulty* with Language items found among the most and least *difficult* items. These differences are not captured using domain sub-scores.

Some limitations of this study are important to consider when interpreting the results. Fundamentally, the sample size of each of the groups analysed here is relatively small, particularly the ‘other frontotemporal lobe degenerative disorders’ group ($n=100$). This sample is very small for all analyses in this study. The decision to include this small sample was made as the diagnoses within this group; semantic dementia and progressive nonfluent aphasia are relatively uncommon and accordingly large numbers of data are difficult to obtain. Therefore while data from this group were included results obtained must be interpreted with caution. Results from this sample and all analyses here require replication in a larger sample. It is also worth noting that the mean age across all three samples is rather low for dementia in general which suggests the results here may be more indicative of cognitive decline in young onset dementia. It would be worth replicating the results in an older sample.

The analysis of a larger total sample would have added greater power to detect any differential domain functioning in the assessment of invariance. It would also have been interesting to compare results from item level CFA and item level Mokken scaling analysis. However, the limited numbers available for analysis restricted the level of CFA permitted. This would be an interesting inclusion in a future study and could provide a better degree of comparability between the different methods.

There has been little research providing minimum sample size requirements for Mokken scaling until recently. A simulation study investigating adequate sample sizes for Mokken scaling determined that the strength of item scalability coefficients, H_i , is inversely

proportional to sample size which serves as a good indicator of adequate sample size (Straat, 2012). In this study's analysis of the smallest sample; 'other frontotemporal lobe degenerative disorders' ten items were excluded due to low H_i values. It is very likely that this is a consequence of the small number of participants in this sample. A larger sample size may have resulted in the inclusion of a greater number of items in the Mokken scale. This extends to all samples analysed here as there were exclusions due to low H_i coefficients in all samples analysed.

Due to the small samples all items excluded in the Mokken scaling process were made tentatively. In a larger sample it is likely that some items excluded from these analyses may well have been retained. With this in mind it is particularly pertinent to take the degree of uncertainty of estimated scalability coefficients into account when using Mokken's heuristic criteria to determine strength of scalability when sample size is low (Kuijpers, van der Ark & Croon, 2013). This degree of uncertainty can be assessed and quantified using the standard errors of the scalability coefficients. For small samples, i.e. below 100, it is important to acknowledge that where standard errors are high the chance of observing scaling error is high (Ringdal et al., 1999). Where the standard error of H is large (for example, 0.08) the probability of the value of H actually being less than 0.3 is reasonable which implies that the items within the scale are unscalable (Kuijpers, van der Ark & Croon, 2013). This extends to standard errors of item-pair and item scalability coefficients. Examining the standard errors of item scalability coefficients here suggests there is high likelihood of scaling errors for some items within each of the three IIO hierarchies.

While it is likely that some exclusions due to low H_i values was related to sample size the low H_i values of items excluded due to low *discrimination*; (e.g. draw overlapping pentagons', 'draw a cube', 'count dot arrays', 'follow written command-close eyes', repetition items and 'reading') could alternatively have been the result of these items

CHAPTER 4: ACE-R ITEM ORDERING BY DIFFIUCILITY

assessing some degree of hearing, vision or speech in addition to cognitive ability and therefore not measuring cognitive impairment as succinctly as the rest of the items. This insight can be used to remove items with poor sensitivity to the latent trait. This emphasises the value of Mokken scaling in assessing the performance of items within established scales. Item exclusions could also be the result of the similar levels of *difficulty* of some test items. When several items in a scale assess the same level of the latent trait IIO cannot be demonstrated (Ligtvoet, 2010; Watson, van der Ark, Lin, Fieo, Deary & Meijer, 2012). A narrow range of *difficulty* could result in item step response functions in close proximity to each other thus making violations of their non-intersection more probable as it is more likely that a particular pattern of response could differ from the expected pattern due to chance alone. While this may be seen as a limitation in the present context and may suggest that some items of the ACE-R could be removed, the similar degree of impairment assessed by some items may be considered an advantage in the context of the test design. As the ACE-R is a screening instrument it is important to increase the amount of information the instrument reveals about cognitive ability at the level of diagnostic threshold.

Finally, the heterogeneity of the sample could have influenced the number of items retained in the IIO hierarchies. As there were insufficient numbers for separate analyses by dementia diagnosis we formed three groups and performed PCA and CFA on the three samples. With PCA analysis showing no significant difference between the three groups the samples were analysed separately for comparison. While this is not ideal it resulted in three distinct groups for analysis. There were some difficulties in choosing the clinical groupings. Originally the groups were termed ‘Alzheimer’s type’ (AD and LPA), ‘Frontal dementia’ (bv-FTD and FTD-MND) and ‘Temporal dementia’ (SD and PNFA). However there were concerns regarding the classification of PNFA as temporal which prompted the labelling of the SD and PNFA group as ‘other frontotemporal lobe degenerative disorders’. While the

groups were formed on the basis of theoretical or structural similarities, (e.g. LPA is considered an atypical presentation of AD (Karantzoulis & Galvin, 2011) the groups are not as distinct as would have been preferred. Future studies should address this limitation by performing aetiology-stratified analyses on specific forms of dementia. Mokken scaling of specific aetiologies may result in a greater number of items retained and could help to establish the consistency and rate at which different abilities are lost in various different patient groups.

Furthermore, inconsistencies between individuals can be assessed using person-fit statistics such as PerFit (Tendeiro & Tendeiro, 2014). Person fit methods allow for the identification of unusual response to test items. Detecting unexpected score patterns could have valuable clinical implications and can be useful in improving the interpretation of test scores (Meijer & Tendeiro, 2014).

High H values (>0.50) of Mokken scales are very seldom reported. The strength of the Mokken scale for the predominantly frontal group ($H=0.54$) raises some concern regarding possible violations of LSI. In some cases elevated H values can reflect LSI violations (Egberink & Meijer, 2011). As discussed earlier, LSI arises where items are linked whereby the response to one item is dependent or impossible without prior response to another item. With regards to the ACE-R some items can be identified as possible sources of LSI violations. For example, '3-item recall' is linked to the earlier registration of the 3 words in '3-item registration' and similarly name and address recall and recognition is related to the initial learning of the name and address. There is a strong possibility that these items are not stochastically independent as it is logical that performance in delayed memory recall is predicated on performance on encoding and learning the information to be recalled. However it is not impossible that a patient could perform better in the recall stage than the learning or repetition stage due to motivational or attentional reasons. Performing Mokken scaling

analysis on the ACE-R excluding these five items could help determine the degree of LSI violations and the impact of these violations on the scalability of the ACE-R items.

4.5 Conclusion

This study of well-phenotyped participants analysed a well-established cognitive test applying novel and robust statistical techniques. The methods applied here, novel in their application to the ACE-R, yielded new and potentially significant findings relevant to both researchers and clinicians. Replication studies of larger samples are required with the present results from this analysis interpreted with caution due to sample size limitations. Mokken scaling analyses applied concurrently with factor analytic methods can provide additional information, offering prognostic value to clinicians assessing patients. A full neuropsychological assessment is the gold standard in assessing cognitive impairment but the ACE-R is frequently used not only to determine the degree of cognitive impairment, but also to inspect the extent of sub-domain deficits to assist differential diagnosis. With different outcomes from factor analysis a further study with sufficient numbers for item level CFA and assessment of invariance is necessary. Despite this limitation it is important to note that the IIO revealed by Mokken scaling suggests a more complex pattern of decline. While further studies are required to further delineate the item orderings in sufficiently large distinct diagnostic groups clinical assessments should expand on merely looking at total scores but should consider the patterns of responses, in particular the order in which the items are failed.

Chapter 5: Mokken scaling analysis in the development of the Mini-Addenbrooke's Cognitive Examination: a new assessment tool for dementia

5.1 Introduction

This chapter is based on my collaboration with the Frontier Research Group at Neuroscience Research Australia (NeuRA) in Sydney where I assisted in the development of a new brief scale. My role in the development of the scale involved the analysis of data collected by the Frontier Research Group, Sydney, the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom. The aim of this analysis was the identification and selection of candidate items from the ACE-III for a new brief scale.

Short cognitive screening tests are particularly valuable in clinical settings (which are often limited in time and resources) as they can be used to identify people who should be targeted for more comprehensive assessments. The ACE-R and its successor the ACE-III are widely used and well validated in both clinical and research settings worldwide (Mioshi et al., 2006, Pigiautile et al., 2011, Fang et al., 2013, Hsieh, Schubert, Hoon, Mioshi, & Hodges, 2013). The most recent version, the ACE-III, has been validated against standard neuropsychological measures as a valuable assessment scale for the detection, differentiation and monitoring of cognitive dysfunction in AD and FTD (Hsieh et al., 2013). However, the administration of the ACE-III takes at least 15-20 minutes, which may stretch the available resources of busy clinic settings. The aim of this analysis was to use data driven scaling methods to acquire additional psychometric information and to use this information to guide item selection for a new brief screening tool. The development of an abbreviated test permitting rapid assessment would have significant clinical relevance as a simple, brief and portable test.

Mokken scaling analysis can be used in test design or in the construction of multi-item questionnaires measuring health constructs (Sijtsma et al., 2008). Mokken scaling is particularly useful in the construction of unidimensional and hierarchical scales (De Jong & Molenaar, 1987; Kempen & Suurmeijer, 1990). Mokken scaling analysis also has useful applications for investigating the suitability and performance of established scales by examining the behaviour of items in response to different levels of the latent trait. With regards to scale development Mokken scaling can be used to identify and select items that *discriminate* well in specific populations.

A good instrument for the assessment of cognitive ability in dementia ideally consists of items with high *discrimination* values of varying *difficulty* levels ensuring sensitive measurement at varying levels of cognitive ability. In this way the instrument would be capable of efficient measurement of individual levels of cognitive ability as well as identifying small differences between patient groups, and responses to interventions and treatment. Taking item *discrimination* and *difficulty* into account can help contribute to this goal of developing an ideal scale.

High levels of *discrimination* are desirable as the higher the *discrimination* the greater the item's contribution to reliable measurement of cognitive impairment (Hambelton & Swaminathan, 1985). Some items within scales only *discriminate* within specific ranges of the latent trait (Reise & Waller, 2009; Meijer & Baneke, 2004). This has significant implications for both researchers and clinicians measuring change in response to treatment or change over time. A scale consisting largely of items with poor *discrimination* across the low *difficulty* range would be likely to fail to detect any effect of the therapy in a clinical sample. Similarly detecting early symptoms in a generally healthy population relies on items that *discriminate* well in the level of latent trait assessed by high *difficulty* items. Without such items a scale would most likely be unable to detect and monitor the early stages of disease

onset. For this reason IRT methods, which take item *discrimination* into account, can be used to guide item selection ensuring that all items *discriminate* well in all required ranges of the latent trait.

To develop a scale with high reliability at a particular level of latent trait highly *discriminatory* items with *difficulty* values assessing the desired level of ability must be identified (Meijer & Baneke, 2004). Scalability statistics provided from Mokken scaling analysis along with graphical analysis of item response functions from TestGraf graphical analysis (Ramsay, 2000) can be used to assess such item properties. On this basis Mokken scaling can identify key items which could be included in a shortened screening tool. This chapter aims to use these techniques and methods to derive a short scale—the Mini-Addenbrooke’s Cognitive Examination (Mini-ACE)—from the ACE-III and to evaluate the performance of this new measure in an independent validation sample.

While these data driven methods are used to identify the most appropriate items across a range of cognitive domains it should also be noted that prior to this analysis a theoretical item selection had been made by the research team in Sydney due to their content and applicability. This selection comprised; ‘name and address learning’, ‘orientation in time’, ‘verbal fluency-letter’, ‘draw a clock’ and ‘name and address recall’. Therefore a secondary consideration in this analysis is to determine whether the Mokken scaling results support the inclusion of these items.

5.2 Method

5.2.1 Participants

Mini-ACE development sample

Participants were recruited from the Frontier Research Group, Sydney, Australia, the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom. The sample of 117 participants included a variety of aetiologies: Alzheimer's disease (AD) ($n=34$), behavioural variant frontotemporal dementia (bvFTD) ($n=25$), corticobasal degeneration (CBD) ($n=9$), progressive primary aphasia (PPA) ($n=49$).

A multidisciplinary team assessed patients and diagnosis was established in line with the current diagnostic criteria (Mathew, Bak & Hodges, 2012, McKhann et al., 2011, Rascovsky et al., 2011, Gorno-Tempini et al., 2011) based on extensive clinical assessments, comprehensive neuropsychological assessment, and evidence of atrophy on structural MRI brain scans. Data from all 117 patients with complete itemised ACE-III data were included in the analysis i.e. including patients with a range of dementia aetiologies. Informed consent was obtained from all participants or a carer where necessary.

Mini-ACE validation sample

This sample, previously described in Chapter 4, comprised 350 patients also from the Frontier Research Group clinic in Sydney with a clinical dementia diagnosis (AD, $n=88$; bvFTD, $n=96$; frontotemporal dementia with motor neurone disease (FTD-MND), $n=23$; logopenic progressive aphasia (LPA), $n=43$, PNFA, $n=39$; semantic dementia (SD), $n=61$). Table 5.1 presents a comparison of the development and validation samples.

Table 5.1 Comparison of the scale development and validation samples

	Mini-ACE development sample	Mini-ACE validation sample
N	117	350
Age (SD)	65.4 (8.5)	65.4 (8.5)
ACE (SD)	63.6 (20)	63.6 (19.4)
Location	Sydney, Oxford, Cambridge	Sydney
Scale	ACE-III	ACE-R
Diagnosis	AD (34), bvFTD (25), CBD (9), PPA (49)	bvFTD (96), AD (88), SD (61), LPA (43), PNFA (39), FTD-MND (23)

Note. Mini-ACE=Mini Addenbrooke's Cognitive Examination, SD=standard deviation, AD=Alzheimer's disease, bvFTD=behavioural variant frontotemporal dementia, CBD= corticobasal degeneration, PPA= progressive primary aphasia, SD=semantic dementia, LPA=Logopenic progressive aphasia, PNFA= progressive nonfluent aphasia, FTD-MND= frontotemporal dementia with motor neurone disease.

5.2.2 Measures

The ACE-III is scored out of 100 and includes 24 items assessing five cognitive domains: attention (18 points), memory (26 points), fluency (14 points), language (26 points), visuospatial (16 points). The scores of each domain are summed to create the total ACE-III score (see Appendix A for a sample of the ACE-III scale).

As there are different score ranges for the ACE-III items (i.e. of the 24 items four had a maximum score of one, three had a maximum score of two, three had a maximum score of three, four had a maximum score of four, five had a maximum score of five, four had a maximum score of seven and one had a maximum of 12) the mean for each ACE-III item score was divided by the maximum number of points available for that item to equate scores for comparison, giving a score with minimum 0 and maximum of 1. For example a mean score of 3.72 for 'orientation in time' was divided by five to produce an equated mean of 0.74.

‘Naming’ with a range of 0-12 was rescaled as 0-9 as the Mokken scaling procedure does not accommodate values above 9. While this rescoring potentially removes some significant variance in responses this was minimized by collapsing the most infrequent item responses; scores of 2, 5 and 7, where the prevalence of responses is very low ($n=6$, 5%).

5.2.3 Analyses

Mokken scaling

Data were entered into the freeware R environment (R Development Core Team) using the “Mokken” package (van der Ark, 2011). The data were explored to assess their unidimensionality and monotonicity before further analyses of IIO were carried out. R’s default method manifest invariant item ordering (MIIO) was applied in the assessment of IIO. Sensitivity analysis was carried out, in which each diagnostic group, one at a time, was excluded from the overall sample (i.e. Mokken scaling was performed on data excluding patients with Alzheimer’s disease, excluding patients with semantic dementia etc.).

Subsequent Mokken scaling analysis was performed on the items chosen as candidate items for the new scale in an independent clinical sample. Data from the sample described in Chapter 4 were used for this analysis.

Graphical analysis-TestGraf

The TestGraf programme (Ramsay, 2000) was used to graphically illustrate item and option effectiveness to aid item selection. TestGraf is a non-parametric-based IRT model that can be used to assess item performance as a function of the latent trait, providing visual illustrations of item and option functioning. Firstly, TestGraf estimates the probability of a specific item option being selected by ranking respondents according to some statistic, typically the summed total score. These rank values are then converted to standard normal scores to establish an estimate of a respondent’s position along the latent trait. Uniformly spaced

evaluation points along the standard normal distribution provide estimates of response curves (Sachs, Law and Chan, 2003). For each respondent, dichotomous values denoting whether or not an option was selected are weighted. A kernel function is used to define the weights. While many respondents contribute to the estimation of the curves the values of those falling near or at an evaluation point are most heavily weighted (Santor & Ramsay, 1998). The options most frequently responded to with the heaviest weights will exert greater influence on the expected item scores. Items with options that are seldom endorsed will not produce expected scores that increase in line with increasing degree of the latent trait (e.g. cognitive impairment). Therefore expected item scores, or the rate of increase of these scores, reflect the sensitivity of the item to the assessment of the latent trait. Summing across expected item scores at evaluation points provides estimated expected total scores. The items and options within with greater sensitivity to changes in latent trait have a greater effect on the expected total score than items with poorer sensitivity. TestGraf calculates an expected total score for each respondent using maximum likelihood estimation methods (Santor, Ramsay & Zuroff, 1994).

The TestGraf program was used to estimate item characteristic curves (ICCs) for the scale items. These ICCs graphically present the expected score of a given ACE-III item as a function of overall cognitive impairment. ICCs were calculated using a Gaussian kernel smoothing technique that calculates the expected item score based on the overall ACE-III distribution of scores.

5.3 Results

One hundred and seventeen participants (60% male) from the Frontier Research Group, Sydney, Australia, the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom with a mean age of 65.4 years ($SD = 8.5$) were included in the Mini-ACE

development sample. The validation sample comprised 350 participants (66% male, mean age 65.4, SD = 8.5) also patients at the Frontier Research Group, Sydney, Australia.

Table 5.1 gives descriptive data on the development and validation samples, and Table 5.2 gives details of the scores within the ACE-III in the development sample.

5.3.1 Mokken scaling analysis

Assessing the unidimensionality of the data showed that all item pair scalability coefficients (H_{ij}) were positive. However the scalability coefficients (H_i) of five of the 24 ACE-III items fell below the 0.3 threshold level ('repetition of single multi-syllabic words', 'count dot arrays', 'draw intersecting infinity loops', 'reading', 'draw a cube') indicating that in this population sample, these items are not homogenous (correlated) enough for inclusion in the scale. Therefore I chose to exclude these items.

No items were excluded in the assessment of monotonicity. The remaining 19 items were sufficiently homogeneous to be considered to be a moderate unidimensional Mokken scale based on the H_i levels combined with H of 0.45.

Finally, in the assessment of IIO, two items ('syntactical comprehension' and 'three item registration') were removed due to violations (intersecting IRFs). The remaining 17 items formed a very reliable (Molenaar Sijtsma (MS, Sijtsma & Molenaar, 1987) statistic=0.91) moderate hierarchical Mokken scale ($H=0.44$) with IIO ($H^T=0.61$). This IIO means the items can be ordered according to their *difficulty* level and items have the same *difficulty* ordering irrespective of the value of the respondent's cognitive ability. This hierarchical subset of ACE-III items is listed in Table 5.3 in descending order of *difficulty* and *discrimination*.

Table 5.2 Raw and equated mean ACE-III item scores for the Mini-ACE development sample ($N=117$) by cognitive domain

Item	Domain	Label	Mean score	Maximum score	Equated Mean
1	Attention	Orientation in time	3.72	5	0.74
2		Orientation in geography	3.75	5	0.75
3		Three item registration	2.79	3	0.93
4		Serial sevens	2.88	5	0.58
5	Memory	Three item recall	1.40	3	0.47
8		Name and address learning	5.44	7	0.78
9		Memory retrograde	2.14	4	0.53
23		Name and address recall	2.28	7	0.33
24		Recognition	3.72	5	0.74
6	Fluency	Verbal fluency-letters	3.20	7	0.46
7		Verbal fluency-animals	2.38	7	0.34
10	Language	Syntactical comprehension	2.37	3	0.79
11		Write two sentences	1.08	2	0.54
12		Repetition of single multi-syllabic words	1.40	2	0.70
13		Repetition-all that glitters is not gold	0.80	1	0.80
14		Repetition-a stitch in time saves nine	0.72	1	0.72
15		Naming	6.15	9*	0.68
16		Semantic comprehension	2.64	4	0.66
17		Reading	0.49	1	0.49
18		Visuospatial	Draw intersecting infinity loops	0.57	1
19	Draw a cube		1.22	2	0.61
20	Draw a clock		3.19	5	0.64
21	Count dot arrays		3.28	4	0.82
22	Identify fragmented letters		3.63	4	0.91
	Domain	Attention (/18)	13.09		
		Memory (/26)	14.97		
		Fluency (/14)	5.58		
		Language (/26)	18.00		
		Visuospatial (/16)	11.96		
		Total (/100)	63.60		

Note. Item=order of item in ACE-III administration, Equated mean=mean item score divided by maximum score available for each item to provide a range of 0-1 for each item with lower scores indicating greater item *difficulty*. *'Naming' maximum score of 12 rescaled to maximum of 9 for analysis.

Table 5.3 Hierarchical subset of ACE-III items revealed by Mokken scaling analysis of the Mini-ACE development sample. Items ordered from most to least *difficult* and most to least *discriminatory*

Difficulty		Discrimination	
Item	Mean	Item	H_i
Name and address recall	0.33	Identify fragmented letters	0.54
Verbal fluency-animal	0.34	Name and address learning	0.52
Verbal fluency-letters	0.46	Name and address recall	0.51
Three item recall	0.47	Verbal fluency-animal	0.51
Memory retrograde	0.53	Orientation to geography	0.50
Write two sentences	0.54	Recognition	0.47
Serial sevens	0.58	Memory retrograde	0.47
Draw a clock	0.64	Naming	0.45
Semantic comprehension	0.66	Three item recall	0.44
Naming	0.68	Orientation in time	0.44
Repetition-a stitch in time saves nine	0.72	Draw a clock	0.42
Recognition	0.74	Serial sevens	0.38
Orientation in time	0.74	Verbal fluency-letters	0.37
Orientation in geography	0.75	Repetition-all that glitters is not gold	0.36
Name and address learning	0.78	Semantic comprehension	0.35
Repetition-all that glitters is not gold	0.80	Write two sentences	0.32
Identify fragmented letters	0.91	Repetition-a stitch in time saves nine	0.32

Note. Mean=mean equated item score with lower values indicating higher *difficulty*. H_i =item scalability coefficient reflecting item *discrimination* with higher values reflecting greater *discrimination*.

5.3.2 Sensitivity analysis

As this analysis was carried out on a heterogeneous sample of patients with various dementia syndromes six separate analyses were performed each excluding one of the patient groups in order to determine whether one of the groups was driving the results. Due to the relatively small sample size, controls ($n=30$) were included in the sensitivity analyses to ensure sufficient numbers in each analysis. Control data were collected by the Frontier Research Group, Sydney. Individual groups were formed by the addition of control data and removal of one of each of the diagnostic groups at a time (e.g. the ‘Minus bvFTD group’ includes control and clinical data minus all patients diagnosed with bvFTD). Table 5.4 presents the mean equated item scores for each group.

Table 5.4 Equated means of ACE-III items and subdomain scores for each group in sensitivity analysis presented in order of the ACE-III

Items	Clinical sample (n=117)	Clinical sample plus controls (n=147)	Minus bvFTD (n=122)	Minus AD (n=113)	Minus SD (n=126)	Minus LPA (n=134)	Minus PNFA (n=134)	Minus CBD (n=138)
Orientation in time	0.74	0.79	0.79	0.85	0.78	0.81	0.78	0.75
Orientation in geography	0.75	0.80	0.80	0.83	0.82	0.82	0.78	0.79
3 item registration	0.93	0.94	0.94	0.95	0.95	0.96	0.94	0.94
Serial sevens	0.58	0.64	0.65	0.68	0.61	0.68	0.63	0.64
Memory Recall	0.47	0.56	0.55	0.64	0.59	0.58	0.53	0.54
Fluency-letter	0.46	0.55	0.56	0.54	0.56	0.58	0.57	0.55
Fluency-animal	0.34	0.44	0.46	0.46	0.47	0.48	0.44	0.44
Name and address learning	0.78	0.82	0.82	0.84	0.81	0.86	0.80	0.80
Memory retrograde	0.53	0.61	0.60	0.64	0.66	0.64	0.59	0.59
Syntactical comprehension	0.79	0.83	0.83	0.83	0.83	0.86	0.85	0.83
Write sentences	0.54	0.63	0.64	0.65	0.61	0.66	0.63	0.64
Repetition 1	0.70	0.76	0.77	0.75	0.73	0.79	0.79	0.77
Repetition 2	0.80	0.84	0.84	0.81	0.86	0.87	0.86	0.83
Repetition 3	0.72	0.77	0.78	0.75	0.78	0.82	0.86	0.77
Naming	0.68	0.74	0.73	0.73	0.81	0.77	0.73	0.73
Semantic comprehension	0.66	0.72	0.73	0.71	0.79	0.72	0.70	0.70
Reading	0.49	0.59	0.56	0.57	0.66	0.63	0.61	0.58
Draw intersecting infinity loops	0.57	0.65	0.65	0.71	0.60	0.67	0.63	0.67
Draw a cube	0.61	0.68	0.69	0.76	0.64	0.70	0.67	0.71
Draw a clock	0.64	0.71	0.69	0.75	0.71	0.73	0.70	0.72
Count dot arrays	0.82	0.84	0.87	0.86	0.82	0.85	0.83	0.86
Identify fragmented letters	0.91	0.93	0.92	0.94	0.93	0.93	0.92	0.93
Name and address recall	0.33	0.43	0.42	0.52	0.47	0.47	0.40	0.41
Recognition	0.74	0.79	0.79	0.84	0.79	0.80	0.77	0.77
Subdomain scores								
Attention (/18)	13.09	13.98	13.96	14.65	13.85	14.43	13.76	13.83
Memory (/26)	14.97	16.87	16.75	18.27	17.39	19.00	16.28	16.40
Fluency (/14)	5.58	6.96	7.18	7.02	7.23	7.45	7.07	6.91
Language (/26)	18.00	19.54	19.43	19.27	20.65	20.15	19.45	19.30
Visuospatial (/16)	11.96	12.69	12.69	13.27	12.49	12.94	12.50	12.79
Total (/100)	63.60	70.04	70.02	72.48	71.60	72.54	69.05	69.23

Note. bvFTD=behavioural variant frontotemporal dementia, AD=Alzheimer's disease, SD=semantic dementia, LPA=Logopenic progressive aphasia, PNFA=progressive nonfluent aphasia, CBD=corticobasal degeneration. Repetition 1=Repetition of single multisyllabic words, Repetition 2=Repetition-all that glitters is not gold, Repetition 3=Repetition-a stitch in time saves nine

The fit of each group to the two Mokken models was determined.

- (i) All diagnostic groups plus controls ($n=147$)

The H_i value of one of the 24 items; ‘count dot arrays’ fell below 0.3. As this indicated poor *discrimination* and scalability this item was removed. Following the removal of this item the remaining 23 items were sufficiently homogeneous to be considered unidimensional. H for this subset of 23 items was 0.52.

Five items were removed due to IIO violations; ‘three item registration’, ‘syntactical comprehensions’, ‘semantic comprehension’, ‘repetition of single multi-syllabic words’ and ‘identify fragmented letters’. The 18 remaining items were retained in a strong and reliable (MS=0.94) hierarchical Mokken scale ($H=0.53$) with high accuracy of item ordering ($H^T=0.73$).

- (ii) Diagnostic groups minus bvFTD, plus controls ($n=122$)

The H_i value of one of the 24 items fell below the 0.3 threshold level (‘count dot arrays’). Following the removal of this item the remaining 23 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.53.

In the assessment of IIO five items were removed due to violations (‘name and address recall’, ‘verbal fluency-animal’, ‘verbal fluency-letter’, ‘repetition of single multi-syllabic words’, ‘semantic comprehension’). The remaining 18 items formed a strong and reliable (MS=0.92) hierarchical Mokken scale ($H=0.50$) with IIO ($H^T=0.78$).

- (iii) Diagnostic groups minus AD, plus controls ($n=113$)

CHAPTER 5: DEVELOPMENT OF MINI-ACE

The H_i value of one of the 24 items fell below the 0.3 threshold level ('count dot arrays').

Following the removal of this item the remaining 23 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.57.

In the assessment of IIO six items were removed due to violations ('verbal fluency-animal', 'name and address recall', 'verbal fluency-letter', 'identify fragmented letters', 'memory retrograde', 'semantic comprehension'). The remaining 17 items formed a strong and reliable (MS=0.92) hierarchical Mokken scale ($H=0.52$) with IIO ($H^T=0.83$).

(iv) Diagnostic groups minus SD, plus controls ($n=126$)

None of the H_i values fell below the 0.3 level; all 24 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.54.

In the assessment of IIO five items were removed due to violations ('identify fragmented letters', 'count dot arrays', 'three item registration', 'semantic comprehension', 'syntactical comprehension'). The remaining 19 items formed a strong hierarchical Mokken scale ($H=0.56$) with IIO ($H^T=0.77$) (MS=0.94).

(v) Diagnostic groups minus LPA, plus controls ($n=134$)

The H_i of three of the 24 items fell below the 0.3 threshold level ('repetition of single multisyllabic words', 'count dot arrays' and 'draw intersecting infinity loops'). Following the removal of these items the remaining 21 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.51.

In the assessment of IIO five items were removed due to violations ('verbal fluency-animal', 'name and address recall', 'three item registration', 'identify fragmented letters', 'syntactical comprehension'). The remaining 16 items formed a strongly reliable (MS=0.91) moderate hierarchical Mokken scale ($H=0.46$) with IIO ($H^T=0.78$).

CHAPTER 5: DEVELOPMENT OF MINI-ACE

(vi) Diagnostic groups minus PNFA, plus controls ($n=134$)

The H_i value of one of the 24 items fell below the 0.3 threshold level ('count dot arrays').

Following the removal of these items the remaining 23 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.54.

In the assessment of IIO six items were removed due to violations ('name and address recall', 'verbal fluency-letter', 'verbal fluency-animal', 'three item registration', 'repetition of single multisyllabic words', 'syntactical comprehension'). The remaining 17 items formed a moderate hierarchical Mokken scale ($H=0.49$) with IIO ($H^T=0.78$) ($MS=0.91$).

(vii) Diagnostic groups minus CBD, plus controls ($n=138$)

None of the H_i values fell below the 0.3 level; all 24 items were sufficiently homogeneous to be considered unidimensional based on the H_i levels combined with H of 0.52.

In the assessment of IIO six items were removed due to violations ('three item registration', 'count dot arrays', 'semantic comprehension', 'syntactical comprehension', 'identify fragmented letters', 'repetition of single multisyllabic words'). The remaining 18 items formed a strong and reliable ($MS=0.94$) hierarchical Mokken scale ($H=0.55$) with IIO ($H^T=0.72$).

The hierarchies from each of these analyses are listed in Table 5.5 from most to least *difficult* based on mean scores. Due to different items being retained in the final hierarchies across the analyses the nine items common ('write sentences', 'draw a clock', 'naming', 'repetition3: a stitch in time saves nine', 'recognition', 'orientation in time', 'orientation in geography', 'name and address learning', 'repetition 2: all that glitters is not gold') to all were examined and compared (see Table 5.6).

CHAPTER 5: DEVELOPMENT OF MINI-ACE

Across the eight analyses the ordering of the items common to all hierarchies is generally similar with ‘repetition- all that glitters is not gold’, ‘name and address learning’ and ‘orientation in geography’ among the least *difficult* items in several of the analyses and ‘write sentences’, ‘draw a clock’ and ‘naming’ among the most *difficult* items across samples.

While the ordering of the smallest subsample excluding AD patients differs slightly from the general trend of *difficulty* for the less *difficult* items (e.g. ‘orientation for time’ is the least *difficult* item for this sample whereas it is consistently more *difficult* for the other samples) the pattern is more consistent for the most *difficult* items.

These similarities in ordering plus the exclusion of many of the same items across samples (e.g. ‘count dot arrays’, ‘semantic comprehension’ and ‘identify fragmented letters’) indicated that no one group was driving the results of the main analysis. Therefore the results of the analysis of the original clinical sample ($N=117$) and the items retained in the IIO hierarchy can form the basis for item selection for the Mini-ACE.

Table 5.5 IIO hierarchies from sensitivity analyses: ordered from most to least *difficult*

Clinical sample (<i>n</i> =117)	Clinical sample plus controls (<i>n</i> =147)	Minus bvFTD (<i>n</i> =122)	Minus AD (<i>n</i> =113)	Minus SD (<i>n</i> =126)	Minus LPA (<i>n</i> =134)	Minus PNFA (<i>n</i> =134)	Minus CBD (<i>n</i> =138)
Name and address recall	Name and address recall	3 item recall	Reading	Name and address recall	3 item recall	3 item recall	Name and address recall
Fluency-animal	Fluency-animal	Reading	3 item recall	Fluency-Animal	Fluency-Letter	Memory retrograde	Fluency-Animal
Fluency-letters	Fluency-letters	Memory retrograde	Write sentences	Fluency-Letter	Reading	Reading	3 item recall
Three item recall	3 item recall	Write sentences	Serial sevens	3 item recall	Memory retrograde	Draw intersecting infinity loops	Fluency-Letter
Memory retrograde	Reading	Serial sevens	Draw infinity loops	Draw intersecting infinity loops	Write sentences	Write sentences	Reading
Write sentences	Memory retrograde	Draw intersecting infinity loops	Naming	Write sentences	Serial sevens	Serial sevens	Memory retrograde
Serial sevens	Write sentences	Draw a cube	Draw a clock	Serial sevens	Draw a cube	Draw a cube	Serial sevens
Draw a clock	Serial sevens	Draw a clock	Repetition 3	Draw a cube	Semantic comprehension	Draw a clock	Write sentences
Semantic comprehension	Draw intersecting infinity loops	Naming	Repetition 1	Reading	Draw a clock	Semantic comprehension	Draw intersecting infinity loops
Naming	Draw a cube	Repetition 3	Draw a cube	Memory retrograde	Naming	Naming	Draw a cube
Repetition 3	Draw a clock	Recognition	Repetition 2	Draw a clock	Recognition	Recognition	Draw a clock
Recognition	Naming	Orientation in time	Syntactical comprehension	Repetition 1	Orientation in time	Orientation in time	Naming
Orientation in time	Repetition 3	Orientation in geography	Orientation in geography	Repetition 3	Repetition 3	Orientation in geography	Repetition 3
Orientation in geography	Recognition	Name and address learning	Recognition	Orientation in time	Orientation in geography	Repetition 3	Recognition
Name and address learning	Orientation in time	Syntactical comprehension	Name and address learning	Recognition	Name and address learning	Name and address learning	Orientation in time
Repetition 2	Orientation in geography	Repetition 2	Orientation in time	Naming	Repetition 2	Repetition 2	Orientation in geography
Identify fragmented letters	Name and address learning	Identify fragmented letters	3 item registration	Name and address learning		Identify fragmented letters	Name and address learning
	Repetition 2	3 item registration		Orientation in geography			Repetition 2
				Repetition 2			
<i>H</i> =0.44	<i>H</i> =0.53	<i>H</i> =0.50	<i>H</i> =0.52	<i>H</i> =0.56	<i>H</i> =0.46	<i>H</i> =0.49	<i>H</i> =0.55
<i>H^T</i> =0.61	<i>H^T</i> =0.73	<i>H^T</i> =0.78	<i>H^T</i> =0.83	<i>H^T</i> =0.77	<i>H^T</i> =0.78	<i>H^T</i> =0.78	<i>H^T</i> =0.72

Note. *H*=scale scalability coefficient with higher values indicating greater strength of Mokken scale. *H^T*=*H* trans with higher values reflecting greater accuracy of item ordering. bvFTD=behavioural variant frontotemporal dementia, AD=Alzheimer's disease, SD=semantic dementia, LPA=Logopenic progressive aphasia, PNFA=progressive nonfluent aphasia, CBD=corticobasal degeneration. Repetition 1=Repetition of single multisyllabic words, Repetition 2=Repetition-all that glitters is not gold, Repetition 3=Repetition-a stitch in time saves nine

Table 5.6 Comparison of items common to all hierarchies from sensitivity analysis: ordered from most to least *difficult*

Clinical sample (<i>n</i> =117)	Clinical sample plus controls (<i>n</i> =147)	Minus bvFTD (<i>n</i> =122)	Minus AD (<i>n</i> =113)	Minus SD (<i>n</i> =126)	Minus LPA (<i>n</i> =134)	Minus PNFA (<i>n</i> =134)	Minus CBD (<i>n</i> =138)
Write sentences	Write sentences	Write sentences	Write sentences	Write sentence	Write sentences	Write sentences	Write sentences
Draw a clock	Draw a clock	Draw a clock	Naming	Draw a clock	Draw a clock	Draw a clock	Draw a clock
Naming	Naming	Naming	Draw a clock	Repetition 3	Naming	Naming	Naming
Repetition 3	Repetition 3	Repetition 3	Repetition 3	Orientation in time	Recognition	Recognition	Repetition 3
Recognition	Recognition	Recognition	Repetition 2	Recognition	Orientation in time	Orientation in time	Recognition
Orientation in time	Orientation in time	Orientation in time	Orientation in geography	Naming	Repetition 3	Orientation in geography	Orientation in time
Orientation in geography	Orientation in geography	Orientation in geography	Recognition	Name and address learning	Orientation in geography	Repetition 3	Orientation in geography
Name and address learning	Name and address learning	Name and address learning	Name and address learning	Orientation in geography	Name and address learning	Name and address learning	Name and address learning
Repetition 2	Repetition 2	Repetition 2	Orientation in time	Repetition 2	Repetition 2	Repetition 2	Repetition 2

Note. bvFTD=behavioural variant frontotemporal dementia, AD=Alzheimer's disease, SD=semantic dementia, LPA=Logopenic progressive aphasia, PNFA=progressive nonfluent aphasia, CBD=corticobasal degeneration. Repetition 2=Repetition-all that glitters is not gold, Repetition 3=Repetition-a stitch in time saves nine

5.4 Item selection

Based on the results on Mokken scaling on the original clinical sample ($N=117$) above, 17 items were available for selection for the Mini-ACE from a formal hierarchy of *difficulty* (see Table 5.3). While the item selection process does involve a subjective element the contribution of Mokken scaling methods here firstly reduced the item pool from 24 to 17 items, all of which were sufficiently *discriminatory*, and added useful item parameters to guide item selection. To maintain content coverage it was desirable that from this hierarchical scale one item from each domain be selected for inclusion (see Table 5.2). Therefore the items were chosen to fulfil two requirements; that the new scale be comprised of highly *discriminatory* items with differing levels of *difficulty*, and that the new scale assess the five cognitive domains of the original ACE; attention, memory, fluency, language, and visuospatial skills. Within each domain the item with the most appropriate item properties identified through Mokken scaling analysis will be considered for inclusion in the new scale. Consequently the results of Mokken scaling analysis of the ACE-III formed the basis for the selection of five key items. That five items were to be selected was based on a conceptually driven decision to maintain content coverage of the main cognitive domains of the ACE-III whilst limiting the number of items and therefore time required for test administration.

The ideal scale will comprise items with high *discrimination* across a range of *difficulty* to ensure all levels of ability are assessed. On this basis Mokken scaling can identify key items that could be included in a shortened screening tool.

5.4.1 Range of difficulty

Focusing on the coverage of a wide range of *difficulty* ‘name and address recall’ with high *discrimination* ($H_i = 0.51$) and high *difficulty* (mean = 0.33) was singled out as a useful item as it may assist in the detection of small changes in milder cases of cognitive decline as this ability is lost quickly at an early stage. As a high *difficulty* item problems with ‘name and address recall’ could alert physicians and carers as it may herald the initial stages of cognitive decline.

With regards to the mid ranges of item *difficulty* ‘draw a clock’ can be identified on the basis of a *difficulty* value adequately assessing the mid-range of *difficulty* (mean=0.64) with good *discrimination* ($H_i=0.42$). This item was therefore identified as a candidate item for the midlevels of dementia severity.

In examining the *difficulty* values of the high *discrimination* items it is apparent that ‘identify fragmented letters’ and ‘name and address learning’ are among the least *difficult* (mean scores=0.91 and 0.78 respectively) and most *discriminatory* ($H_i=0.54$ and $H_i=0.52$ respectively). These items *discriminate* well at the assessment of the lower end of the hierarchy- i.e. these items may help to indicate differences in ability in the more advanced stages of disease. Patients with problems with these items are unlikely to be able to correctly respond to any of the more *difficult* items. In this way knowing the ordering of item *difficulty* can provide a quick gauge of a patient’s level of functioning. Due to the similar item properties of these items they contribute similar information. Only one of these would be required in a brief screening tool. As ‘name and address learning’ provides similar information to ‘orientation in time’ the inclusion of this item may add more reliability at this level with two items assessing a similar degree of cognitive decline. However, ‘identify fragmented letters’ offers measurement at a slightly more severe level of impairment with only 15% of participants scoring less than full marks in comparison to 50% of participants

scoring less than full marks for ‘name and address learning’ which may make ‘identify fragmented letters’ a better floor item.

Graphical analysis using TestGraf used in conjunction with the findings from Mokken analysis can be particularly useful in examining the response probabilities and *discrimination* of each of the item options. Figure 5.1 depicts the ICCs for ‘identify fragmented letters’ and ‘name and address learning’. These ICCs graphically represent the mean item score by means of the ‘expected score’ and confidence interval as a function of cognitive impairment. The curves also illustrate the *discriminatory* power of both items; ‘identify fragmented letters’ and ‘name and address learning’. The slope reflecting the rate of change indicates the degree of item effectiveness at any point along the latent trait (De Jong & Molenaar, 1987).

From Figure 5.1 it is evident that for both items the short vertical lines, reflecting 95% confidence regions, are larger for more severe cognitive decline due to the relatively low number of participants scoring poorly available to estimate the curve, reflecting the relative lack of *difficulty* of these items. This is particularly evident for the less *difficult* of the two ‘identify fragmented letters’.

Importantly, from this figure it can also be noted that while ‘identify fragmented letters’ demonstrates good *discrimination* at lower levels of ability this sharply diminishes after an approximate expected score of 10 the slope for ‘name and address learning’ remains steep until a far lesser degree of impairment. These slopes illustrate that ‘name and address learning’ *discriminates* well at a similar level of ability to ‘identify fragmented letters’ but also has a wider range of effective measurement.

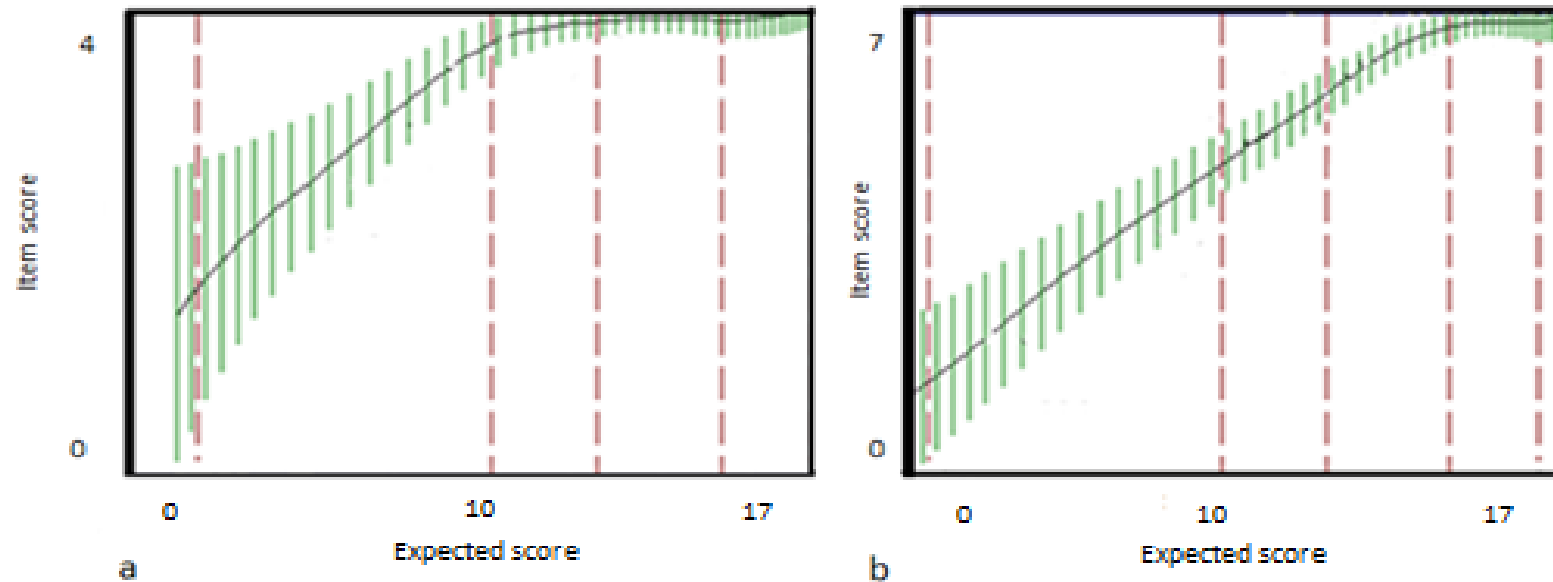
The probability of obtaining the maximum points in ‘identify fragmented letters’ increases at a lower ‘expected score’ range than ‘name and address learning’. At an approximate score of 10 there is almost 100% probability of obtaining the full score for this item. For ‘name and address learning’ the increase in probability of scoring maximum points

CHAPTER 5: DEVELOPMENT OF MINI-ACE

occurs at a higher 'expected score' range and does not reach the same high levels of probability even at an 'expected score' of 15.

'Identify fragmented letters' is less *difficult* and contributes more heavily to the lower limits of ability. However with most subjects performing well on this item even within very low expected score ranges this item does not contribute much to the measurement of any but the few who score at floor range. 'Name and address learning', whilst not quite as efficient at the extreme low end of ability, does provide more information beyond this level. 'Name and address learning' was therefore selected to indicate differences in ability in the more advanced stages of disease.

Figure 5.1 Item Characteristic Curves providing graphical representation of *discriminatory* power for: ‘identify fragmented letters’ (a) and ‘name and address learning’ (b).



Note. The x-axis indicates the expected ACE-III score on the 17 Mokken scale items. The y-axis reflects the item score. The vertical short lines along the curve reflect 95% regions of confidence for the position of the population curve at that trait level.

5.4.2 Cognitive domains

As ‘name and address recall’ demonstrated good *discrimination* at a high level of *difficulty* it was selected for the assessment of high *difficulty* within the memory domain. For the assessment of visuospatial skills ‘draw a clock’ identified as *discriminatory* at the mid-ranges of *difficulty* was selected.

With regards to the items within the attention domain both ‘orientation in time’ and ‘orientation in geography’ were identified as good candidate items. The mean scores of both items are very similar (0.74 and 0.75). The *discrimination* value of ‘orientation in geography’ was slightly higher ($H_i = 0.50$) than ‘orientation in time’ ($H_i = 0.44$) making this the preferred choice of attention items assessing the mid-severe ranges of *difficulty* based on the Mokken scaling results. However the pre-analysis preference for ‘orientation in time’ prevailed and this item was consequently chosen to assess attention in the middle-severe range of item *difficulty*.

To maintain content coverage one of the verbal fluency items must be included. ‘Verbal fluency-animal’ has a higher *discriminatory* value ($H_i = 0.51$) than ‘verbal fluency-letters’ ($H_i = 0.37$) and was therefore considered the more suitable choice of the two based on its greater sensitivity to the assessment of the latent trait. The larger discrepancy between *discrimination* values (0.51-0.37) for both fluency items as opposed to the smaller difference between that of the orientation items (0.50-0.44) was sufficient to alter the pre-analysis selection of ‘verbal fluency-letters’. Therefore ‘verbal fluency-animals’ was selected as the fluency item in the short scale assessing the mild-moderate levels of disease severity.

5.4.3 Practical consideration

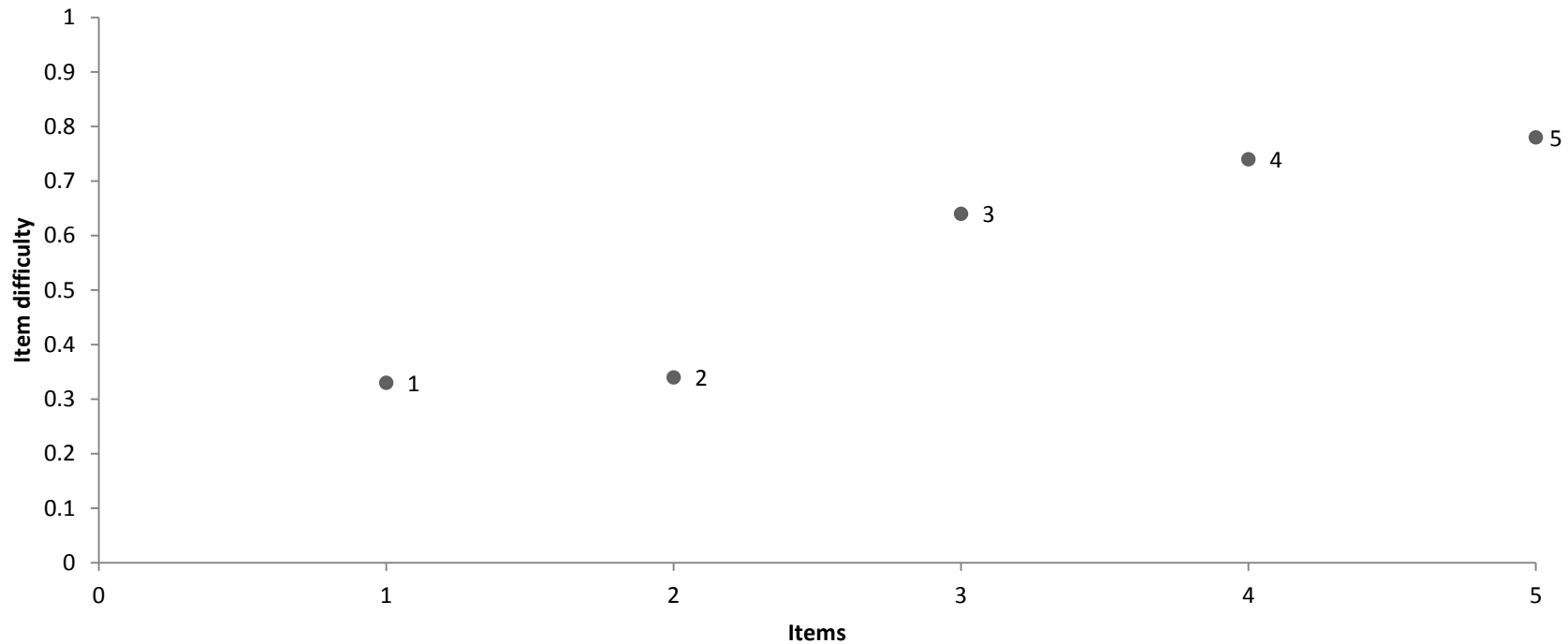
Throughout this analysis it had been overlooked that the inclusion of ‘name and address recall’ required the additional inclusion of ‘name and address learning’. Therefore despite

CHAPTER 5: DEVELOPMENT OF MINI-ACE

having used these analytical techniques to determine ‘name and address learning’ was the more appropriate item for selection the practicalities of testing also necessitated and supported the inclusion of ‘name and address learning’ as the item assessing the advanced stages of dementia.

Therefore the five items identified as candidate items for the Mini-ACE were; ‘name and address learning’: subject given a name and address to repeat three times and told they will be asked about it again later, the third trial of repeating “Harry Barnes, 73 Orchard Close, Kingsbridge, Devon” is scored (maximum score =7); ‘orientation in time’: “what day, date, month, year and season is it?” (maximum score=5); ‘verbal fluency-animal’: subject given one minute to generate name as many animals as possible (maximum score=7); ‘draw a clock’: subject asked to draw a clock face with numbers with the hands at ten past five (maximum score=5), and ‘name and address recall’: subject asked to recall the name and address learned previously (Harry Barnes, 73 Orchard Close, Kingsbridge, Devon) (maximum score=7). This selection of items was then assessed with regards to the intended use and goal of the new scale. Figure 5.2 presents the breadth of *difficulty* levels the five candidate items assess to ensure that the Mini-ACE measures a wide spectrum of impairment (see Appendix A for the scale in full).

Figure 5.2 Range of *difficulty* coverage of item selection for the Mini-ACE



Note. The y-axis represents the mean item scores reflecting item *difficulty* (with higher mean values indicating lower *difficulty*). The x-axis represents the number of items in the scale.

1=Name and address recall (0.33); 2=Fluency-Animal (0.34); 3=Draw a clock (0.64); 4=Orientation in time (0.74); 5=Name and address learning (0.78).Mini-ACE=Mini-Addenbrooke's Cognitive Examination.

5.5 Validation of the Mini-ACE

An additional Mokken scaling analysis was carried out on the five Mini-ACE candidate items. This analysis was performed using the sample ($N=350$) from Chapter 4. This sample was described fully in Chapter 4. As the Mini-ACE items are common to the ACE-R the scores for the Mini-ACE were extracted from the ACE-R data.

Assessing the unidimensionality of the data showed that the scalability coefficients (H_i) of all five items were above the 0.3 threshold indicating that in this population sample, these items demonstrate good *discrimination* and are homogenous (correlated) enough for inclusion in the scale. All item pair scalability coefficients (H_{ij}) were positive.

As the Mokken scaling procedure did not identify any violations none of the scale items were excluded in the assessment of monotonicity or IIO. Therefore the five item scale formed a reliable ($MS=0.75$) strong hierarchical Mokken scale ($H=0.53$) with IIO ($H^T=0.52$). This means the items can be ordered according to their *difficulty* level and items have the same *difficulty* ordering irrespective of the value of the respondent's cognitive ability. Table 5.7 lists the Mini-ACE items ordered by *difficulty* and *discrimination* for this sample.

The ordering by *discrimination* is the same for both samples. The ordering by *difficulty* is the same as the ordering for the Mini-ACE development sample ($N=117$) for more *difficult* items; 'name and address recall', 'verbal fluency-animal' and 'draw a clock' but the ordering is reversed for the two least *difficult* items; 'name and address learning' and 'orientation in time' (see Table 5.8 for a comparison of Mini-ACE item *difficulty* and *discrimination* between the two samples).

The differences in *difficulty* ordering between the two samples are likely to be a result of the different patient groups included in each. The largest patient group in the larger Chapter 4

sample was bvFTD ($n=96$). Patients with FTD generally exhibit preserved orientation in time and place (Neary & Snowden, 1996) which could be contributing to the finding of ‘orientation in time’ as the least *difficult* item in this sample.

Table 5.7 Mini-ACE items ordered by *difficulty* and *discrimination* for Mini-ACE validation sample

Item	Mean	Item	H_i
Name and address recall	0.30	Name and address learning	0.56
Verbal fluency-animal	0.31	Name and address recall	0.56
Draw a clock	0.72	Verbal fluency-animal	0.52
Name and address learning	0.72	Orientation in time	0.49
Orientation in time	0.76	Draw a clock	0.48

Note. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*.

Table 5.8 Mini-ACE items ordered by *difficulty* and *discrimination* (from most to least) for the Mini-ACE development and Mini-ACE validation samples

Mini-ACE development sample ($N=117$)				Mini-ACE validation sample ($N=350$)			
Item	Mean	Item	H_i	Item	Mean	Item	H_i
Name and address recall	0.33	Name and address learning	0.52	Name and address recall	0.30	Name and address learning	0.56
Verbal fluency-animal	0.34	Name and address recall	0.51	Verbal fluency-animal	0.31	Name and address recall	0.56
Draw a clock	0.64	Verbal fluency-animal	0.51	Draw a clock	0.72	Verbal fluency-animal	0.52
Orientation in time	0.74	Orientation in time	0.44	Name and address learning	0.72	Orientation in time	0.49
Name and address learning	0.78	Draw a clock	0.42	Orientation in time	0.76	Draw a clock	0.48

Note. Mini-ACE=Mini Addenbrooke's Cognitive Examination, Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*.

5.6 Discussion

This study demonstrates the application of Mokken analysis as a data driven scaling method in the development of a new instrument. The Mini-ACE developed using Mokken scaling analysis has been validated in patients with varying dementia diagnoses (Hsieh et al., 2015). The further analysis of the Mini-ACE in the independent validation sample in this chapter also demonstrates that the new scale performs well in this additional larger sample. All items were sufficiently *discriminatory* indicating good sensitivity to the assessment of cognitive decline in dementia. The five items were retained in an invariantly ordered hierarchical Mokken scale.

The Mini-ACE offers several advantages in clinical applications. Firstly the Mini-ACE was largely empirically derived from a Mokken scaling analysis. This item selection formed a highly sensitive instrument which is less likely to have ceiling effects making it particularly valuable for assessing patients with mild cognitive impairment (Hsieh et al., 2015). It is brief, taking under five minutes to administer and can be administered and scored without formal specialised training. As the ACE-III takes at least 15-20 minutes to complete the reduction in time taken to administer the test is significant as in clinics and hospital settings where time for assessing patients is limited even seconds can count. Another advantage of the Mini-ACE is the fact that it was derived from the ACE-III items, items which are also present in the ACE-R, which enables clinicians to extract Mini-ACE scores from pre-existing data from the ACE-III and ACE-R. Additionally, with coverage of the domains of the full length ACE-III the Mini-ACE is also capable of providing somewhat distinctive diagnostic profiles across AD, FTD and corticobasal syndromes (Hsieh et al., 2015).

The assessment of orientation in the Mini-ACE differs slightly from the more comprehensive ACE, ACE-R and ACE-III measures; the embedded item ‘orientation to season’ was removed as there are issues regarding the universality of this item due to geographical differences (e.g. the tropics where there are only two seasons: the wet season and the dry season). The removal of this embedded orientation item reduced the maximum orientation score from five to four and gives the Mini-ACE a total score of 30, with higher scores indicating better cognitive ability.

While item *discrimination* was considered in item selection it should be noted that as Mokken scaling is designed to create scales of items that are sufficiently *discriminatory* between respondents all items within this hierarchical Mokken scale can be considered as having good *discriminatory* value. However as only five items were required those items with higher *discriminatory* value within the hierarchical scale were focused on for item selection.

5.6.1 Limitations and future directions

Some study limitations require comment. The Mini-ACE was developed using data from patients at a specialised clinic. This sample overrepresented some of the less commonly occurring forms of dementia such as progressive primary aphasia which comprised 42% of the sample. Prospective testing of new measures of dementia assessment in representative samples of patients with dementia is essential to ensure the general applicability of these scales (Borson, Scanlan, Chen & Ganguli, 2003). The secondary validation analysis in this chapter also used data collected at the same specialist multidisciplinary tertiary referral centre (Frontier Research Group) as the Mini-ACE development sample which raises the question of the generalizability and widespread application of the scale. Hsieh et al. (2015) also validated the new scale in an independent sample of 242 participants from the Frontier Research Group including 164 with a dementia diagnosis (PPA, $n=82$; AD, $n=38$; bv-FTD, $n=23$; CBS, $n=21$)

CHAPTER 5: DEVELOPMENT OF MINI-ACE

and 78 controls. However, like the Mini-ACE development sample both of these validation samples have a preponderance of bv-FTD and PPA patients due to the research interests of the clinics where the data were collected. Another concern here is the size of the sample used to develop the Mini-ACE. Therefore further research to examine the wider applicability and performance of the Mini-ACE in a larger sample from a more general and less acute community environment is necessary.

The Mini-ACE was derived from analysis of the ACE-III whereas the subsequent validation analysis was performed using data from the ACE-R. The original validation analysis of the Mini-ACE was also carried out using ACE-R data (Hsieh et al., 2015). While this appears incongruous, the Mini-ACE items are common to both the ACE-R and the ACE-III both in question wording and scoring which means it is unlikely that this inconsistency had an effect on the results of the validation analyses.

Further work should also be undertaken to determine the effects of age, education and sex on Mini-ACE performance and to evaluate this new scale against alternative brief tools such as the Montreal Cognitive Assessment (Nasreddine et al., 2005). Also the brevity of the scale means the time delay between learning the name and address and subsequent recall is shorter (i.e. less than five minutes time delay compared to approximately 15 minutes delay in the full version). However reduction in this timespan is unlikely to have much of an effect on the memory score in the Mini-ACE of most patients, particularly those with AD, as problems with retention of new material have been established following a very brief interference tasks in AD (Benson, Slavin, Tran, Petrella & Doraiswamy, 2005; Fillenbaum, Wilkinson, Welsh & Mohs, 1994; Chandler et al., 2004). Alternate versions of the Mini-ACE should be developed for clinical use particularly for 'name and address learning' where the repeated use of the same memory stimuli could help patients to improve their recollection on subsequent clinic visits.

Mokken scaling methods can be used to reduce the subjectivity of item selection by providing important item parameters. *Discrimination* for example helps to ensure that the scale consists of sufficiently *discriminatory* items, which is a crucial consideration in the development of a new scale. For example, the greater *discrimination* value of ‘verbal fluency-animal’ prompted the inclusion of this item instead of ‘verbal fluency-letters’, which had been identified for selection prior to the analysis. However, Figure 5.2 demonstrates the similarity of the ‘name and address learning’ and ‘verbal fluency-animal’ in terms of *difficulty*. Given the greater coverage offered and the selection of ‘verbal fluency-letter’ prior to the consideration of the results of Mokken scaling analysis it might be worth assessing performance of the scale with ‘verbal fluency-letter’ included in place of ‘verbal fluency-animal’. While the results demonstrate the weaker *discriminatory* value of this item the *difficulty* level would provide greater range of coverage.

Therefore despite analysis and consideration of item properties the ultimate choice of items to form new scales from this 17 item hierarchical scale had a considerable subjective element. Between items of a similar level of *discrimination* and *difficulty* the item selection can be determined by the aim of the scale. For example, ‘identify fragmented letters’ or ‘name and address learning’ could have been selected to assess the more severe levels of impairment. The choice of ‘name and address learning’ was made based consideration of item properties and the graphical assessment of item functioning with regards to the intended goal of the Mini-ACE. While ‘name and address learning’ was deemed the more appropriate choice based on the results of Mokken scaling on reflection of the scale as a whole, the inclusion of ‘name and address recall’ requires the inclusion of ‘name and address learning’ in order to test recall. This provides an example of the importance of taking practical considerations into account and maintaining sight of the overall aim of the scale whilst selecting items with the most appropriate item properties for inclusion.

The combination of a data driven approach with the influence of clinical experience should be noted. The practical experience will mean that subjective decisions can be made that fit with prior clinical practice but these experiences may unconsciously bias the selection. For example, the knowledge that certain items such as ‘name and address recall’, contribute heavily to the detection of early identification of memory decline could influence the preference for selection of this item. While the results of Mokken scaling support the inclusion of this item the possibility of bias in item selection should be considered in scale development.

It would be interesting to determine the performance of the scale comprising the choice of item assessing attention based on the results of the Mokken scaling analysis; ‘orientation in geography’. While both items assess the same level of *difficulty* ‘orientation in geography’ displayed greater *discrimination* than ‘orientation in time’ implying that at this level of impairment ‘orientation in geography’ demonstrates a closer association to the assessment of cognitive impairment. Additionally the selection of ‘orientation in geography’ would not require the removal of any of the embedded items as in the case of ‘orientation in time’ where orientation to season was removed.

These issues pertain to the element of subjectivity regarding the choice of items from within an IIO hierarchy. All items within the hierarchy demonstrate good item properties and on this basis alone are already acceptable choices for item selection. Examining the items that failed to meet the criteria for inclusion in the formal Mokken hierarchy provides further example for the benefits of using Mokken scaling methods to identify items for inclusion in a new scale. For example, in this sample the item ‘count dot arrays’ was excluded from each of the eight separate analyses within this chapter. Six of these exclusions were based on poor *discrimination* and scalability. This indicates that this item would be a very poor choice of

item in a new scale. This emphasises how Mokken scaling methods used here eliminate such items from consideration.

5.6.2 Conclusion

This chapter demonstrates the use of IRT methods in scale analysis and development. Examining item responses and characteristics can identify key items that contribute meaningfully to the scale and at known levels of ability. This knowledge can be valuable in the development of a shorter scale. While further validation analyses in larger more representative samples are required the Mini-ACE developed through an empirical data-driven Mokken scaling approach is available for use as a brief and sensitive cognitive instrument allowing for rapid assessment of patients.

Chapter 6: Hierarchical patterns of decline in the Addenbrooke's Cognitive Examination-Revised

6.1 Introduction

This chapter applies Mokken scaling analyses to the ACE-R to investigate item properties within and the hierarchical structure of the scale. This is similar to Chapter 4 where Mokken scaling analyses were applied to the ACE-R collected in the Sydney cohort: a sample derived from a tertiary memory clinic specialising in the study frontotemporal dementia and related disorders. In this chapter, however, the sample is derived from the Scottish Dementia Research Interest Register (SDRIR) the majority of whom have AD. A novel element of the present chapter is the concurrent analysis of the sample as a whole along with three diagnostic subsamples (late onset AD; early and late onset AD; mixed AD and vascular dementia). This permits the examination of potential differential item functioning by diagnosis and the scope to determine whether the item ordering by *difficulty* within the ACE-R is impervious to the different cognitive impairments associated with different types of dementia.

The sample in this Chapter predominantly comprises patients diagnosed with Alzheimer's disease. Therefore while the analyses of this study include three diagnostic groups the focus is largely on the course of decline in Alzheimer's disease. This allows for a more precise level of Alzheimer's disease-specific analysis to be carried out. Alzheimer's disease is the most common cause of dementia accounting for an estimated 60% to 80% of all cases of dementia (Thies & Bleiler, 2013).

Enhanced knowledge of the pattern of decline in Alzheimer's disease would be an important advance permitting further understanding of the trajectory of impairment and could facilitate the development and application of treatments.

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

Alzheimer's disease is typically differentiated from other non-Alzheimer's disease forms of dementia by neuropsychological examination and detailed clinical history. It would be valuable if distinct patterns of impairment within one of the most commonly used measures of cognitive functioning in dementia could be discerned. These unique patterns could assist in discriminating Alzheimer's disease from other variants of dementia. For example, assessments of patterns of performance on various tests of memory have proven valuable in discriminating between Alzheimer's disease and frontotemporal dementia (Wicklund, Johnson, Rademaker, Weitner & Weintraub, 2006).

The typical clinical course of Alzheimer's disease first manifests in problems with memory with the inability to retain recently acquired information and culminates in difficulty with over-learned associations and early-learned verbal mimicking (McKhann et al., 2011; Ashford, Kolm, Colliver, Bekian & Hsu, 1989). As the disease progresses other domains of cognition (language, executive functioning, visuospatial abilities) are affected to varying degrees. With regards to the ACE-R this sequence of impairment suggests that the items that will first detect the onset of cognitive decline will be the recall items, and there will be later problems with repetition and registration. To determine whether this pattern is common to all patients with Alzheimer's disease of similar level of impairment or whether there is some inter-diagnostic variation Mokken scaling analysis can be applied to investigate invariant item ordering. If all items conform to an IIO hierarchy this suggests that there is a generally consistent order to the decline in Alzheimer's disease whereas many violations of IIO imply that some items are more *difficult* for some patients with AD than for others with the same diagnosis.

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

Mokken scaling analyses can add value to cognitive assessments in many ways: (i) where assessments are used to screen for cognitive impairment Mokken scaling can help by identifying which items are likely to reveal initial changes in cognition in various types of dementia; (ii) where measures are applied to assist in differential diagnosis of cause invariantly ordered items can help to establish unique sequences of decline for different types of dementia, and (iii) rating severity or monitoring disease progression can be improved by using Mokken scaling to develop hierarchical scales whereby the score from an isolated item can be used to quickly gauge a patient's level of functioning and to anticipate subsequent decline. In these ways Mokken scaling analyses can provide clinicians with important information on the symptomology of different dementia syndromes by the location of specific items on the continuum of disease severity for different clinical groups.

Mokken scaling analyses in Chapter 4 revealed invariantly ordered subsets of ACE-R items. However the low numbers of participants forming diagnostic subgroups restricted these analyses. The sample size limits the reliability of IIO. Furthermore several items were identified in Chapter 4 which could violate the assumption of local stochastic independence. Therefore the aim of this chapter was to analyse the pattern of decline in both a large mixed patient sample, with and without the stochastically dependent items, in addition to larger diagnostic sub-samples to examine how this method of item level analysis can add value to the application of cognitive assessments in dementia.

6.2 Method

6.2.1 Participants

This sample is drawn from the Scottish Dementia Research Interest Register (ethics approval from Scotland A REC 08/MRE00/49). This national case register was set up under the aegis of the Scottish Dementia Clinical Research Network (SDCRN) in 2008 to facilitate dementia research by providing a central database of people with dementia who are interested in participating in research studies investigating the causes and consequences of dementia. Participants were referred by their clinicians having been diagnosed with dementia or a related cognitive disorder making the SDRIR a clinical as opposed to an epidemiological sample. As a voluntary database it does not include all (or indeed necessarily a representative sample of) people with dementia.

Participants were 1248 individuals with dementia who had been recruited to the register by March 2014. All had consented, or where consent cannot be given due to lack of capacity consent was sought from the participant's legal representative, to the storage of demographic information along with data from cognitive, functional and behavioural assessments, including the ACE-R. Clinical studies officers, all of whom had undertaken training and validation to guarantee consistency across assessments, assessed SDRIR participants.

Data were entered by the clinical studies officers directly onto a secure laptop and then uploaded to a secure central server at the Health Informatics Centre in Dundee, Scotland. Diagnoses were established by the independent classification by an old-age psychiatrist and physician. Differences in opinion were resolved by referring to original records and case notes. The dementia diagnoses represented on the register

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

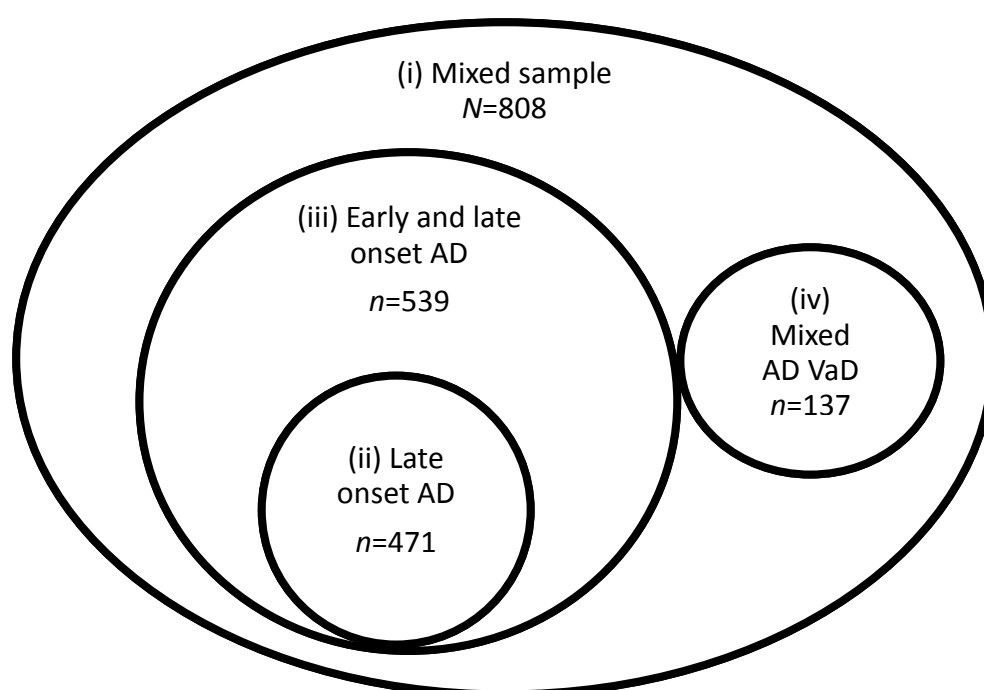
were classified as: (i) late-onset AD (ii) young onset AD, (iii) VaD, (iv) mixed AD and cerebrovascular disease, (v) FTD, (vi) dementia with Lewy bodies (DLB), (vii) Parkinson's disease dementia (PDD), (viii) other dementia, (ix) mild cognitive impairment, and (x) uncertain diagnosis.

From this register those with complete itemised ACE-R data along with along with age, sex, and diagnosis variables ($n = 921$) were isolated for Mokken analysis. Patients diagnostically classified as 'other dementia' ($n = 35$), 'uncertain diagnosis' ($n = 53$) and 'mild cognitive impairment' ($n = 14$) were excluded from analysis. A further 11 participants were excluded from the final dataset due to inaccuracies in age or scoring. The final dataset ($N = 808$) included several dementia aetiologies; late onset AD ($n = 471$), mixed AD VaD ($n = 137$), VaD ($n = 89$), early onset AD ($n = 68$), DLB ($n = 20$), FTD ($n = 14$) and PDD ($n = 9$).

This whole sample (i) SDRIR ACE-R sample ($N = 808$), was further split to provide three sub-samples (ii) late onset AD ($n = 471$), (iii) combined early and late onset AD ($n = 539$), (iv) mixed AD VaD ($n = 137$). These groups were formed to provide sufficiently large and diagnostically heterogeneous groups for the analysis of item ordering in different types of dementia. Figure 6.1 provides a visual overview of the samples analysed here and the relationships between them. It should be noted that the combined early and late onset AD sample (sample iii) comprises the same participants as the late onset AD sample (sample ii) with the addition of 68 early onset AD patients. As there were insufficient number for separate stratified analysis of the item ordering of early onset AD these 68 early onset patients were added to the late onset AD group to determine whether the inclusion of these patients changed the item ordering of the late onset AD group. Any discrepancies between the item ordering of the groups could be considered the influence of a differing symptomology of early

onset AD. The full sample plus the three diagnostic subgroups were analysed to investigate the item properties and ordering within each group. Demographic and cognitive information for each group is presented in Table 6.1.

Figure 6.1 Venn diagram illustrating the relationship between the different samples used in these analyses



Note. AD=Alzheimer's disease, VaD= Vascular dementia.

Table 6.1 Demographic and cognitive information for diagnostic SDRIR

groups

	N	Sex (% male)	Mean age (SD)	Mean ACE-R (SD)
Data as a whole	808	425 (52.6%)	77.5 (7.8)	63.0 (16.8)
Late onset AD	471	227 (48.2%)	79.5 (5.7)	62.9 (16.8)
Early onset AD	68	37 (54.4%)	63.3 (5.7)	60.9 (21.4)
Combined early and late onset AD	539	264 (48.9%)	77.4 (7.9)	62.7 (17.5)
Mixed AD VaD	137	75 (54.7%)	79.0 (6.8)	64.6 (14.4)

Note. SDRIR=Scottish Dementia Research Interest Register. N=number of participants in each sample, SD=standard deviation, ACE-R=Addenbrooke's Cognitive Examination-Revised, AD=Alzheimer's disease, VaD=vascular dementia

6.2.2 Statistical analysis

ACE-R item scores were equated for analysis by dividing each mean item score by the range of each item to give a score between 0 and 1 (e.g. for late onset AD

'Orientation in time' mean score of 3.4 was divided by the maximum score possible for this item, 5). Item scores for each of the four samples is presented in Table 6.2.

ACE-R items in each sample were analysed using Mokken scaling analysis using the Mokken package in R (van der Ark, 2007). The fit of the items to each of the two Mokken models was assessed by examining the fit of the items to the four assumptions of Mokken scaling; unidimensionality, local independence, monotonicity and non-intersection.

Both confirmatory and exploratory Mokken scaling analysis was applied in these analyses. Mokken scaling can be performed in either an exploratory or confirmatory manner. In both applications the same criteria to determine presence and strength of Mokken scales are used but the methods differ in terms of what is entered

into the analysis (Watson, Wang & Thompson, 2014). While both methods are equally useful and the choice of approach is flexible, both exploratory and confirmatory Mokken scaling modes were applied here to determine whether there were subscales of the ACE-R according to Mokken scaling conditions and whether these subscales would be identified by scalability coefficients in the confirmatory analyses.

Exploratory Mokken scaling analysis applies the automated item selection procedure (AISP). AISP uses an iterative hierarchical process to select items and partition them to scales starting with the pair of items with the highest item-pair scalability coefficient (H_{ij}). Starting with the items with the highest scalability the process continues and builds as many scales as necessary until no additional items can be allocated to a scale. Any unscalable items, those with item scalability coefficients (H_i) less than the chosen lower bound threshold (0.30 by default) for example, remain out with any scale.

While the AISP uses scalability coefficients to guide its item allocations, i.e. the algorithm starts with the items with the highest item pair scalability coefficients and continues selecting items on this basis until no items remain than meet Mokken scaling criteria; the confirmatory approach provides coefficients for examination with heuristic guidelines for item exclusions. Scalability coefficients of items, item pairs and the scale along with their associated standard errors at item-pair, item and scale level are examined. Large standard errors (for example, 0.8) indicate a reasonable likelihood that the scale H is less than 0.3 making the scale items unscalable (Kuijpers, van der Ark & Croon, 2013). The implications of standard error values extend also to item-pair and item scalability coefficients.

As the only difference between exploratory and confirmatory methods is what is entered into the analysis once this initial assessment of scalability coefficients was performed, IIO can be examined in the same manner. The interpretation of IIO was based on calling the function *check.iio* and visual inspection of item-pair plots. Item rest-score regression plots were visually inspected to identify item overlap or ‘outlying’ items: items located far away from the cluster of the other scale items. These items can cause artificially exaggerated IIO and can result in the misleading appearance of IIO (Meijer & Egberink, 2012).

An additional analysis was carried out excluding items with potential violations of local stochastic independence (‘3-item registration’, ‘3-item recall’, ‘name and address learning’, ‘name and address recall’ and ‘name and address recognition’). This additional analysis was carried out using the largest available sample ($N = 808$).

6.3 Results

The sample comprised of 808 SDRIR participants (425 male, 383 female) with a mean age of 77.5 (SD=7.8) years, diagnosed with dementia were included in the analysis (see Table 6.1). The sample comprised different dementia diagnoses; AD ($n = 471$), mixed AD VaD ($n = 137$), VaD ($n = 89$), early onset AD ($n = 68$), DLB ($n = 20$), FTD ($n = 14$) and PDD ($n = 9$). From this sample three subsamples were also isolated for analysis; late onset AD ($n = 471$), late and early onset AD ($n = 539$) and mixed AD VaD ($n = 137$). Equated ACE-R item scores were used to designate item *difficulty* in Mokken scaling. These scores are presented for each of the four clinical groups in Table 6.2.

Table 6.2 Mean equated ACE-R item scores for each SDRIR sample. Items presented in order of test administration

Data as a whole (N=808)		Late onset AD (n=471)		Combined early and late onset AD (n=539)		Mixed AD VaD (n=137)	
Item	Mean	Item	Mean	Item	Mean	Item	Mean
Orientation in time	0.57	Orientation in time	0.68	Orientation in time	0.55	Orientation in time	0.57
Orientation in geography	0.90	Orientation in geography	0.91	Orientation in geography	0.90	Orientation in geography	0.90
3 item registration	0.95	3 item registration	0.95	3 item registration	0.96	3 item registration	0.96
Serial sevens	0.77	Serial sevens	0.80	Serial sevens	0.76	Serial sevens	0.80
3 item recall	0.24	3 item recall	0.22	3 item recall	0.22	3 item recall	0.27
Name and address learning	0.66	Name and address learning	0.66	Name and address learning	0.65	Name and address learning	0.73
Memory retrograde	0.44	Memory retrograde	0.39	Memory retrograde	0.42	Memory retrograde	0.45
Fluency-letters	0.54	Fluency-letters	0.56	Fluency-letters	0.56	Fluency-letters	0.52
Fluency-animals	0.36	Fluency-animals	0.37	Fluency-animals	0.37	Fluency-animals	0.38
Follow written command	0.95	Follow written command	0.95	Follow written command	0.95	Follow written command	0.97
Syntactical comprehension	0.93	Syntactical comprehension	0.93	Syntactical comprehension	0.93	Syntactical comprehension	0.92
Write a sentence	0.86	Write a sentence	0.89	Write a sentence	0.87	Write a sentence	0.88
Repetition 1	0.85	Repetition 1	0.85	Repetition 1	0.85	Repetition 1	0.85
Repetition 2	0.91	Repetition 2	0.91	Repetition 2	0.90	Repetition 2	0.94
Repetition 3	0.67	Repetition 3	0.69	Repetition 3	0.68	Repetition 3	0.65
Naming 1	0.95	Naming 1	0.95	Naming 1	0.95	Naming 1	0.97
Naming 2	0.78	Naming 2	0.77	Naming 2	0.77	Naming 2	0.81
Semantic comprehension	0.72	Semantic comprehension	0.70	Semantic comprehension	0.72	Semantic comprehension	0.70
Reading	0.87	Reading	0.88	Reading	0.87	Reading	0.91
Draw intersecting pentagons	0.56	Draw intersecting pentagons	0.61	Draw intersecting pentagons	0.57	Draw intersecting pentagons	0.58
Draw a cube	0.44	Draw a cube	0.46	Draw a cube	0.45	Draw a cube	0.48
Draw a clock	0.63	Draw a clock	0.65	Draw a clock	0.64	Draw a clock	0.66
Count dot arrays	0.86	Count dot arrays	0.88	Count dot arrays	0.87	Count dot arrays	0.89
Identify fragmented letters	0.92	Identify fragmented letters	0.93	Identify fragmented letters	0.91	Identify fragmented letters	0.97
Name and address recall	0.07	Name and address recall	0.06	Name and address recall	0.06	Name and address recall	0.08
Recognition	0.51	Recognition	0.49	Recognition	0.50	Recognition	0.53

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised, SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer's disease, VaD=Vascular dementia, mean=mean item scores reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. Repetition: repeat multi-syllabic words, Repetition 2: repeat 'Above, beyond and below', Repetition 3: repeat 'No ifs, ands or buts'. 'Naming 1': name pencil and watch, 'Naming 2': name 10 pictures, Follow written command: Follow written command-close eyes.

6.3.1 Mokken scaling analyses of diagnostic groups

Data from each of the 26 items of the ACE-R from SDRIR participants in the complete dataset ($N = 808$) plus each of the three diagnostic groups (late onset AD, early and late onset AD and mixed AD VaD) were analysed separately using Mokken scaling methods. Additional separate analysis was carried out on 21 of the ACE-R items using the full ($N=808$) dataset to investigate potential local stochastic independence violations.

(i) SDRIR mixed diagnosis sample ($N=808$)

In the exploratory assessment of unidimensionality in the Monotone Homogeneity Model (MHM) two ACE-R items were identified for exclusion; ‘repetition 1’ was unscalable, not conforming to any scale using the automated item selection procedure (AISP). ‘Repetition 3’ was also identified due to its low scalability coefficient. This item’s low *discrimination* value indicates its poor contribution to the assessment of cognitive impairment in dementia. Due to the poor contribution of these items to the measurement of dementia in this sample both items were excluded from further analysis. No items violated monotone homogeneity. This means that the probability of a correct response to each of the remaining items is a monotonically nondecreasing function of the latent trait. These 24 ACE-R items meet MHM criteria and form a moderate Mokken scale ($H=0.41$, $SE=0.02$).

Before analysing this 24 item subset for IIO, the items were analysed using a confirmatory Mokken approach. This assessment of scalability coefficients prompted the exclusion of ‘repetition 1’ and ‘repetition 3’ due to their low item scalability coefficients ($H_i=0.25$, $SE=0.03$ and $H_i=0.29$, $SE=0.03$ respectively). No further items

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

warranted exclusion on the basis of scalability coefficients or assessment of monotonicity. The 24 items remaining had a scale coefficient of 0.41 (SE=0.02).

The 24 items were then assessed for invariant item ordering. The backward selection procedure suggested the removal of 12 items ('orientation in time', '3 item registration', 'serial sevens', '3 item recall', 'name and address learning', 'memory retrograde', 'fluency-letters', 'fluency-letters', 'syntactical command', 'draw a cube', 'draw a clock' and 'name and address recall'). Following these exclusions an examination of item scalability coefficients of the remaining 12 items resulted in the exclusion of 'follow written command-close eyes' due to its low H_i value (0.27).

The remaining 11 items (Table 6.3) formed a reliable (MS=0.81), moderate hierarchical scale ($H=0.42$, SE=0.02) with IIO ($H^T=0.87$). The invariantly ordered items for the SDRIR ACE-R mixed sample are presented in Table 6.3 in order of *difficulty*, represented by equated mean item scores, and *discrimination* indexed by item scalability coefficients (H_i). This information is valuable as within the ACE-R these 11 items follow the same pattern of decline for all patients in this mixed sample. These IIO items are associated with similar levels of severity for the various dementia diagnoses represented in this sample.

Table 6.3 SDRIR mixed diagnosis sample: IIO hierarchy items listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
Recognition	0.51	Identify fragmented letters	0.47	0.03
Draw intersecting pentagons	0.56	Naming 1	0.46	0.04
Semantic comprehension	0.72	Semantic comprehension	0.46	0.03
Naming 2	0.78	Naming 2	0.45	0.03
Count dot arrays	0.86	Orientation in geography	0.43	0.03
Write a sentence	0.86	Reading	0.38	0.03
Reading	0.87	Write a sentence	0.38	0.03
Orientation in geography	0.90	Count dot arrays	0.37	0.03
Repetition 2	0.91	Recognition	0.36	0.03
Identify fragmented letters	0.92	Repetition 2	0.35	0.04
Naming 1	0.95	Draw intersecting pentagons	0.34	0.03

Note. SDRIR=Scottish Dementia Research Interest Register, IIO=Invariant item ordering, Mean= mean item score reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. H_i =item scalability coefficient indicating item *discrimination* with higher values associated with higher *discrimination*. SE=standard error. Repetition 2: repeat 'Above, beyond and below', 'Naming 1': name pencil and watch, 'Naming 2': name 10 pictures

- (ii) SDRIR mixed diagnosis sample excluding possible stochastically dependent items ($N=808$)

As some concerns about the stochastic dependence of five ACE-R items were raised in Chapter 4 an additional analysis was carried out to determine whether items identified as potential sources of local stochastic violations; '3-item registration', '3-item recall', 'name and address learning', 'name and address recall' and 'name and address recognition' were exaggerating the scalability of the ACE-R. To assess the effect of these items on the strength of the ACE-R these five items were excluded from this analysis. The remaining 21 items were analysed in the same manner and using the same data as described above.

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

Of these 21 items, ‘repetition 1’ and ‘repetition 3’ were allocated to a separate scale by AISP. ‘Orientation in time’ violated monotonicity and was also identified for removal at this point.

From a confirmatory approach the same two items were again identified for removal this time due to low scalability coefficients; ‘repetition 1’ ($H_i=0.25$, $SE=0.03$), ‘repetition 3’ ($H_i=0.28$, $SE=0.03$). In the assessment of monotonicity ‘orientation in time’ was again identified for exclusion due to violations.

Based on these results the decision to exclude these three items was made. The remaining 18 items could be considered unidimensional according to Mokken scaling criteria, all were partitioned to the same scale using AISP, all item-pair scalability coefficients were nonnegative, item scalability coefficients were greater than 0.3 and the scale coefficient was 0.44. These exclusions left 18 items meeting the assumptions of the MHM with $H=0.44$.

Seven items violated IIO. Due to these violations these items were removed leaving an 11 item Mokken scale ($H=0.42$, $SE=0.02$) with IIO ($H^T=0.86$) (see Table 6.4). This subset of items within the ACE-R demonstrates the same ordering in terms of *difficulty* for all patients in this sample. The scalability coefficients are similar to those in the previous analysis of all 26 items in the same sample ($H=0.42$, $H^T=0.87$) indicating that the item chains identified in Chapter 4 and excluded from this analysis are not driving the scalability of the ACE-R.

Table 6.4 SDRIR mixed diagnostic sample excluding possible stochastically dependent items: IIO hierarchy items listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
Memory retrograde	0.44	Identify fragmented letters	0.48	0.03
Draw intersecting pentagons	0.56	Semantic comprehension	0.46	0.02
Semantic comprehension	0.72	Orientation in geography	0.45	0.03
Serial sevens	0.77	Naming 2	0.45	0.02
Naming 2	0.78	Memory retrograde	0.43	0.02
Count dot arrays	0.86	Write a sentence	0.41	0.03
Write a sentence	0.86	Reading	0.40	0.03
Reading	0.87	Count dot arrays	0.37	0.03
Orientation in geography	0.90	Serial sevens	0.35	0.03
Repetition 2	0.91	Repetition 2	0.35	0.04
Identify fragmented letters	0.92	Draw intersecting pentagons	0.35	0.03

Note. SDRIR=Scottish Dementia Research Interest Register, IIO=Invariant item ordering, Mean=mean item score reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. H_i =item scalability coefficient indicating item *discrimination* with higher values associated with higher *discrimination*, SE=standard error, Naming 2: name 10 pictures, Repetition 2: repeat 'Above, beyond and below'

(iii) SDRIR late onset AD sample ($N=471$)

Using AISP to investigate item clusters within the scale determined that two items formed a separate item cluster; '3 item registration' and 'repetition 1' and 'follow written command-close eyes' did not fit conform to any cluster. These items were excluded.

Of the remaining 23 items all item-pair scalability coefficients were nonnegative and item scalability coefficients were greater than 0.3. With no violations of monotonicity these 23 ACE-R items meeting MHM criteria formed a moderate Mokken scale ($H=0.43$, $SE=0.02$).

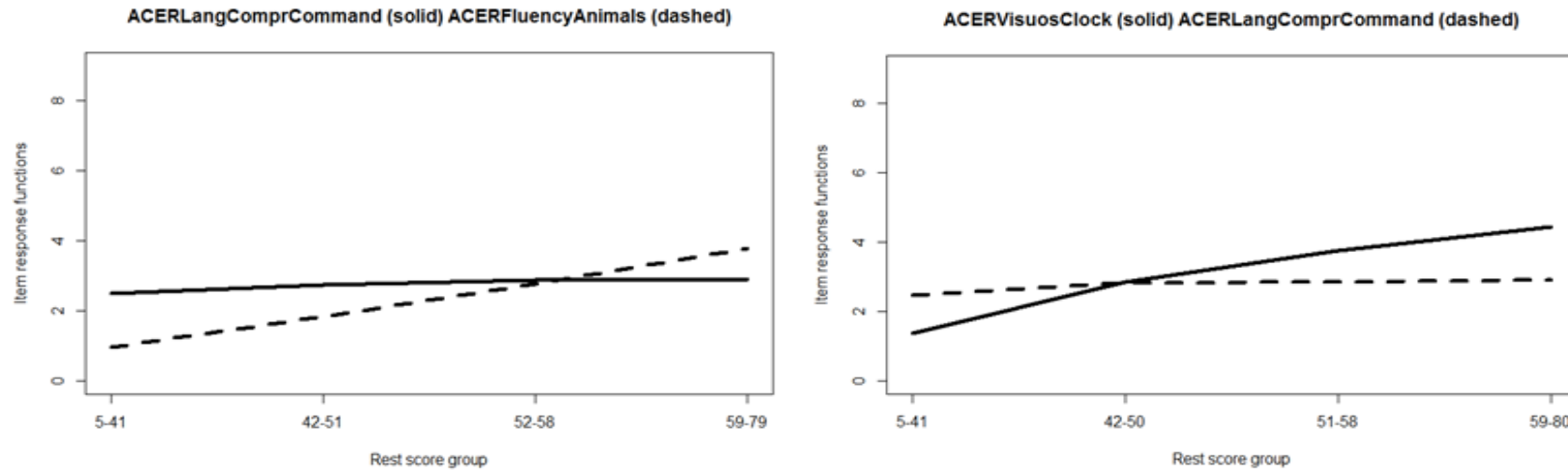
Confirmatory Mokken scaling analysis prompted the exclusion of two items due to low scalability coefficients; 'follow written command-close eyes' ($H_i=0.24$,

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

SE=0.06) and ‘repetition 1’ ($H_i=0.27$, SE=0.05). The item pair scalability coefficient of ‘3 item registration’ and ‘naming 1’ was negative. ‘3 item registration’ was selected for exclusion initially due to its lower scalability coefficient of the two (‘3 item registration’ $H_i=0.33$, ‘naming 1’ $H_i=0.43$). Following the removal of this item all item scalability coefficients were greater than 0.3 and all item-pair scalability coefficients were nonnegative. No exclusions were made in the assessment of monotonicity. The 23 remaining items formed a moderate Mokken scale with $H=0.43$ (SE=0.02).

The assessment of non-intersection led to the exclusion of nine items (see Figure 6.2 for examples of IIO violations resulting in exclusion of ‘draw a clock’ and ‘fluency-animals’ and ‘syntactical comprehension’). The remaining 14 items, (see Table 6.5) formed a reliable (MS = 0.85) moderate hierarchical scale ($H=0.43$, SE=0.02) with IIO ($H^T=0.81$).

Figure 6.2 Item pair plots showing IIO violations between ‘Syntactical comprehension’ (ACERLangComprCommand) and ‘Fluency-animals’ (ACERFluencyAnimals) and ‘Draw a clock’ (ACERVisuosClock) in late onset AD analysis



Note. IIO=Invariant item ordering, AD=Alzheimer’s disease, X-axis reflecting the ‘rest score group’ \approx latent trait

Table 6.5 SDRIR late onset AD: IIO hierarchy items listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
Name and address recall	0.06	Name and address recall	0.57	0.04
Memory retrograde	0.39	Identify fragmented letters	0.52	0.05
Recognition	0.49	Memory retrograde	0.48	0.02
Draw intersecting pentagons	0.61	Draw intersecting pentagons	0.48	0.04
Name and address learning	0.66	Semantic comprehension	0.46	0.03
Orientation in time	0.68	Naming 2	0.45	0.03
Repetition 3	0.69	Name and address learning	0.44	0.03
Semantic comprehension	0.70	Reading	0.42	0.05
Naming 2	0.77	Write a sentence	0.42	0.05
Count dot arrays	0.88	Recognition	0.39	0.03
Reading	0.88	Count dot arrays	0.38	0.04
Write a sentence	0.89	Repetition 2	0.37	0.06
Repetition 2	0.91	Orientation in time	0.36	0.03
Identify fragmented letters	0.93	Repetition 3	0.33	0.04

Note. SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer's disease, IIO=Invariant item ordering, Mean=mean item scores reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. H_i =item scalability coefficient indicating item *discrimination* with higher values associated with higher *discrimination*. SE=standard error, Repetition 2: repeat 'Above, beyond and below', Repetition 3: repeat 'No ifs, ands or buts', Naming 2: name 10 pictures

(iv) SDRIR combined early and late onset AD sample ($N=539$)

Three items were excluded in the assessment of the MHM; 'repetition 1' failed to form a cluster with any other ACE-R item in AISP, 'follow written command-close eyes' H_i coefficient fell below the 0.3 threshold and '3 item registration' was excluded due to a violation of monotonicity with a crit value greater than 40. The remaining 23 items meeting the criteria of the MHM formed a moderate Mokken scale ($H=0.45$).

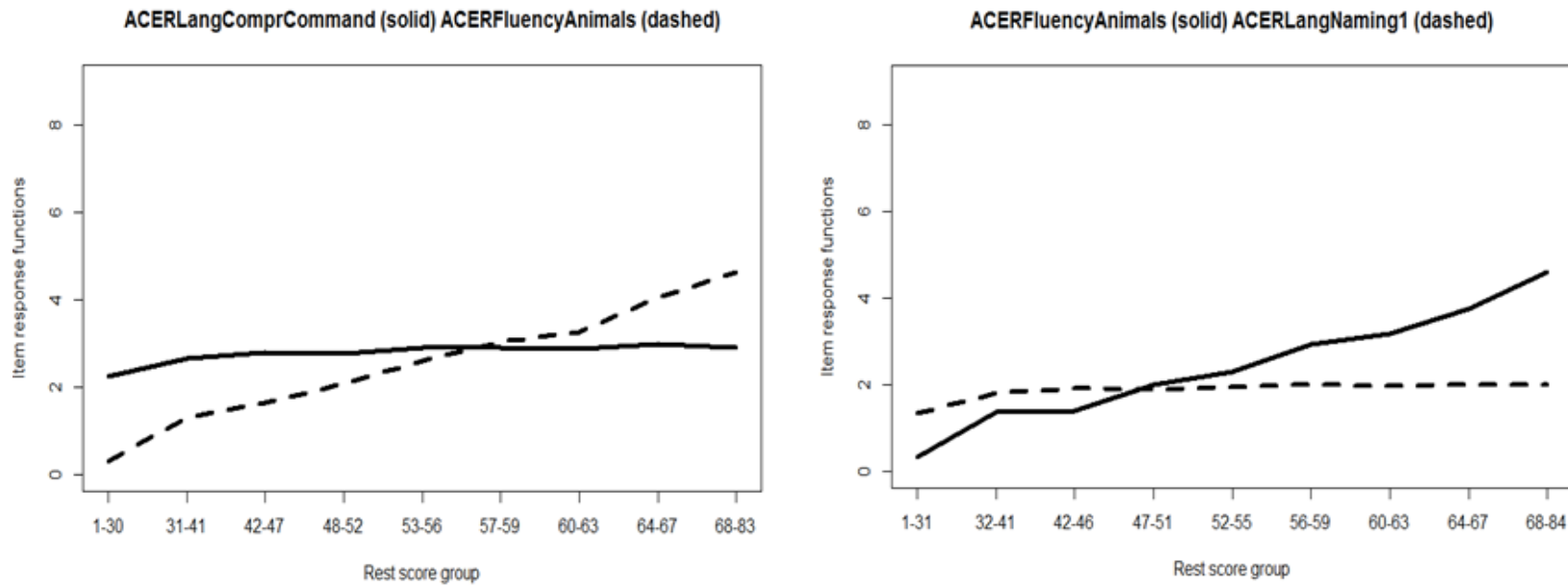
From a confirmatory scaling approach 'follow written command -close eyes' and 'repetition 1' were identified for removal from the scale due to their scalability coefficients ($H_i=0.29$, $SE=0.06$ and $H_i=0.28$, $SE=0.04$ respectively). The assessment

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

of monotonicity prompted the exclusion of '3 item registration' which left 23 items in a moderate scale ($H=0.45$, $SE=0.02$).

Thirteen items were excluded in the assessment of non-intersection (see Figure 6.3 for examples of items: 'fluency-animals', 'syntactical comprehension', 'naming 1' excluded due to IIO violations). Following these exclusions the remaining 10 items formed a reliable ($MS=0.83$), moderate hierarchical scale ($H=0.44$, $SE=0.02$) with IIO ($H^T=0.83$). These invariantly ordered items are listed in Table 6.6 from most *difficult* to least *difficult* and from most *discriminatory* to least *discriminatory*.

Figure 6.3 Item pair plots showing IIO violations between ‘Syntactical comprehension’ (ACERLangComprCommand) and ‘Fluency-animals’ (ACERFluencyAnimals) and ‘Fluency-animals’ and ‘Naming 1’ (ACERLangNaming 1) in combined early and late onset AD analysis



Note. IIO=Invariant item ordering, AD=Alzheimer’s disease, X-axis reflecting the ‘rest score group’ \approx latent trait

Table 6.6 SDRIR combined late and early onset AD: IIO hierarchy items from most to least *difficult* and most to least *discriminatory*

Items	Mean	Items	H_i	SE
Memory retrograde	0.42	Draw a clock	0.49	0.02
Recognition	0.50	Name and address learning	0.48	0.03
Draw intersecting pentagons	0.57	Write a sentence	0.47	0.04
Draw a clock	0.64	Reading	0.46	0.04
Name and address learning	0.65	Naming 2	0.45	0.03
Repetition 3	0.68	Serial sevens	0.43	0.03
Serial sevens	0.76	Memory retrograde	0.42	0.03
Naming 2	0.77	Recognition	0.38	0.03
Write a sentence	0.87	Draw intersecting pentagons	0.38	0.03
Reading	0.87	Repetition 3	0.36	0.03

Note. SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer's disease, IIO=Invariant item ordering, Mean=mean item score reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. H_i =item scalability coefficient indicating item *discrimination* with higher values associated with higher *discrimination*. SE=standard error, Repetition 3: repeat 'No ifs, ands or buts', Naming 2: name 10 picture

(v) SDRIR mixed AD VaD sample ($N=137$)

In the analysis of item properties in this sample nine items were initially identified for removal as they formed separate item clusters; scale 2: 'count dot arrays', 'identify fragmented letters', scale 3; 'repetition 3', 'reading', scale 4; 'name and address recall', 'recognition', scale 5; 'naming 1', 'follow written command-close eyes', 'syntactical comprehension'. Two items ('repetition 1' and 'repetition 2') failed to form any item cluster and as such were considered unscalable. These nine items were excluded due to the failure of these items to form a single Mokken scale. There were no further exclusions due to low scalability coefficients and all items met monotonicity assumptions. The 15 items partitioned to the main cluster-scale 1- formed a weak Mokken scale ($H=0.38$).

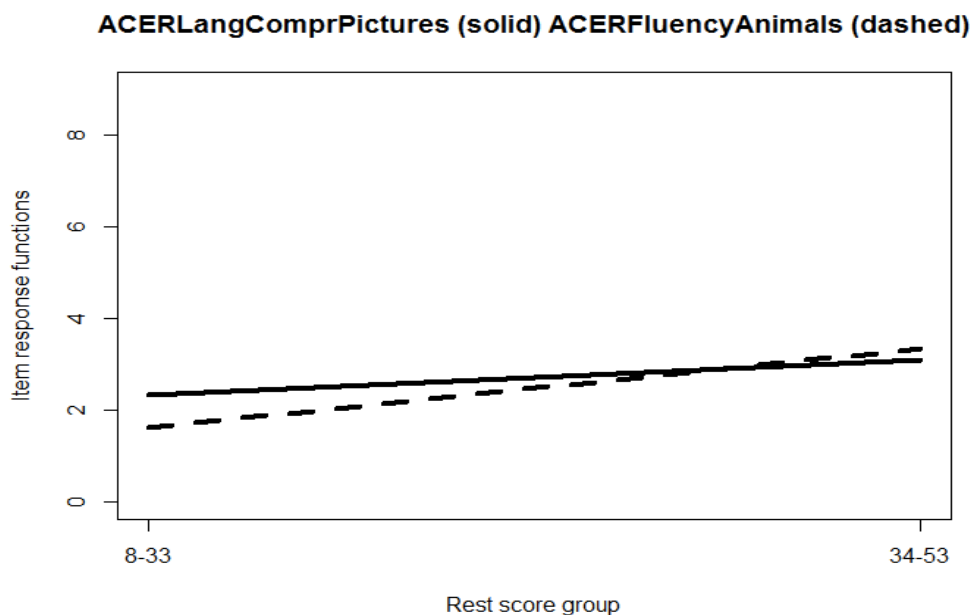
CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

Examining the items in this sample using confirmatory methods resulted in the exclusion of 10 items due to low scalability coefficients; ‘follow written command-close eyes’ ($H_i=0.27$, $SE=0.07$), ‘syntactical comprehension’ ($H_i=0.22$, $SE=0.09$), ‘repetition 1’ ($H_i=0.11$, $SE=0.06$), ‘repetition 2’ ($H_i=0.20$, $SE=0.09$), ‘repetition 3’ ($H_i=0.19$, $SE=0.06$), ‘naming 1’ ($H_i=0.15$, $SE=0.10$), ‘reading’ ($H_i=0.28$, $SE=0.07$), ‘count dot arrays’ ($H_i=0.19$, $SE=0.07$), ‘identify fragmented letters’ ($H_i=0.25$, $SE=0.05$) and ‘recognition’ ($H_i=0.26$, $SE=0.05$). The item pair scalability coefficient of ‘name and address recall’ and ‘write a sentence’ was negative. ‘Name and address recall’ with the lower H_i value (0.32 , $SE=0.07$) was removed first. Subsequent analysis of the scalability coefficients of the remaining items determined that the removal of ‘name and address recall’ was sufficient to increase the scalability coefficient of ‘write a sentence’. Therefore following these 11 exclusions the remaining items met MHM assumptions with $H=0.38$ ($SE=0.04$).

In the assessment of invariant item ordering one item (‘3 item registration’) was found to violate IIO. This item was removed accordingly. The remaining 14 items formed a reliable ($MS=0.86$) but weak hierarchical scale ($H=0.38$) with IIO ($H^T=0.77$). Inspection of the item pair plots revealed a slight intersection between ‘semantic comprehension’ and ‘fluency-animal’ (see Figure 6.4). These items were removed one at a time to inspect the effect of this IIO violation on H^T and H . Removing ‘semantic comprehension’ and re-analysing raised H^T to 0.79 and increased H to 0.39. The removal of ‘fluency-animal’ also increased H^T to 0.79 but lowered H to 0.36. Removing both items resulted in an increased accuracy of IIO ($H^T=0.80$) but decreased scalability ($H=0.37$). ‘Semantic comprehension’ was selected for removal as it resulted in higher scalability. The remaining 13 items shown in Table 6.7 formed a weak scale ($H=0.39$, $SE=0.04$) with very high accuracy of IIO

($H^T=0.79$). The invariantly ordered items for each sample are presented in Table 6.8 for comparison.

Figure 6.4 Item pair plot demonstrating intersection between 'Semantic Comprehension' (ACERLangComprPictures) and 'Fluency-Animals' (ACERFluencyAnimals) in the mixed AD VaD analysis.



Note. IIO=Invariant item ordering, AD=Alzheimer's disease, VaD= Vascular dementia, X-axis reflecting the 'rest score group' \approx latent trait

Table 6.7 SDRIR mixed AD VaD: IIO hierarchy items ordered from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
3 item recall	0.27	Fluency-animals	0.44	0.04
Fluency-animal	0.38	Write a sentence	0.43	0.07
Memory retrograde	0.45	Memory retrograde	0.43	0.04
Draw a cube	0.48	Orientation in geography	0.42	0.07
Fluency-letters	0.52	3 item recall	0.40	0.05
Orientation in time	0.57	Name and address learning	0.40	0.05
Draw intersecting pentagons	0.58	Fluency-letter	0.39	0.05
Draw a clock	0.66	Draw a clock	0.38	0.05
Name and address learning	0.73	Serial sevens	0.37	0.07
Serial sevens	0.80	Orientation in time	0.34	0.05
Naming 2	0.81	Naming 2	0.33	0.04
Write a sentence	0.88	Draw intersecting pentagons	0.32	0.06
Orientation in geography	0.90	Draw a cube	0.31	0.06

Note. SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer's disease, VaD= Vascular dementia, IIO=Invariant item ordering, Mean=mean item score reflecting item *difficulty*. Item scores range from 0-1 with lower scores indicating poor ability. H_i =item scalability coefficient indicating item *discrimination* with higher values associated with higher *discrimination*. SE=standard error, Naming 2: name 10 pictures

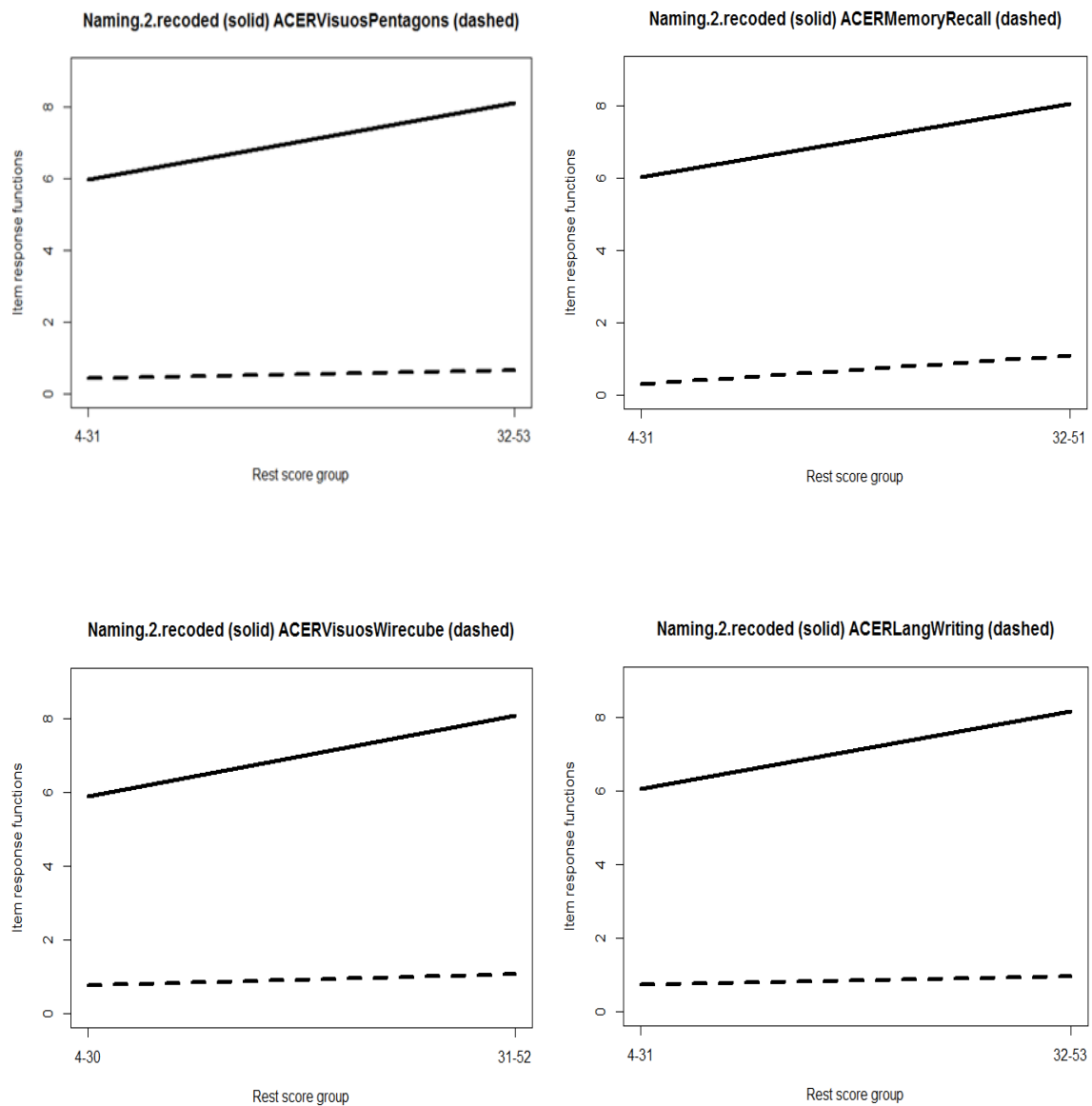
Table 6.8 ACE-R IIO hierarchies from most to least *difficult*: comparison across SDRIR groups

Full sample (N=808)	Late onset AD (n=471)	Combined late and early onset AD (n=539)	Mixed AD VaD (n=137)
Recognition	Name and address recall	Memory retrograde	3 item recall
Draw intersecting pentagons	Memory retrograde	Recognition	Fluency-animal
Semantic comprehension	Recognition	Draw intersecting pentagons	Memory retrograde
Naming 2	Draw intersecting pentagons	Draw a clock	Draw a cube
Count dot arrays	Name and address learning	Name and address learning	Fluency-letters
Write a sentence	Orientation in time	Repetition 3	Orientation in time
Reading	Repetition 3	Serial sevens	Draw intersecting pentagons
Orientation in geography	Semantic comprehension	Naming 2	Draw a clock
Repetition 2	Naming 2	Write a sentence	Name and address learning
Identify fragmented letters	Count dot arrays	Reading	Serial sevens
Naming 1	Reading		Naming 2
	Write a sentence		Write a sentence
	Repetition 2		Orientation in geography
	Identify fragmented letters		

Note. ACE-R=Addenbrooke’s Cognitive Examination-Revised, IIO= Invariant item ordering, SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer’s disease, VaD= Vascular dementia. Items common across all samples listed in bold. ‘Naming 1’: name pencil and watch, Naming 2: name 10 pictures, Repetition 2: repeat ‘Above, beyond and below’, Repetition 3: repeat ‘No ifs, ands or buts’,

Inspection of the item pair plots in each analysis identified that ‘naming 2’ (identifying 10 pictures, e.g. anchor, crown, penguin) is located some distance from the other items in all analyses. Outlying items can exaggerate the effect of IIO (Meijer & Egberink, 2012). Examination of these plots suggests that this apparently ‘outlying’ item could be causing an exaggerated IIO (see Figure 6.5 for example of item pair plots from the mixed AD VaD analysis). This does not make a lot of sense clinically as while it has been observed that both very low and high *difficulty* items can cause IIO due to the large difference between the mean scores of these extreme items and the remaining items, the mean score for ‘naming 2’ is neither the lowest or highest mean score in any of the analyses and consequently would not be considered ‘outlying’. However this item was removed from each analysis to determine whether this apparent outlying item was contributing to IIO. In the full sample removal and re-analysis raised the H^T value from 0.87 to 0.91. The removal of ‘naming 2’ from each of the other analyses resulted in a small decrease in the strength of IIO; late onset AD H^T decreased from 0.81 to 0.79, combined late and early onset AD H^T lowered from 0.83 to 0.76 and mixed AD VaD H^T lowered from 0.79 to 0.73 following the removal of ‘naming 2’ from the analysis.

Figure 6.5 Item pair plots demonstrating distance between ‘Naming 2’ (solid line) and other ACE-R items (dashed line) (‘Draw intersecting pentagons’, ‘3-item recall’, ‘Draw a cube’ and ‘Write a sentence’) in mixed AD VaD analysis.



Note. ACE-R=Addenbrooke’s Cognitive Examination-Revised, AD=Alzheimer’s disease, VaD= Vascular dementia, X-axis reflecting the ‘rest score group’ \approx latent trait, Naming.2.recoded=Naming 2, ACERVisuosPentagons=Draw intersecting pentagons, ACERVisuosWireCube=Draw a cube, ACERLangWriting=Write a sentence.

6.4 Discussion

This study used data collected from the SDRIR to investigate item properties and ordering of cognitive decline in dementia in a volunteer group in Scotland. The 808 included participants with a range of diagnoses, but the majority had been diagnosed with AD (early or late) or mixed dementia, which were analysed as separate subsamples. This allowed delineation of the sequence of cognitive impairment in different dementia groups. Results of the assessment of the fit of the items to the MHM using exploratory and confirmatory methods converged across all analyses. This supports the use of either method depending on the intended goal of the analyses; exploratory analyses can be applied to a large group of items where nothing is known about the existence of Mokken scales and confirmatory methods can be applied to test established Mokken scales against the criteria for Mokken scales (Watson et al., 2014). These analyses indicate that both modes of analysis are useful in gaining insight into these scales.

Of the 26 ACE-R items 11 items formed an IIO hierarchy in the full group, 14 items in the late onset AD sample, 10 in the combined AD sample and 13 in the mixed AD VaD sample. Within these IIO hierarchies the order of *difficulty* is the same for all respondents regardless of their cognitive ability or severity of dementia. For example according to these findings both patients with mild and severe late onset AD will find an assessment of retrograde memory more *difficult* than writing a sentence and both patients with mild and severe mixed AD VaD will find category fluency more *difficult* than letter fluency.

Hierarchical measures add another dimension to a scale's applicability.

Establishing an invariant pattern of decline can help to anticipate further decline and importantly can provide clinicians with a means to assess a patient from a few items within a scale. For example, a patient having difficulty drawing intersecting pentagons is likely to have trouble performing any of the more *difficult* tasks from the hierarchy. Equally patients recalling or recognising the name and address with ease are unlikely to have any trouble performing any of the other less *difficult* tasks in the hierarchical scale. This information can be applied to reduce the burden of testing on participants.

6.4.1 Invariant item ordering across diagnostic subgroups

Due to several exclusions in each analysis the comparison of item ordering across diagnostic groups is hampered by the lack of IIO items common across groups (see Table 6.8). Only three items feature in every IIO hierarchy; 'draw intersecting pentagons' (copy a drawing of intersecting pentagons), 'naming 2' (naming 10 items) and 'write a sentence' (make up and write a sentence). These items are in the same order of *difficulty* in each IIO hierarchy. However comparing the pattern of item *difficulty* across items common to two different groups can reveal double dissociations which can help to characterise distinctive cognitive profiles. Although this sample is biased towards the cognitive profile of AD there are some interesting differences in *difficulty* patterns, most notably between the mixed AD VaD and late onset AD groups. The patients with mixed AD VaD performed worse on the visuospatial task of 'drawing intersecting pentagons' than the patients with AD whereas the patients with AD performed worse on an assessment of 'orientation in time' (what is the day, date, month, year and season). These disparities can help to differentiate patients and contribute to differential diagnosis. According to these

results where ‘draw intersecting pentagons’ is more *difficult* than ‘orientation in time’ the pattern of item *difficulty* suggest mixed AD VaD may be more likely than late onset AD. Also AD patients performed worse on ‘name and address learning’ (learn name and address: Harry Barnes, 73 Orchard Close, Kingsbridge, Devon), whereas patients with mixed AD VaD performed worse on both ‘orientation in time’. This suggests that when ‘name and address learning’ is a more *difficult* item than ‘orientation in time’ late onset AD is more likely than mixed AD VaD.

However, these differences are slight in terms of mean scores which imply that while there are differences in item orderings these discrepancies may not remain in a larger analysis. Where item mean scores are close the order of *difficulty* may change in a different sample. It is also important to consider that differences between samples in terms of disease severity may be responsible for some of these findings. The mixed AD VaD group are slightly less impaired than the AD group as demonstrated by total ACE-R scores (see Table 6. 1). Replication with participants matched for disease severity is necessary.

The addition of patients diagnosed with early onset AD did not result in any variations in the item ordering from that found in patients with late onset AD. However, this may be a result of the greater proportion of patients with late onset AD who may have been driving the pattern of item performance. Fewer items conformed to an IIO scale in the combined analysis of early and late onset AD which suggests that the addition of patients with early onset AD led to increased variation in item responses. Further numbers of patients with early onset AD are required for separate analysis to identify whether this variability results in a different IIO pattern from late onset AD.

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

That 11 items conformed to an IIO hierarchy in the full data set suggest that the item ordering in this hierarchy may be impervious to the different diagnoses represented here suggesting the order of decline of these items in the full sample is common across all groups analysed. Consistent ordering between the full SDRIR group and both the combined AD and mixed AD VaD groups supports this. Between the full data and late onset AD there is only one slight change in the item ordering by *difficulty* with writing a sentence slightly more *difficult* than ‘reading’ (reading: sew, pint, soot, dough, height) (0.88-0.89) for the full group than late onset AD. This difference is so small it is unlikely to be clinically significant.

Comparing the item ordering here with that observed in Chapter 4’s analyses of IIO of the ACE-R in a different patient population reveals a consistent item ordering between the analysis of patients in the AD type sample with Alzheimer’s disease or logopenic progressive aphasia in Chapter 4 and the current chapter’s item ordering for late onset-AD. The five items common to each of these IIO hierarchies establish a sequence of decline starting with ‘name and address recall’ (recollection of the previously learned name and address), followed by decline in ‘memory retrograde’ (name the current Prime Minister, the woman who was Prime Minister, name of USA president and name of USA president who was assassinated in the 1960’s), ‘memory recognition’ (recognition of the previously learned name and address), ‘orientation in time’ and lastly ‘write a sentence’. The consistency of decline between these two samples offers a trajectory of decline that may be common to all with Alzheimer’s disease. However the small sample sizes and large standard errors of scalability coefficients in Chapter 4 should be noted here. Further analyses are required to determine the validity of this seemingly consistent pattern of decline in Alzheimer’s disease.

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

The IIO hierarchy of ACE-R items for the SDRIR full sample is restricted to the less *difficult* range of measurement with a mean score of 0.51 for the most *difficult* item ('recognition') conforming to the IIO scale. The hierarchical scales from the other three samples include items with a wider range of *difficulty* (late onset AD: 0.06-0.93, mixed AD VaD: 0.27:0.90 combined AD: 0.42-0.87). That the full sample IIO hierarchy does not include the more *difficult* items such as 'name and address recall', '3 item recall' (recall of lemon, key, ball) or 'fluency-animals' (naming as many animals as possible in one minute) could be due to the greater heterogeneity of the full sample which may have introduced greater variability of the pattern of decline in the more *difficult* items in the scale measuring the earlier stages of dementia.

Item-pair plots were examined visually to determine if there were any intersecting items or whether any item was driving IIO due to it being located far away from the other items. Inspection of item pair plots confirmed that 'naming 2' was located at some distance from the other scale items. Visual inspection also revealed intersection between 'semantic comprehension' (from an array of pictures identify the picture associated with the monarchy, the picture of a marsupial, the one which is found in the Antarctic, the one with the nautical connection) and 'fluency-animals' in the mixed AD VaD analysis. This emphasises the importance of considered analysis with visual inspection of each item within a Mokken scale before establishing IIO. The removal of 'naming 2' and 'semantic comprehension' lowered the strength of IIO from 0.79 to 0.73 in the mixed AD VaD analysis.

Items located far away from the other items should not necessarily be removed from a scale due to IIO violations as to do so would restrict the breadth of measurement of the scale which could result in the failure to observe important information and to detect changes in latent trait level. Removing these 'outlying'

items however can help to assess and improve the psychometric quality of the scale; allowing removal or alteration of items where necessary (Watson, Wang, Thompson & Meijer, 2014).

It should be noted that the differentiation of the clinical groups was based on independent diagnoses made by an old age-psychiatrist and physical supported by medical records and case notes which may not consistently correlate with neuropathological assessment (Harper et al., 2008). While ACE-R scores were not used in isolation to diagnose patients its application as a diagnostic tool for this sample and the subsequent analysis of ACE-R performance in the various diagnostic groups in these analyses raise some concerns regarding circularity of diagnosis and the primary study measure. To determine whether this had any influence on these findings additional research where ACE-R scores are not considered in the diagnosis of participants should be carried out.

6.4.2 Assessment and interpretation of item parameters

These findings demonstrate the advantage of examining individual items within cognitive scales in dementia through both exploratory and confirmatory Mokken scaling analyses. Identifying the *discrimination* of individual tasks and the level of ability they assess can add to understanding of disease progression as well as helping clinicians to quickly assess patients with key screening items; items with good *discrimination* at known levels of *difficulty*. For example, ‘name and address recall’ is the most *discriminatory* and most *difficult* item for the AD group which means for this group the ability to recall verbal information is lost very early and quickly at an early stage of decline. At the other end of the spectrum ‘identify fragmented letters’ is the least *difficult* and one of the most *discriminatory* items in this group, which means

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

that this item is very effective at measuring the advanced stage of disease. This ability is lost very late but quickly once this stage has been reached.

Examining the item properties of the IIO hierarchy for the diagnostically mixed sample reveals that the low *difficulty* items assessing high levels of severity ('naming 1' (naming pencil and watch) and 'identify fragmented letters' (identify degraded letters K, M, A, T) demonstrate high levels of *discrimination*. Towards the more *difficult* range of the IIO hierarchy; 'recognition' and 'draw intersecting pentagons' demonstrate adequate but low *discrimination*. While these items are the most *difficult* in the hierarchy the mean scores indicate that these deficits appear with moderately severe dementia. This means that the abilities assessed at this level are not lost as rapidly as the items assessing the more severe levels of dementia severity.

The items lost late in the combined early and late onset AD sample ('reading', 'write a sentence' and 'naming 2') are lost rapidly at this stage as indicated by the high levels of *discrimination*. The pattern of item *discrimination* is more mixed towards the mid-range of *difficulty* with these items associated with low *discrimination* ('draw intersecting pentagons' and 'repetition 3: no ifs, ands or buts') and high levels of *discrimination* ('draw a clock' and 'name and address learning').

In the mixed AD VaD IIO hierarchy the items at either range of the *difficulty* spectrum are associated with high levels of *discrimination*. For example, 'verbal fluency-animals' and 'memory retrograde', the mean scores and scalability coefficients both indicate that these impairments appear quickly towards the less severe levels of dementia. The mean scores of 'write a sentence' and 'orientation in geography' (which building, floor, town, county and country are you in?) along with their associated *discriminatory* values suggest that these abilities are retained until

high levels of dementia severity but are lost quickly once this stage is reached. These examples demonstrate how consideration of item parameters can provide clinically valuable insight.

The assessment of item properties and ordering is also of value psychometrically and can contribute towards the development of meaningful and effective assessment tools by identifying candidate items and also those which add little to the sensitive assessment which should be removed. Items removed due to insufficient *discrimination* include; ‘following a written command-close eyes’ and the verbal repetition items. These items were excluded from each of the current analyses. ‘Follow written command-close eyes’ and verbal repetition have been shown to lack sensitivity to cognitive impairment (Brugnolo et al., 2009). These items were among those excluded due to low scalability in Chapter 4. ‘Follow written command-close eyes’ has been removed from the latest version of the ACE, the ACE-III, due to this poor sensitivity (Hsieh et al., 2013). It may be that these items are more susceptible to other factors such as hearing or vision and as such are not as sensitive to cognitive impairment and by consequence would not be as predictive of performance in the other items. Item level analysis can reveal and confirm such weaknesses within established scales and these results demonstrate how Mokken scaling analysis can identify which items should be removed from the scale due to their low contribution or poor association with the latent construct.

IRT methods providing item *difficulty* and *discrimination* parameters permit the examination of item *difficulty* distribution which helps to confirm the range of cognitive abilities the scale assesses and how well it can differentiate cognitive impairment levels in a particular range (Spector & Fleishman, 1998). Identifying large gaps between item *difficulties* can help to determine regions where the latent trait is

assessed with relatively lower levels of precision and reliability. This can prompt the addition of items with the required *difficulty* levels to provide measurement across the range of abilities. Inspecting the item *difficulties* of the ACE-R suggests that the scale is well equipped in the identification of cognitive impairment and changes in ability in moderate-severe dementia. This range of ability is assessed by many low *difficulty* items such as ‘3 –item registration’ (repeat lemon, key ball), ‘identify fragmented letters’ and ‘orientation in geography’.

As ‘name and address recall’ consistently had the lowest mean score the use of this item contributes heavily to the identification of changes in cognition in the early stages of cognitive decline. Following this item there is a gap in measurement with a mean item differences of between 0.16 and 0.19 between this item and the item with the next highest level of *difficulty*; ‘3 item recall’. This gap in assessment could result in the failure to detect subtle changes in ability which the addition of more high *difficulty* items could help to address.

6.4.3 Limitations and methodological considerations

The analyses in this chapter were carried out using larger sample sizes than those analysed in Chapter 4, which resulted in fewer items, excluded due to low scalability coefficients. In the three largest samples analysed in the present chapter no greater than three items were excluded due to violations of MHM assumptions. The number of items with low scalability coefficients can reflect sample size (Straat, 2012).

Examining scalability parameters to determine adequate sample size here indicates that the sample size of the mixed AD VaD analysis in this Chapter where 11 items failed to meet the minimum criteria for MHM is too small. Furthermore, the standard errors for several items within this sample were large. Replication in a larger sample

is required to determine the reliability of the results here and whether these item exclusions are due to the small sample size.

The exploratory procedure applied here determined that ‘repetition 1: repeat multi-syllabic words’ was unallocated to any scale in the full sample, combined late and early onset AD and mixed AD VaD samples and the poor scalability of this item was confirmed by examination of its item scalability coefficient in the confirmatory analyses. This indicates that this item is a weak item (Smits, Timmerman & Meijer, 2012). Other items warranting further investigation with regards to their contribution to the accurate assessment of cognitive impairment in dementia include ‘repetition 2: repeat Above, beyond and below’ and ‘follow written command-close eyes’. These items were also among those meeting criteria for exclusion due to low scalability coefficients using confirmatory Mokken scaling in Chapter 4. This demonstrates that practically speaking both analytic approaches are complementary.

Due to some concerns regarding violations of local stochastic independence items where suspected violations were likely to occur were removed from the analysis to determine whether LSI was influencing the high H^T values found. However, the H^T value of the 11 item IIO scale confirmed from this analysis was not any lower than the values from the analyses of all ACE-R items. Results from this additional analysis suggest that the items removed (‘3 item registration’, ‘3 item recall’, ‘name and address learning’, ‘name and address recall’ and ‘name and address recognition’) are not responsible for the elevated H^T values.

From the SDRIR data available ($N = 1248$) data from 327 participants were excluded from analyses due to missing ACE-R data leaving 921 participants. As the majority of exclusions were due to a complete lack of any ACE-R data it was not

possible to determine whether there were any important differences between those analysed and those who were excluded. A further 113 cases were excluded due to their diagnostic classification (other dementia, uncertain diagnosis or mild cognitive impairment). As the sample in this Chapter is a selected clinical sample, rather than a representative epidemiological sample this does not change the validity of analyses within the sample however it may limit the generalizability of the findings.

6.5 Conclusion

This chapter is primarily concerned with the pattern of cognitive decline in patients with dementia and establishing if this progressive deterioration differs in patients with Alzheimer's disease both early and late onset and mixed Alzheimer's disease and vascular dementia. Methodological concerns regarding sample size and violations of local stochastic independence were also addressed. The results of this chapter demonstrate the potential for 'name and address recall', as a highly *discriminatory* high *difficulty* item, to be used as an indicator of the initial stages of cognitive decline in Alzheimer's disease. This makes sense clinically as most clinicians will expect poor performance on this item first where AD is suspected. 'Identify fragmented letters' was also identified as an effective item assessing the advanced stages of Alzheimer's disease.

Mokken analysis indicates that an 11-item subset of ACE-R items form a common hierarchy of cognitive decline for a heterogeneous dementia sample. These hierarchical patterns can provide predictive insights to clinicians monitoring patients and the sequences of decline and variations between them can help to characterise cognitive differences among diagnostic subgroups. This 11 item subset of ACE-R items could also be examined to determine whether this hierarchy could yield a brief

CHAPTER 6: HIERARCHICAL PATTERNS IN ACE-R

shortened ACE-R scale as in the case of the Mini-ACE development. This will be explored in Chapter 7.

Chapter 7: Development and validation of the Short ACE-R

7.1 Introduction

In Chapter 5 Mokken scaling methods were used to derive the five item Mini-ACE, a shortened version of the ACE-III, from a clinical sample attending the Frontier Research Group, Sydney, the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom. The properties of this shortened scale, the Mini-ACE, were then assessed by Mokken scaling analysis of the scale using data from an independent clinical sample, also collected by the Sydney group.

In this Chapter the same methodology was used with data from the ACE-R measured in the Scottish Dementia Research Interest Register (SDRIR) to derive a shortened version of the ACE-R. The objective was to determine whether the items selected for the Mini-ACE would be selected for this new Short ACE-R scale and if not, to examine the item properties of each scale to determine which scale had the best clinical application as a brief screening tool. This new five-item scale will also be validated using the same independent clinical sample used to validate the Mini-ACE.

In some ways this chapter replicates Chapter 5 in terms of design and methodology; that is; it uses Mokken scaling analysis to select items for a new brief scale and performs Mokken scaling analysis to validate this new scale using data from an independent sample. However, the sample used to develop the Short ACE-R in this Chapter differs from the population used in Chapter 5 to derive the Mini-ACE in terms of dementia diagnoses and severity, geographical location, age, and sample size. Table 7.1 provides a comparison of the different samples. The Mini-ACE was developed using a sample from specialised tertiary

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

memory clinics in Sydney, Cambridge and Oxford, which due to the particular research interests of these clinics, comprised a greater preponderance of less common forms of dementia such as frontotemporal dementia and progressive aphasia. The scale development sample used in this Chapter is drawn from the SDRIR, which is more representative of the general population with the majority of patients on the register diagnosed with Alzheimer's disease (AD).

This Scottish sample also has a higher mean age (77.5 years) and a substantially larger sample size ($N=808$) than the Mini-ACE development sample (mean age; 65.4 years, $N=117$). Furthermore the Mini-ACE was developed from analysis of the ACE-III with subsequent validation performed using data from the ACE-R, whereas the scale developed in this Chapter will be derived from the ACE-R and analysed using data from the ACE-R.

7.2 Method

The SDRIR sample from which this new scale is to be selected from will be referred to as the Short ACE-R development sample in this Chapter. This sample was used previously in Chapter 6 in the Mokken scaling analysis of the ACE-R. The findings of the analysis in Chapter 6 form the basis for item selection in the current Chapter. The sample used to validate the new Short ACE-R was originally used in Chapter 4 in the Mokken scaling analysis of the ACE-R and to validate the Mini-ACE in Chapter 5 and will be referred to in this Chapter as the Short ACE-R validation sample. Figure 7.1 presents a visual overview of the relevant samples in this Chapter. Figure 7.2 presents a flowchart illustrating the processes of the Short ACE-R development in this Chapter.

Table 7.1 Comparison of the different scale development and validation samples

	Development Samples		Validation samples
	Mini-ACE	Short ACE-R	Mini-ACE & Short ACE-R
N	117	808	350
Age (SD)	65.4 (8.5)	77.5 (7.8)	65.4 (8.5)
ACE (SD)	63.6 (20)	63 (16.8)	63.6 (19.4)
Location	Sydney, Oxford, Cambridge	Scotland	Sydney
Scale	ACE-III	ACE-R	ACE-R
Diagnosis	AD (34), bvFTD (25), CBD (9), PPA (49)	Late AD (471), mixed AD/VaD (137), VaD (89), early AD (68), DLB (20), FTD (14), PDD (9)	bvFTD (96), AD (88), SD (61), LPA (43), PNFA (39), FTD-MND (23)

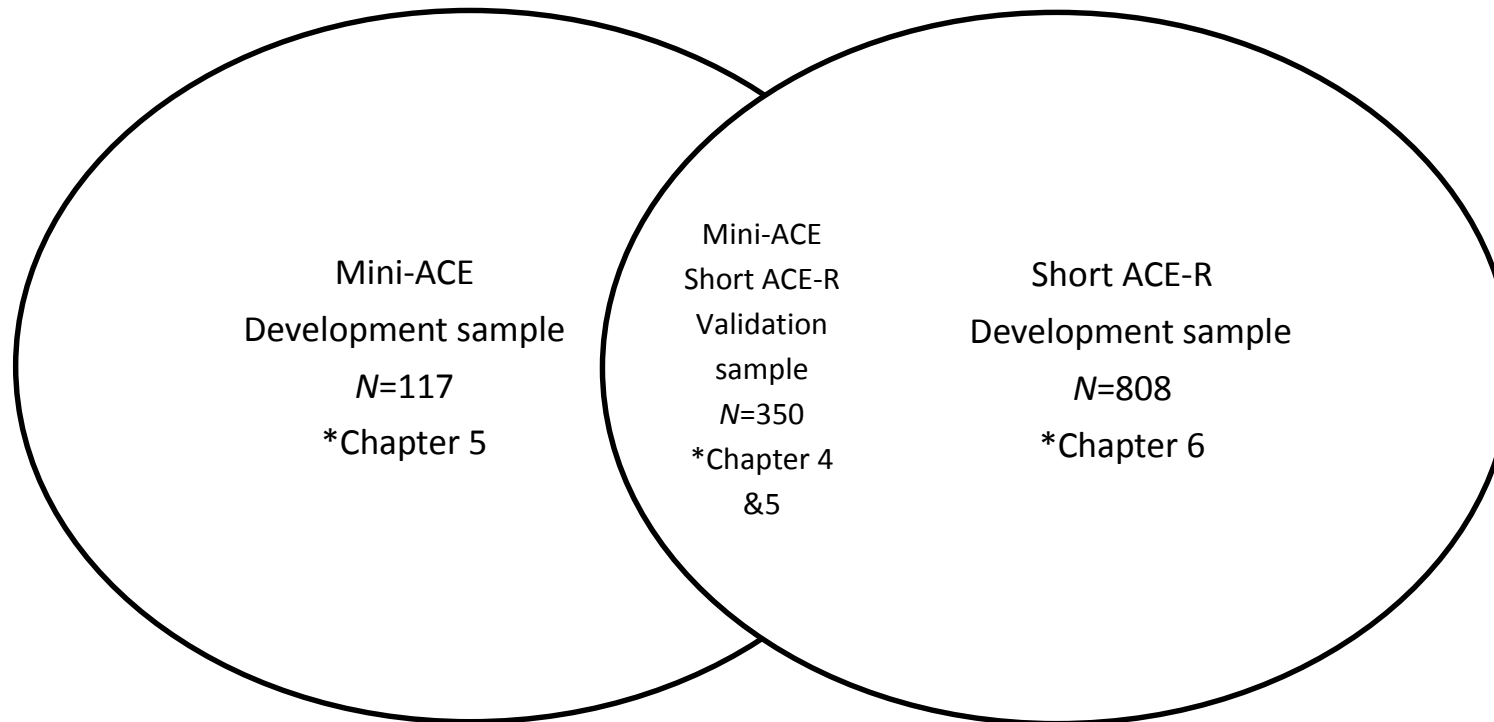
Note. Mini-ACE=Mini Addenbrooke's Cognitive Examination, Short ACE-R=Short Addenbrooke's Cognitive Examination-Revised, ACE-III= Addenbrooke's Cognitive Examination III. ACE-R= Addenbrooke's Cognitive Examination-Revised, AD=Alzheimer's disease, bvFTD=behavioural variant frontotemporal dementia, CBD= corticobasal degeneration, PPA= progressive primary aphasia, Late AD=late onset Alzheimer's disease, mixed AD/VaD=mixed Alzheimer's disease Vascular dementia, VaD=Vascular dementia, early AD=early onset Alzheimer's disease, DLB= dementia with Lewy bodies, FTD=frontotemporal dementia, PDD= Parkinson's disease dementia, SD=semantic dementia, LPA= logopenic progressive aphasia, PNFA= progressive nonfluent aphasia, FTD-MND=frontotemporal dementia with motor neurone disease.

7.2.1 Participants

Short ACE-R development sample

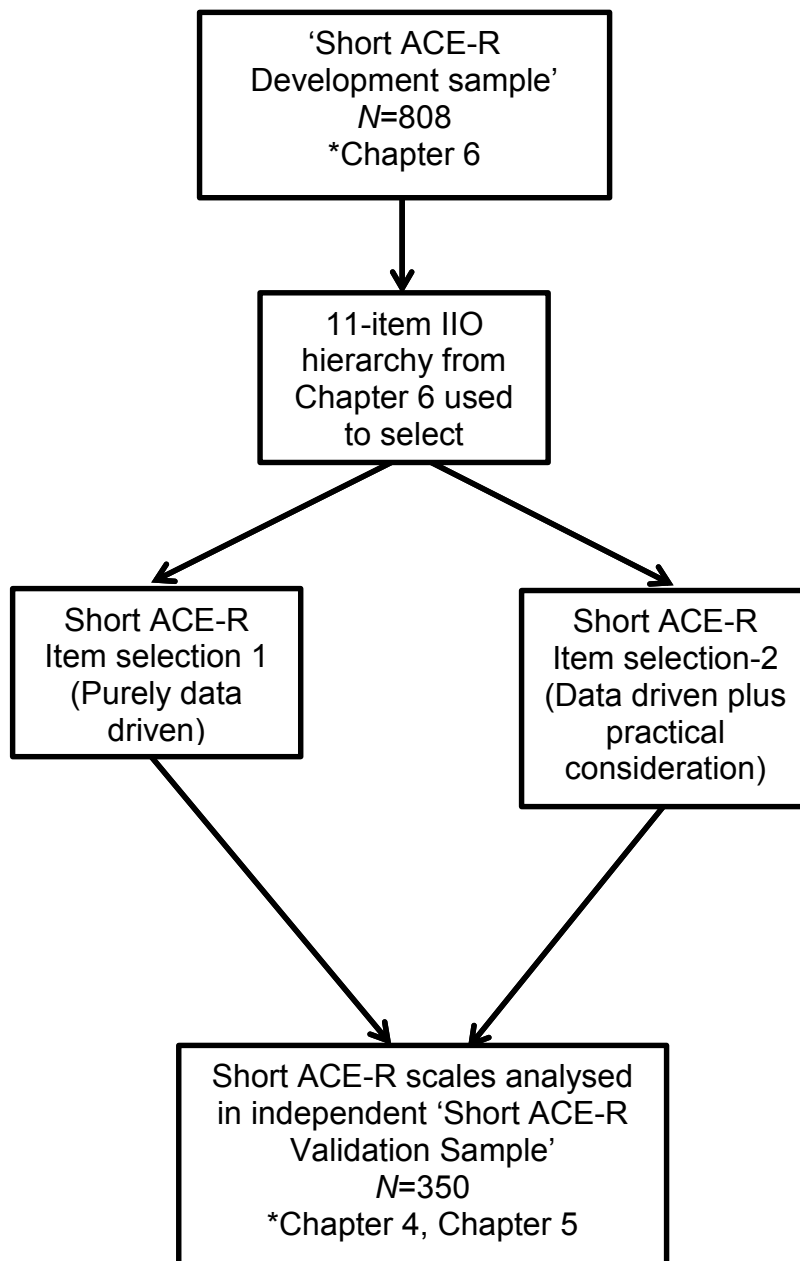
The Short ACE-R development sample was drawn from the Scottish Dementia Clinical Research Register (SDRIR). This sample was previously described in Chapter 6. This sample ($N=808$; mean age=77.5 (7.8), 425 (52.6%) male) comprised patients in seven diagnostic groups; late-onset Alzheimer's disease (AD), $n=471$; mixed Alzheimer's disease/Vascular dementia (mixed AD/VaD), $n=137$; VaD, $n=89$; early-onset AD, $n=68$; frontotemporal dementia (FTD), $n=14$; and Parkinson's disease dementia (PDD), $n=9$; mean ACE-R=63 (16.8)

Figure 7.1 Venn diagram illustrating the different samples used in this Chapter



Note. Mini-ACE=Mini Addenbrooke’s Cognitive Examination, Short ACE-R=Short Addenbrooke’s Cognitive Examination-Revised, *=Chapter where the sample was originally used, Mini-ACE development sample collected in Sydney, Oxford and Cambridge was originally used in the development of the Mini-ACE in Chapter 5, Short ACE-R development sample collected in Scotland from the SDRIR was originally used in Chapter 6, the validation sample for both Mini-ACE and Short ACE-R collected in Sydney was originally used in Chapter 4 and was also used in Chapter 5 for the Mini-ACE validation analysis

Figure 7.2 Flowchart illustrating the methods and samples used in this Chapter



Note. IIO= Invariant item ordering, ACE-R= Addenbrooke's Cognitive Examination-Revised. *=Chapter were sample previously used. Short ACE-R Development Sample of SDRIR participants was originally used in Chapter 6. The Short ACE-R Validation sample collected in Sydney was originally used in Chapter 4 and was also used in Chapter 5 for the Mini-ACE validation analysis.

Short ACE-R validation sample

The validation sample for this Chapter was also used to validate the Mini-ACE in Chapter 5. It comprises 350 participants from the Frontier Research Group in Sydney with dementia; bvFTD, $n = 96$; AD, $n = 88$; semantic dementia (SD), $n=61$; logopenic progressive aphasia (LPA), $n=43$; PNFA, $n=39$ and frontotemporal dementia with motor neurone disease, $n=23$; mean age=65.4 (8.5); mean ACE-R=63.6 (19.4).

7.2.2 Measures

The ACE-R, scored out of 100, includes 26 items across five cognitive domains: attention, memory, fluency, language and visuospatial (see Appendix A). Again as in previous analyses of ACE data (Chapters 4, 5 and 6) the response to each of the 26 items was equated for analysis whereby the mean item scores were divided by the range. For example, the mean score for ‘identify fragmented letters’ (3.68) from the Short ACE-R development sample was divided by 4, the maximum number of points available for this item to provide an equated mean score of 0.92.

7.3 Item selection

Mokken scaling analysis of the full SDRIR sample ($N=808$) (Chapter 6) identified 11 items that conformed to an IIO hierarchy (see Table 7.2). These 11 items were used as the basis for item selection for the Short ACE-R in the same way as the 17-item IIO hierarchy of ACE-III items was used to derive the Mini-ACE in Chapter 5. As in the development of the Mini-ACE the criteria for retaining items for the Short ACE-R were i) including one item from each cognitive domain (attention, memory, fluency, language and visuospatial skills) and ii) ensuring high *discrimination* at various levels of *difficulty*.

From this IIO scale two selections of five items were chosen. In both cases item selection was driven by the desire to cover a broad range of the cognitive domains of the ACE-R. In addition to this goal the first selection was based solely on assessment of item *difficulty* and *discrimination* whereas the second selection also considered the practicalities of test administration.

7.3.1 Short ACE-R selection 1

In the memory domain there was only one item from this domain in the IIO scale—‘recognition’ (see Table 7.2). This item was therefore chosen for inclusion in the Short ACE-R. ‘Recognition’ is a suitable item for inclusion as it is the most *difficult* item in the IIO hierarchy (mean=0.51) and therefore is valuable in assessing the initial stages of impairment. The *discrimination* value is among the lowest in the hierarchy ($H_i=0.36$). This means that while ‘recognition’ measures the upper end of the hierarchy it may not help to indicate differences in ability in this more advanced stage of disease as well as items with higher *discrimination*. However no alternative high *difficulty* item was available and furthermore the item’s inclusion in a hierarchical Mokken scale implies it has adequate *discrimination*. Therefore ‘recognition’ was selected to assess memory at the upper range of dementia severity.

Again focusing on selecting items assessing the breadth of the domains, ‘orientation in geography’ was identified as being the only item from the attention domain respectively (Table 7.2). ‘Orientation in geography’ was a relatively low *difficulty* item (mean=0.90) and had good *discrimination* ($H_i=43$). This item was selected as it assesses attention with good *discrimination* at a level of relatively low *difficulty*. Therefore ‘orientation in geography’ adds support to the assessment of a more severe level of impairment.

Table 7.2 IIO hierarchy of ACE-R items (from analysis of SDRIR data ($N=808$) in Chapter 6) listed in descending order of *difficulty* and *discrimination*.

Domain	Item	Mean	Domain	Item	H_i	SE
Memory	Recognition ^{1,2}	0.51	Visuospatial	Identify fragmented letters ^{1,2}	0.47	0.03
Visuospatial	Draw intersecting pentagons	0.56	Language	Naming 1	0.46	0.04
Language	Semantic comprehension ^{1,2}	0.72	Language	Semantic comprehension ^{1,2}	0.46	0.03
Language	Naming 2	0.78	Language	Naming 2	0.45	0.03
Visuospatial	Count dot arrays	0.86	Attention	Orientation in geography ^{1,2}	0.43	0.03
Language	Write a sentence	0.86	Language	Reading	0.38	0.03
Language	Reading	0.87	Language	Write a sentence	0.38	0.03
Attention	Orientation in geography ^{1,2}	0.90	Visuospatial	Count dot arrays	0.37	0.03
Language	Repetition 2	0.91	Memory	Recognition ^{1,2}	0.36	0.03
Visuospatial	Identify fragmented letters ^{1,2}	0.92	Language	Repetition 2	0.35	0.04
Language	Naming 1 ¹	0.95	Visuospatial	Draw intersecting pentagons	0.34	0.03

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised. SDRIR=Scottish Dementia Clinical Research Interest Register. SE=standard error. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*. ¹=Item selection 1, ²=Item selection 2

In the language domain ‘naming 1’ was identified as a candidate item for the new scale. ‘Naming 1’ is the least *difficult* item in the hierarchical scale (mean=0.95) and is also amongst the most *discriminatory* ($H_i=0.46$). This item is a good candidate for selection as it extends the breadth of measurement of the Short ACE-R to the lower ranges of ability and measures this level of severe impairment very succinctly as is reflected by its high *discrimination* value (see Table 7.2).

The two visuospatial items in the IIO hierarchy; ‘draw intersecting pentagons’ and ‘identify fragmented letters’ differ considerably in their *discrimination* (0.34 and 0.47 respectively) and *difficulty* (0.56 and 0.92 respectively). ‘Identify fragmented letters’ was deemed the more suitable for selection due to its higher *discrimination*.

Neither of the fluency items was present in the IIO hierarchy. ‘Semantic comprehension’ was considered the most appropriate substitution for a fluency item. This item was chosen to balance out the coverage in terms of *difficulty* (mean=0.72) and as it had the highest *discrimination* of the remaining items ($H_i=0.46$).

Therefore the first data driven selection comprised: ‘orientation to geography’, ‘naming 1’, ‘semantic comprehension’, ‘identify fragmented letters’ and ‘recognition’. These items were selected as the best choice of items from a range of cognitive domains that spanned a reasonable range of *difficulty* (range: 0.51-0.95) with good *discrimination* (all >0.36) (see Table 7.2).

The item *difficulty* and *discrimination* of items from the Short ACE-R selection 1 is presented in Table 7.3 with items listed from most to least *difficult*. The level of *difficulty* assessed by the Short ACE-R item selection 1 is presented in Figure 7.3. This figure in

addition to the mean scores in Table 7.3 demonstrates the restricted range of measurement with the absence of high *difficulty* items assessing the upper ranges of ability.

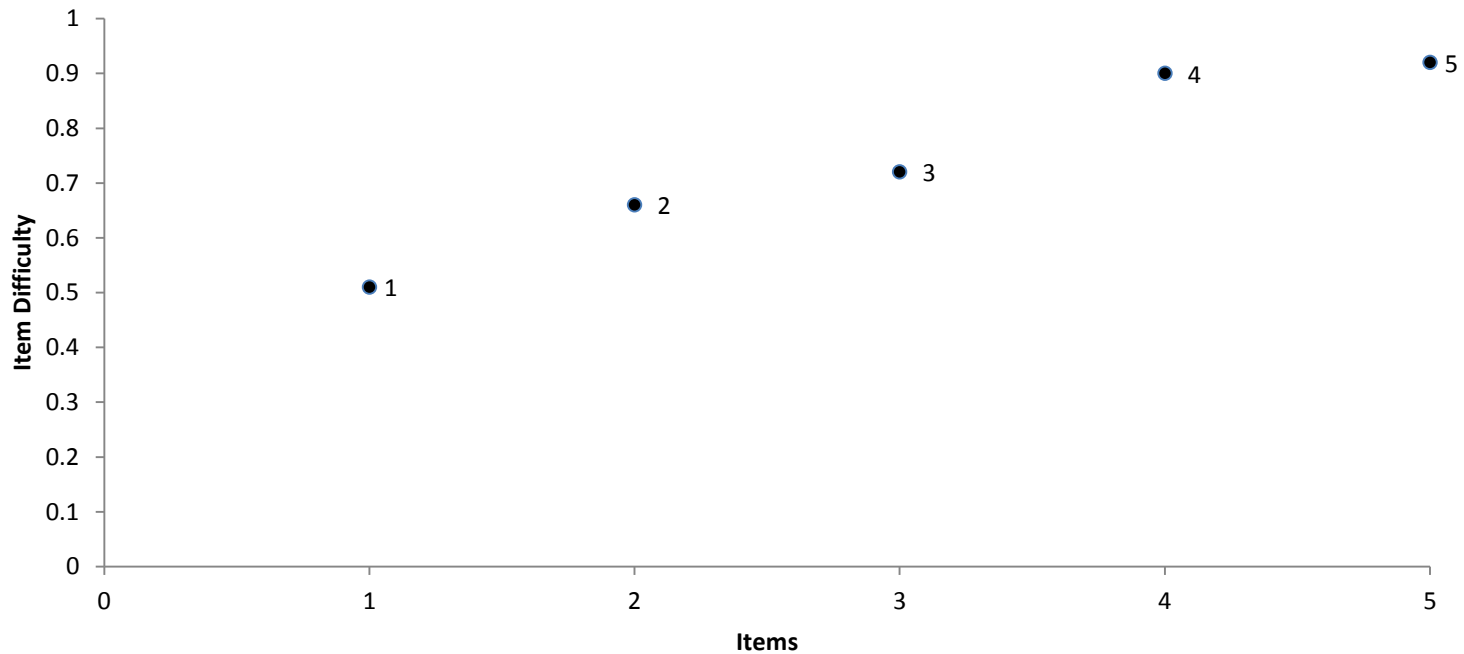
Item selection 1 was purely data driven, using the psychometric criteria to guide item selection. However, when the content and practical consequences of the items were considered, a fundamental issue was identified which meant that these items did not form a practical scale. ‘Recognition’ assesses the patients’ ability to recognise a previously learned name and address. Testing ‘recognition’ of the name and address necessitates the inclusion of ‘name and address learning’. As the data driven method of item selection does not include ‘name and address learning’, the Short ACE-R with only the 5 items identified in Table 7.3 could not be administered. ‘Name and address learning’ could be added to the items, giving the Short ACE-R six items, which means it would not fulfil the initial objectives of selecting a five item scale. This would have reduced the comparability between the Short ACE-R and the Mini-ACE. Therefore an additional item selection was made; one which considered both practical issues and psychometric item properties as any scale development process should.

Table 7.3 Short ACE-R item selection 1 based on results of Mokken scaling analysis of SDRIR data ($N=808$)

Domain	Item	Mean	H_i
Memory	Recognition	0.51	0.36
Language	Semantic comprehension	0.72	0.46
Attention	Orientation in geography	0.90	0.43
Visuospatial	Identify fragmented letters	0.92	0.47
Language	Naming 1	0.95	0.46

Note. ACE-R=Addenbrooke’s Cognitive Examination-Revised. SDRIR=Scottish Dementia Clinical Research Interest Register. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*. Naming 1=Naming pencil and watch. See Appendix A for full ACE-R item format and question wording.

Figure 7.3 Range of item *difficulty* of the Short ACE-R item selection 1 for the Short ACE-R development sample.



Note. The y-axis represents the mean item scores reflecting item *difficulty* (with higher mean values indicating lower *difficulty*). The x-axis represents the number of items in the scale. 1=Recognition (0.51); 2=Semantic Comprehension (0.72); 3=Orientation in Geography (0.90); 4=Identify fragmented letters (0.92); 5=Naming 1 (0.95). Short ACE-R=Short Addenbrooke's Cognitive Examination-Revised.

7.3.2 Short ACE-R selection 2

Item selection 1 focused solely on selecting items the most *discriminatory* items assessing the breadth of *difficulty* across the domains (i.e. selecting the most appropriate item—the most *discriminatory*—from each cognitive domain with an adequate spread of item *difficulty* to provide measurement along the spectrum of abilities).

Considering test practicalities supported the decision to include ‘identify fragmented letters’ as opposed to the alternative visuospatial item; ‘draw intersecting pentagons’ in item selection 1. The inclusion of ‘draw intersecting pentagons’ would necessitate the use of a pencil and paper which adds to the complexity of testing and may prevent certain patient groups (stroke, arthritis, those in severe pain) being able to complete the test.

However it is apparent that item selection 1 is of no practical use as one of the items; ‘recognition’ (the recognition of a name and address) cannot be responded to as the item selection does not provide respondents an opportunity to familiarise themselves with this name and address (i.e. it does not include the item ‘name and address learning’).

Therefore a second item selection taking the practicalities of test administration into account was made. Firstly ‘name and address learning’ was incorporated into the scale. It should be noted that this item did not conform to the IIO hierarchy revealed in Chapter 6 which formed the basis for item selection here. However the *discrimination* value of this item revealed in the analysis in Chapter 6 ($H_i=0.43$) is sufficiently high to avoid concerns regarding this item’s ability to contribute to meaningful assessment in the scale. As the objective was to identify five candidate items the addition of ‘name and address learning’ meant one of the previously selected items needed to be excluded.

Examining the remaining items from item selection 1 (‘semantic comprehension’, ‘orientation in geography’, ‘identify fragmented letters’ and ‘naming 1’) the decision to

remove one of the two language items was made as to eliminate any of the other items would exclude the assessment of one of the cognitive domains of the ACE. Of these two items; ‘naming 1’ and ‘semantic comprehension’, ‘naming 1’ had the poorer *discrimination*. Also the very low *difficulty* of ‘naming 1’ was very similar to that of ‘orientation in geography’ and ‘identify fragmented letters’. This cluster of item *difficulty* at the more severe levels of impairment is not appropriate for a brief dementia-screening tool. Therefore ‘naming 1’, rather than ‘semantic comprehension’ which has a higher *difficulty* level, was deemed the more appropriate item for removal from the scale.

With this item substitution, the second, more practically considered selection comprised: ‘orientation in geography’, ‘name and address learning’, ‘semantic comprehension’, ‘identify fragmented letters’ and ‘recognition’. This changes the hierarchy of *difficulty*. The least *difficult* item in this scale is now ‘identify fragmented letters’ and the *difficulty* level of the newly added ‘name and address learning’ places it as the second most *difficult* item in the hierarchy. However it must be noted that not all of these items were selected from a formal IIO hierarchy. ‘Name and address learning’ was included for content coverage despite not conforming to the IIO hierarchy revealed in the analysis in Chapter 6.

The item *difficulty* and *discrimination* of items from the Short ACE-R selection 2 is presented in Table 7.4 with items listed from most to least *difficult*. Figure 7.4 illustrates the measurement range of this scale and, like Figure 7.3, demonstrates the gap in measurement beyond a *difficulty* level of 0.50, which limits the assessment of the early stages of cognitive impairment.

Table 7.4 Short ACE-R item selection 2 based on results of Mokken scaling analysis of SDRIR data ($N=808$)

Domain	Item	Mean	H_i
Memory	Recognition	0.51	0.36
Memory	Name and address learning*	0.66	0.43
Language	Semantic comprehension	0.72	0.46
Attention	Orientation in geography	0.90	0.43
Visuospatial	Identify fragmented letters	0.92	0.47

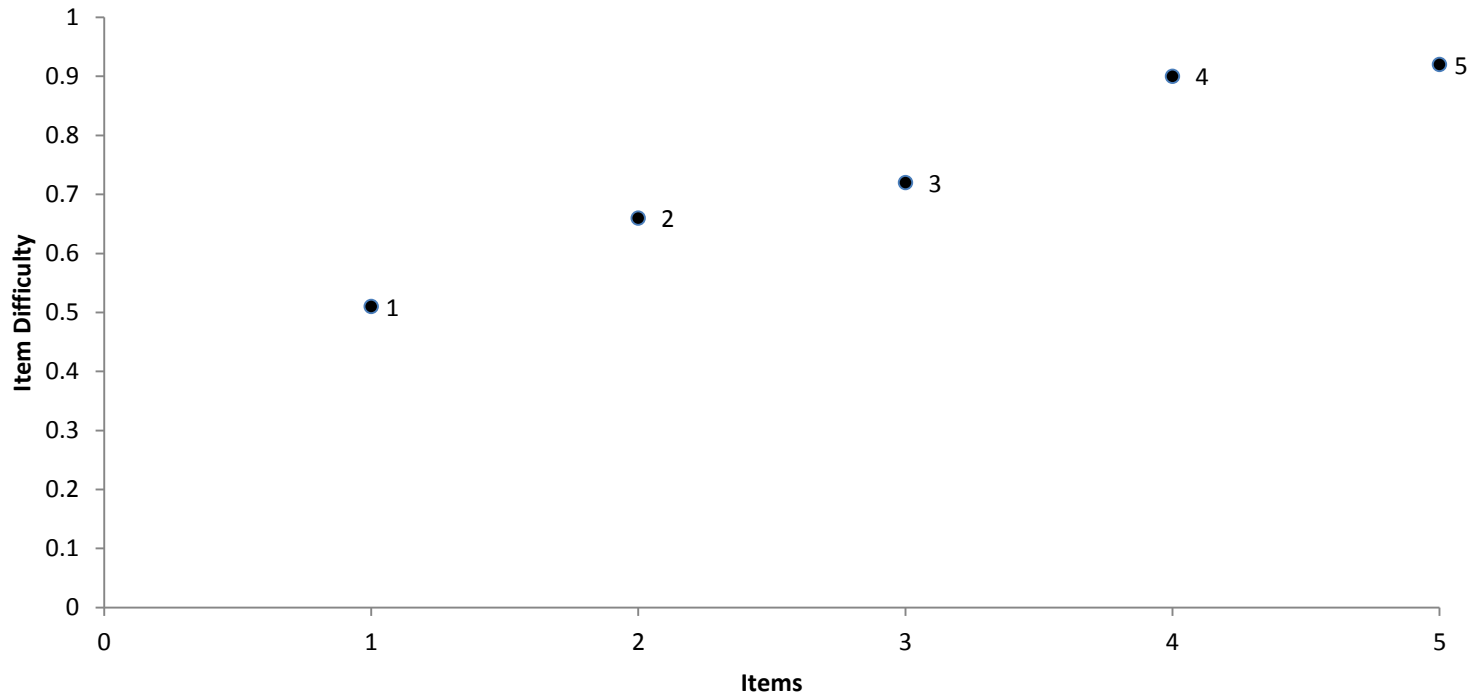
Note. ACE-R=Addenbrooke's Cognitive Examination-Revised. SDRIR=Scottish Dementia Clinical Research Interest Register. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*. *Item taken from MHM Mokken scale

7.4 Validation of the Short ACE-R

The two proposed versions of the Short ACE-R derived from the results of Mokken scaling analysis of the SDRIR data ($N=808$) in Chapter 6 were validated using Mokken scaling analyses in an independent sample, Sydney data ($N=350$) which were originally used in Chapter 4 and also in Chapter 5 in the validation of the Mini-ACE. This sample in this Chapter will be referred to as the Short ACE-R validation sample.

The mean item scores for the Short ACE-R validation sample ($N=350$) for both Short ACE-R scales are presented in the order of ACE-R administration in Table 7.5. The mean item scores for both scales are quite different from the mean scores from the SDRIR Short ACE-R development sample (see Tables 7.3 and 7.4). This disparity reflects the difference in dementia groups and severity between these two samples.

Figure 7.4 Range of item *difficulty* of Short ACE-R item selection 2 for the Short ACE-R development sample



Note. The y-axis represents the mean item scores reflecting item *difficulty* (with higher mean values indicating lower *difficulty*). The x-axis represents the number of items in the scale. 1=Recognition (0.51); 2=Name and address learning (0.66);3= Semantic comprehension (0.72);4= Orientation in geography (0.90);5=Identify fragmented letters (0.92). Short ACE-R=short Addenbrooke’s Cognitive Examination-Revised.

Table 7.5 Mean equated ACE-R scores from the Short ACE-R validation sample ($N=350$) for both SDRIR derived Short ACE-R scales

Short ACE-R selection 1		Short ACE-R selection 2	
Item	Mean	Item	Mean
Orientation in geography	0.78	Orientation in geography	0.78
Naming 1	0.83	Name and address learning	0.72
Semantic comprehension	0.66	Semantic comprehension	0.66
Identify fragmented letters	0.95	Identify fragmented letters	0.95
Recognition	0.70	Recognition	0.70

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised. SDRIR=Scottish Dementia Clinical Research Interest Register. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*.

(i) Short ACE-R item selection 1

Exploratory Mokken analysis was performed on the five scale items. The automated item selection procedure (AISP) allocated all items to the same scale. The item-pair and item scalability coefficients were nonnegative and greater than the lower bound 0.3 respectively. Examining the standard error of the scalability coefficients reveals a considerable degree of uncertainty with the scalability of 'identify fragmented letters' ($H_i=0.38$, $SE=0.07$).

There were no exclusions in the assessment of monotonicity. The assessment of IIO prompted the removal of 'identify fragmented letters'. Additional exploration of scalability coefficients confirmed the item's poor *discrimination* ($H_i=0.19$, $SE=0.06$). The four remaining items formed a reliable ($MS. = 0.76$), moderate hierarchical scale ($H=0.49$) with IIO ($H^T=0.66$) (see Table 7.6).

Table 7.6 Short ACE-R item selection 1: items listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
Semantic comprehension	0.66	Orientation in geography	0.50	0.04
Naming 1	0.83	Semantic comprehension	0.50	0.04
Recognition	0.70	Recognition	0.48	0.04
Orientation in geography	0.78	Naming 1	0.48	0.04

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination* SE=standard error.

(i) Short ACE-R item selection 2

Exploratory Mokken analysis was carried out on the second Short ACE-R selection. The AISP partitioned all items into one scale. The item-pair and item scalability coefficients all met the necessary requirement for inclusion in a Mokken scale; H_{ij} were nonnegative and $H_i > 0.3$. Again, the high standard error of 'identify fragmented letters' ($H_i=0.37$, SE=0.07) introduced some concern about the scalability of this item.

There were no exclusions due to violations of monotonicity. The assessment of IIO prompted the exclusion of 'identify fragmented letters'. The remaining items formed a reliable (MS = 0.77), moderate hierarchical scale ($H=0.49$, SE=0.03) with IIO ($H^T=0.50$). This four item subscale is presented in Table 7.7 ordered by *difficulty* and *discrimination*.

Table 7.7 Short ACE-R item selection 2 items listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i	SE
Semantic comprehension	0.66	Recognition	0.54	0.03
Recognition	0.70	Orientation in geography	0.51	0.04
Name and address learning	0.72	Name and address learning	0.49	0.04
Orientation in geography	0.95	Semantic comprehension	0.42	0.04

Note. Short ACE-R=Short Addenbrooke's Cognitive Examination-Revised. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*. H_i =item scalability coefficient, higher values reflecting greater item *discrimination*. SE=standard error.

7.5 Comparison of Short ACE-R and Mini-ACE

As Short ACE-R item selection 1 does not make clinical sense due to the inclusion of 'recognition' in the absence of 'name and address learning' this scale was disregarded. The comparison of scales will focus on the Mini-ACE and the Short ACE-R item selection 2. This Short ACE-R as a brief quantitative measure includes five items and has a total score of 27; 'name and address learning': subject given a name and address to repeat three times and told they will be asked about it again later, the third trial of repeating "Harry Barnes, 73 Orchard Close, Kingsbridge, Devon" is scored (maximum score =7); 'orientation in geography; "which building, floor, town, county and country are we in?" (maximum score = 5) ; 'semantic comprehension': from an array of 12 drawings the subject is asked to identify the one which is associated with the monarchy, the one which is a marsupial, the one which is found in the Antarctic and the one which has a nautical connection (maximum score = 4) ; 'identify fragmented letters': the subject is asked to identify four degraded letters: K, M, A and T (maximum score = 4) ; 'recognition': subject is given hints to the name and address learned earlier, each recognised item scores one point: "was the name Jerry Barnes, Harry Barnes or Harry Bradford?", "was the number 37, 73 or 76?", "was the street Orchard Place, Oak Close or Orchard Close?", "was the town Oakhampton, Kingsbridge or Dartington?" and "was the county Devon, Dorset or Somerset?" (maximum score = 7) .

The validation of the Mini-ACE in Chapter 5 confirmed that all five items selected ('orientation in time', 'name and address learning', 'verbal fluency-animals', 'draw a clock' and 'name and address learning') formed a formal hierarchy with IIO. The current analyses showed that the items of the Short ACE-R developed in this Chapter did not perform as well with only four of the five items retained in a formal hierarchy with IIO. However it is important to consider whether excluding an item purely for this reason is appropriate. For

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

example, in this case excluding the item would remove assessment of the visuospatial domain. It is important to take the content and relevance of items excluded due to IIO violations into account as they may offer a unique contribution to the scale. Therefore in this case the contribution of all five items selected for inclusion in the brief scales will be examined and discussed.

Item means scores of the Mini-ACE and the Short ACE-R for both the Short ACE-R development sample and Short ACE-R validation sample are presented in Table 7.8. This table demonstrates how the scales perform differently in the two samples. Figures 7.5 and Figure 7.6 illustrate the differences between the two scales in terms of their range of *difficulty* in the development and validation samples respectively; in both samples the Mini-ACE focusses on the assessment of the more *difficult* end of the spectrum whereas the focus of the Short ACE-R is towards the less *difficult* range of assessment.

While psychometrically, a spread across the whole range of item *difficulty* would be useful, from a practical point of view the ideal coverage of the scale in terms of *difficulty* is dependent on what the purpose of the scale is. A scale designed to alert clinicians to possible disease very early in the disease onset would require items of very high *difficulty*. Ideally these high *difficulty* items would also be highly *discriminatory*. A scale with such items would be capable of measuring the cognitive abilities that are lost early and quickly at this stage of impairment. A scale permitting test completion even in very severe stages of disease would require items of very low *difficulty*. A scale designed to be able to monitor change in performance where incremental scores are helpful would need highly *discriminatory* items, which would be capable of assessing small degrees of change. This is less important in a scale where the aim is to determine whether any errors are made or not. These different objectives will have a consequence on the selection of the most appropriate items for a scale.

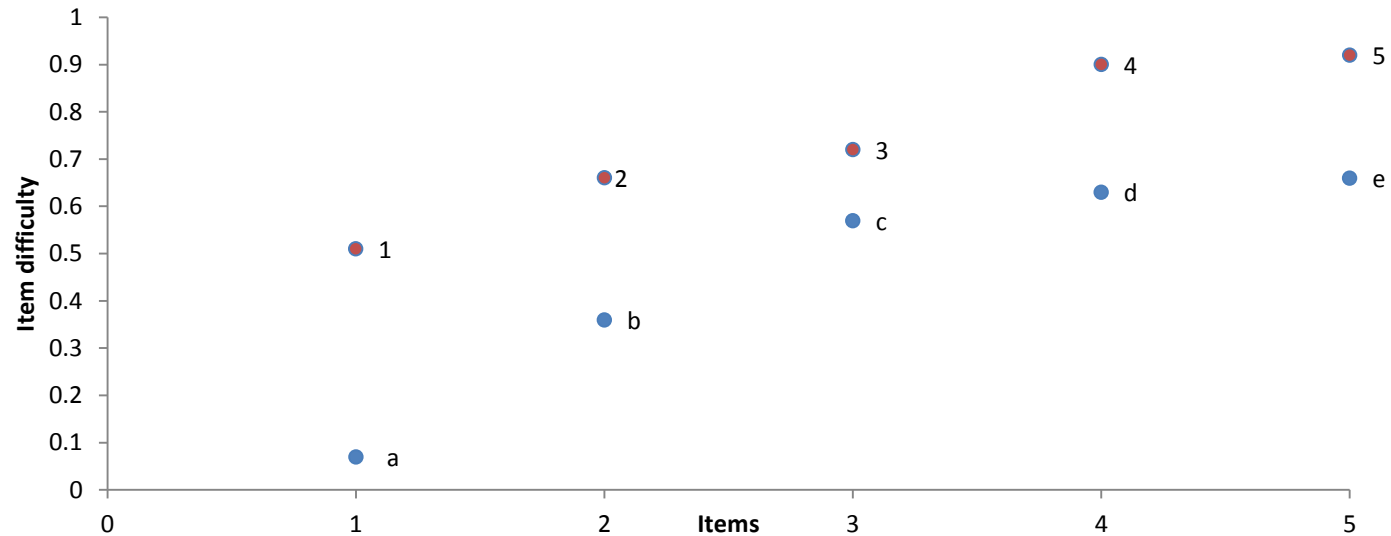
CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

Table 7.8 Mean item scores for the Short ACE-R development sample ($N=808$) and Short ACE-R validation sample ($N=350$) for two shortened versions of the ACE derived using Mokken scaling methods in order of item *difficulty* from most to least *difficult*

Short ACE-R				Mini-ACE			
Short ACE-R Development sample ($N=808$)		Short ACE-R Validation sample ($N=350$)		Short ACE-R Development sample ($N=808$)		Short ACE-R Validation sample ($N=350$)	
Item	Mean	Item	Mean	Item	Mean	Item	Mean
Recognition	0.51	Semantic comprehension	0.66	Name and address recall	0.07	Name and address recall	0.30
Name and address learning	0.66	Recognition	0.70	Verbal fluency-animal	0.36	Verbal fluency-animal	0.31
Semantic comprehension	0.72	Orientation in geography	0.78	Orientation in time	0.57	Draw a clock	0.72
Orientation in geography	0.90	Name and address learning	0.72	Draw a clock	0.63	Name and address learning	0.72
Identify fragmented letters	0.92	Identify fragmented letters	0.95	Name and address learning	0.66	Orientation in time	0.76

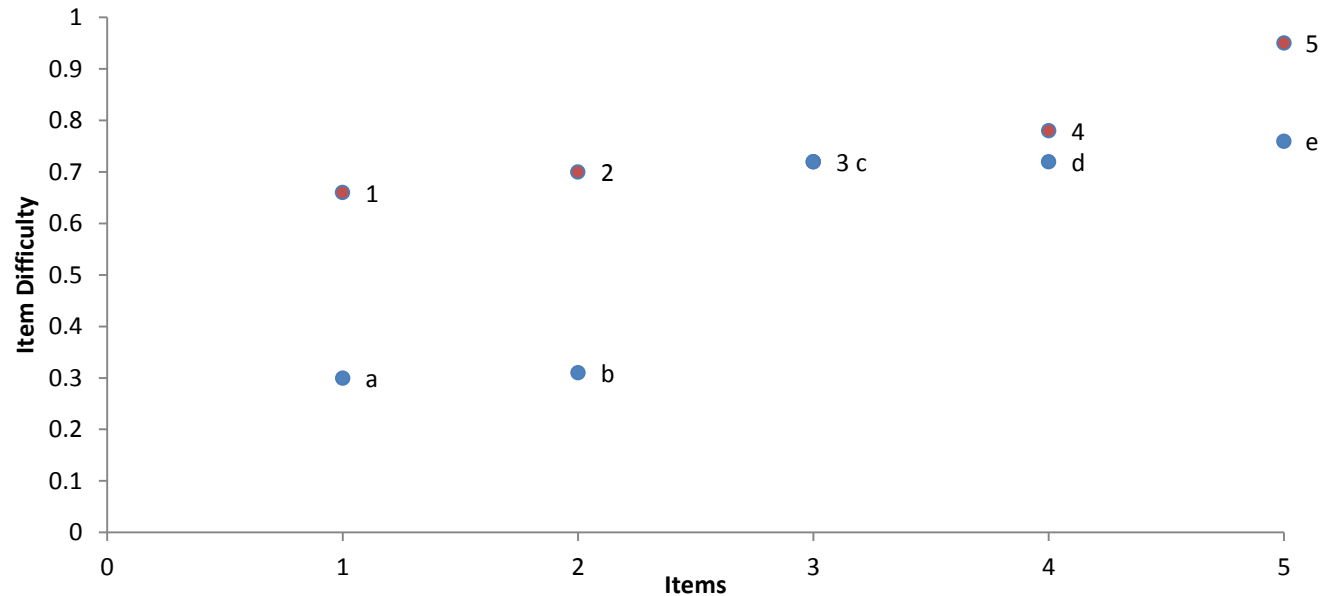
Note. Short ACE-R=Short Addenbrooke’s Cognitive Examination-Revised. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*.

Figure 7.5 Comparison of the range of item *difficulty* of the Short ACE-R and Mini-ACE for the Short-ACE-R development sample (N=808)



Note. The y-axis represents the mean item scores reflecting item *difficulty* (with higher mean values indicating lower *difficulty*). The x-axis represents the number of items in the scale. Red dots=Short ACE-R items: 1=Recognition (0.51); 2=Name and address learning (0.66); 3=Semantic comprehension (0.72); 4=Orientation in Geography (0.90); 5=Identify fragmented letters (0.92). Blue dots=Mini-ACE items: a=Name and address recall (0.07); b=Verbal fluency-animal (0.36); c=Orientation in time (0.57); d=Draw a clock (0.63); e=Name and address learning (0.66). Short ACE-R=Short Addenbrooke's Cognitive Examination-Revised. Mini-ACE=Mini-Addenbrooke's Cognitive Examination

Figure 7.6 Comparison of the range of item *difficulty* of the Short ACE-R and Mini-ACE for the Short ACE-R validation sample (N=350).



Note. The y-axis represents the mean item scores reflecting item *difficulty* (with higher mean values indicating lower *difficulty*). The x-axis represents the number of items in the scale. Red dots=Short ACE-R item: 1=Semantic comprehension (0.66); 2=Recognition (0.70); 3=Name and address learning (0.72); 4=Orientation in geography (0.78); 5=Identify fragmented letters (0.95) Blue dots=Mini-ACE items: a=Name and address recall (0.30); b=Verbal fluency-animal (0.31); c=Draw a clock (0.72); d=Name and address learning (0.72), e=Orientation in time (0.76). Short ACE-R=Short Addenbrooke's Cognitive Examination-Revised. Mini-ACE=Mini-Addenbrooke's Cognitive Examination

7.5.1 Assessment of mild impairment

The Mini-ACE is the superior screening scale for milder cognitive impairment. While the most *difficult* item in the Short ACE-R ('recognition') has a mean score of 0.51 (development sample) and 0.70 (validation sample) the Mini-ACE extends the range of assessment at the lower levels of impairment with the inclusion of 'name and address recall'. 'Name and address recall' has a mean score of 0.07 and 0.30 for the development and validation samples respectively. The memory tested in the Short ACE-R is recognition as opposed to recall which has a higher mean score. The inclusion of an item assessing memory recall is important due to the significance of episodic memory impairment in the early identification of AD (Sperling et al., 2011).

Comparatively the assessment of mild impairment using the Short ACE-R is rather limited due to the lack of high *difficulty* items. The mean values for the most *difficult* items of the Short ACE-R for the development and validation samples are 0.51 and 0.66 respectively. Figures 7.4 and 7.5 visually present the differences between the two scales in terms of the levels of *difficulty* assessed in both samples. It can be seen that the Mini-ACE would be the more appropriate choice of scales to detect early cognitive decline due to the inclusion of high *difficulty* items.

7.5.2 Assessment of severe impairment

The Mini-ACE however has a more limited range in the assessment of severe levels of impairment (see Figures 7.4 and 7.5). The least *difficult* items in the Mini-ACE have mean scores for the development and validation sample of 0.66 ('name and address learning') and 0.76 ('orientation in time') respectively. In comparison the Short ACE-R is well equipped in the measurement of the severe ranges with a mean score for the least *difficult* item ('identify fragmented letters') in the development and validation samples of 0.92 and 0.95 respectively.

Therefore while this level of impairment is restricted using the Mini-ACE, the Short ACE-R's inclusion of a very low *difficulty* item ('identify fragmented letters') means the Short ACE-R adequately assesses this level of impairment. Figures 7.4 and 7.5 displays this extended level of coverage offered by the Short ACE-R by the tendency of the items in this scale to gravitate towards to upper ranges of the graph reflecting the high mean scores (i.e. low *difficulty*) of these items.

7.5.3 Importance of samples used

The two samples used to develop the Short ACE-R and Mini-ACE differ considerably from each other. The most significant of these differences is the dementia diagnoses, severity and age of the participants of each sample. The Short ACE-R development sample is more representative of the general dementia population with a preponderance of patients with AD (58 %) whereas the sample from which the Mini-ACE was derived is less typical in terms of its diagnostic make up with a majority of primary progressive aphasia (42 %) and a range of less common dementia syndromes such as corticobasal syndrome and behavioural variant frontotemporal dementia. The mean age for the Short ACE-R development sample (77.5 years) is more typical of a dementia register whereas the mean age of the Mini-ACE development sample (65.4 years) and the validation sample (65.4 years) is more characteristic of an earlier onset dementia sample. It is not clear whether the results of the analysis of the validation sample would apply equally to an older patient sample. The selection of items based on results of Mokken scaling analysis of these two different samples will be influenced by these sample characteristics and differences. The selection of the Mini-ACE was made from a larger pool of IIO items (17 vs. 11 for the Short ACE-R) which increased the range of items available. The 11 items of the ACE-R conforming to an IIO hierarchy, which formed the basis for item selection, had a restricted range at the higher levels of ability (i.e. no item with a mean score <0.50). The exclusion of these high *difficulty* items due to violations of IIO

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

could have arisen due to the greater number of diagnostic groups in the Short ACE-R development sample. This sample comprised seven different clinical groups whereas the sample used to derive the Mini-ACE was made up of four groups. This greater heterogeneity is likely to have been the source of IIO violations which saw a greater proportion of items excluded (12 items were excluded from the Short ACE-R development sample in comparison to only two items in the Mini ACE development analysis). However, it is also worth considering whether the heterogeneity of the sample from which the Short ACE-R was derived may make the scale more generalizable to other clinical groups as opposed to the Mini-ACE which was developed using more homogenous data. This is supported by the difference in mean item scores between the development and validation samples for each scale. The scores for the Short ACE-R differ less between the development and validation samples than those of the Mini-ACE (see Table 7.8).

The samples used within this Chapter to develop and validate the Short ACE-R also differ in terms of diagnoses and disease severity. The development sample is more impaired than the higher functioning validation sample. This is reflected by the mean scores of the new proposed scales for both of these samples; while the mean Short ACE-R scores are relatively similar: 19.2/27 and 20.3/27, for the development and validation samples respectively, there is a considerable difference in Mini-ACE performance: 13.6/31 and 16.7/31 for the development and validation samples respectively. The greater difference in Mini-ACE performance is largely driven by the inclusion of 'name and address recall', which has a significantly lower mean equated score for the SDRIR sample (0.07) than the Sydney sample (0.30).

These differences will have affected the performance of the two scales. The item ordering of both scales differs in the two samples. The item ordering of the Short ACE-R in the development sample was different with a lower mean score for the most *difficult* item

(‘recognition’ = 0.51) than for the validation sample (0.70). This is most likely due to the greater number of patients with AD in the development sample where memory problems are likely. That ‘semantic comprehension’ is a more *difficult* item in the validation sample (0.66 vs. 0.72) is unsurprising due to the preponderance of primary progressive aphasia where semantic impairments are common (Mesulam, 2001) in this sample.

‘Orientation in geography’ was more *difficult* for the validation sample (0.78) than for the development sample (0.90). As patients with AD are known to experience disorientation at an early stage of the disease (Morris, 1999) this is unexpected as the validation sample has a smaller proportion of patients diagnosed with AD (25%) compared to the development sample (58 %). However looking at the mean score for ‘orientation in time’ from the Mini-ACE the patients in the development sample perform worse than those in the validation sample which is the expected pattern given the diagnostic make up of these samples.

Orientation may serve as a potentially valuable diagnostic difference between patients with AD and FTD (Yew, Alladi, Shailaja, Hodges & Hornberger, 2013). With regards to item selection for a brief scale ‘orientation in time’ may be a more useful measure of orientation as both AD and FTD patients perform worse in measures of temporal orientation than spatial orientation (Yew et al., 2013) and from the current results this item is more *difficult* for a sample with a majority of AD patients. The Mini-ACE’s inclusion of ‘orientation in time’ may therefore offer greater opportunity for differentiation between AD and FTD patients.

7.5.4 Practical implications

The breadth of measurement of the Short ACE-R reflects the range of the items in the ACE-R IIO hierarchy revealed in Chapter 6, which formed the basis for item selection for the Short ACE-R. This hierarchy comprises items from the less *difficult* range of the breadth of measurement of the ACE-R. The restricted range of items available for selection restricts the

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

applicability of these shortened scales (Short ACE-R item selection 1 and 2) to the assessment of more severe stages of dementia which limits its clinical applicability as a brief screening tool where early cognitive impairment would hope to be detected. For this reason using items from the ACE-R IIO hierarchy from Chapter 6 to derive a shortened scale is not recommended for a brief screening tool. See Table 7.9 where items are presented in order of test administration with mean item scores for both short scales in the validation sample.

The Short ACE-R may have better applicability as a measure of moderate to severe levels of impairment due to its bias towards the low *difficulty* items. However while the inclusion of an item of very low *difficulty* item—‘identify fragmented letters’—increases the Short ACE-R’s range of assessment to the more severe levels of dementia examining the mean scores (0.92-0.95) demonstrates that as a very low *difficulty* item almost every participant in both samples scored highly on this item which raises the question of the value of this item. In a scale with so few items having one item where almost all patients score well is not very helpful. It would be useful to examine this item in a sample with more severe dementia than the validation sample to establish if there were a range of responses and therefore whether this item had high levels of *discrimination* as well as low *difficulty*.

The Mini-ACE, on the other hand, has the most appropriate application as a ‘screen’ in higher functioning samples, as it is more sensitive to early cognitive impairment. The mean scores for both samples demonstrate this where the mean item scores range from 0.07-0.66 and 0.30 to 0.76 for the development and validation samples respectively. The Mini-ACE performs quite well in the Scottish development sample but the Short ACE-R does not perform as well in the higher functioning Australian validation sample.

The use of the Mini-ACE is preferable for several reasons: (i) it performed better in the additional Mokken scaling analysis performed in Chapter 5 than the Short ACE-R

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

analysis using the same sample in this Chapter; (ii) all items of the Mini-ACE were selected from a hierarchical subscale of ACE-III items whereas in the development of the Short ACE-R ‘name and address learning’ was selected from the subset of items meeting the less restrictive MHM; (iii) the Mini-ACE includes an item from each of the five cognitive ACE domains whereas the Short ACE-R does not include a measure of verbal fluency and (iv) the inclusion of a high *difficulty* item (‘name and address recall’) in the Mini-ACE extends the range of measurement to a higher level of cognitive functioning providing the opportunity to detect early cognitive decline which is important in any cognitive screen for dementia (see Table 7.9).

Table 7.9 Mean item scores for the Short ACE-R validation sample (N=350) for the Short ACE-R and Mini-ACE. Items are presented in test order

Short ACE-R			Mini-ACE		
Domain	Item	Mean	Domain	Item	Mean
Memory	Name and address learning	0.72	Memory	Name and address learning	0.72
Attention	Orientation in geography	0.78	Attention	Orientation in time	0.76
Language	Semantic comprehension	0.66	Fluency	Verbal fluency-animal	0.31
Visuospatial	Identify fragmented letters	0.95	Visuospatial	Draw a clock	0.72
Memory	Recognition	0.70	Memory	Name and address recall	0.30

Note. Short ACE-R= Short Addenbrooke’s Cognitive Examination-Revised. Mini-ACE=Mini Addenbrooke’s Cognitive Examination. Mean item scores (range: 0-1) reflect item *difficulty* with higher scores indicating lower *difficulty*

7.6 Discussion

This Chapter describes the development and analysis of a new proposed scale, the Short ACE-R. This analysis was performed to determine whether using the same methods as the Mini-ACE development analysis in Chapter 5 in a larger sample would produce similar results and if not to explore the reasons for any differences. The analyses in this Chapter did not identify the same items for inclusion in the new brief scale. Therefore the performance of both the Short ACE-R and the Mini-ACE were compared in two different samples to determine which had the superior application clinically as a brief screening tool.

While these analyses introduce a new brief scale, the Short ACE-R, the clinical applications of this new scale are limited. The psychometric properties of the Short ACE-R preclude its application as an effective screening test. However these analyses demonstrate the value of Mokken scaling analysis to examine the item properties of existing scales and its application to scale development, and in addition the essential role of clinical expertise to ensure a scale meets its required function.

Mokken scaling analyses of both Short ACE-R scales failed to retain an item assessing visuospatial skills in a formal IIO hierarchy. That the five items of the Short ACE-R do not form an IIO hierarchy means the score of individual items cannot be used to quickly gauge a patient's cognitive status (i.e. responses to individual items cannot be used to provide information about a respondent's likely level of cognitive impairment as would be the case had the five items conformed to an IIO hierarchy). However as the scale was designed to be very brief there is little need for this psychometric property. The failure of 'identify fragmented letters' to conform to a hierarchical Mokken scale in the additional analysis could be attributed to the poor *discrimination* of this item reflecting its poor relation to cognitive impairment in dementia. The heterogeneity of the validation sample could be contributing to

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

the low *discriminatory* power of this item. Previous results where visuospatial items also showed poor *discrimination* (Chapter 4) suggests that in this sample these items do not measure cognitive impairment as effectively as other ACE-R items possibly through variability of response introduced by the additional dependence on vision and motor skills for item performance. The inclusion of patients diagnosed with frontotemporal dementia with motor neurone disease in the validation sample where motor symptoms could impair the ability of this group to draw the pentagons, which could be responsible for driving the poorer *discrimination* in the present results. In the analyses of SDRIR data in Chapter 6 the visuospatial items performed much better with ‘identify fragmented letters’ and ‘draw intersecting pentagons’ retained in three and all four of the IIO hierarchies respectively. While perhaps this implies that ‘drawing intersecting pentagons’ may have been the more appropriate choice of visuospatial item for the Short ACE-R the results of the Mokken scaling analysis used to select the items did not suggest this as in this sample ‘identify fragmented letters’ had a higher *discriminatory* value and its inclusion does not require additional practical considerations.

Further analysis of the Short ACE-R is required to determine the scale’s ability to provide specific cognitive profiles for different types of dementia. The Mini-ACE is sensitive to decline in memory in early AD and has been found to demonstrate somewhat distinctive diagnostic profiles across AD, FTD and corticobasal syndromes (Hsieh et al., 2015). It is doubtful that the Short ACE-R would perform as well in the detection of early AD due to the lesser *difficulty* of the items within it. The absence of a verbal fluency item in the Short ACE-R may also restrict the ability of the scale to reveal executive function impairments and to differentiate between different dementia types, for example patients with progressive aphasia retrieved significantly fewer words on the animal fluency item than all other dementia groups ($p < 0.01$) (Hsieh et al., 2015). The Mini-ACE’s inclusion of verbal fluency

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

item will add to the scale's ability to discriminate between these types of dementia. The application of the Mini-ACE in further differentiation of progressive aphasia subtypes (non-fluent, Logopenic and semantic variant) could be explored further using larger samples of these patient groups. While a brief screening test such as the Mini-ACE does not replace a comprehensive assessment including the thorough and detailed neuropsychological, medical and imaging evaluation necessary for an accurate differential diagnosis of dementia it is helpful that, as in the case of the MMSE, the Mini-ACE also produces different group profiles for different diagnoses.

While the Short ACE-R was derived from analysis of ACE-R data the items selected are also common to the ACE-III. This is the same case for the Mini-ACE where the item selection derived using ACE-III data resulted in the selection of items common to both ACE-III and ACE-R. This is significant for both scales for two reasons. Firstly, scores for both the Mini-ACE and the Short ACE-R can be produced from patients pre-existing data from either the ACE-R or ACE-III. This is helpful as it would allow further validation of these scales and also it enables clinicians to derive the scores of the shorter scales. Secondly, that the items of the Short ACE-R are common to the ACE-III avoids any uncertainty regarding copyright of the MMSE.

There are significant advantages to the examination of item properties; for example, with regards to scale development this level of analysis permits the selection of items of high *discrimination* and at specific levels of *difficulty* to form a new hierarchical scale. Using an invariantly ordered hierarchical scale from which to select items for a new briefer scale (as in the case with the development of the Mini-ACE and for four of the five items of the Short ACE-R) implies that all items are sufficiently *discriminatory*, and have monotonic and nonintersecting item response functions. However the choice of items from a hierarchical scale ultimately has a considerable subjective element. Between items of a similar level of

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

discrimination and *difficulty* the item selection is determined by the aim of the scale or motivation of the scale developer. For example, the choice of the visuospatial item for the Short ACE-R from the two visuospatial items of the IIO hierarchy; ‘draw intersecting pentagons’ and ‘identify fragmented letters’ was ultimately made based on the higher level of *discrimination* of ‘identify fragmented letters’. Additional practical considerations were taken into account in this choice as for a brief screening tool where brevity is the goal, the inclusion of ‘draw intersecting pentagons’ would add to the test administration time. This illustrates that while there are subjective and practical components to the item selection the examination of item parameters provided by Mokken scaling analyses such as *discrimination*, can be valuable in scale development by guiding item selection.

Mokken scaling can also be used to analyse larger measures to identify specific profiles of deterioration to help to accurately diagnosis and differentiate patients. This knowledge can then be used to develop disease specific subscales. There is a large involvement of language and verbally based items in both the Short ACE-R and the Mini-ACE which may increase the tests sensitivity to semantic dementia as these patients tend to perform worse on language-based measures (Libon et al., 2007). Mokken scaling methods could be used to develop more disease specific subscales with the identification of cognitive profiles for different dementia types. For example, a subscale with visuospatial items could be useful for assessing patients with corticobasal syndromes (Tang-Wai et al., 2003), or episodic memory items for AD (Libon et al., 2007).

Some limitations of these analyses should be considered. The better scalability of the Mini-ACE in comparison to the Short ACE-R according to the additional Mokken scaling analyses could be caused by the samples in which the analyses were performed. While both scales were subsequently analysed using the same sample the samples each were developed from differed from this sample to different degrees. The Mini-ACE validation analysis was

CHAPTER 7: DEVELOPMENT OF THE SHORT ACE-R

performed in a sample which is very similar to the one from which it was developed. The Mini-ACE was derived using data from Sydney, Oxford and Cambridge ($N=117$) and was assessed by a larger sample of patients from one of the same research groups in Sydney ($N=350$). The Short ACE-R was derived using data from a large Scottish case register (SDRIR) ($N=808$) and then assessed using data from a smaller sample from Sydney ($N=350$). While further analysis would be required to confirm, it is possible that the differences between development and validation sample could have contributed to the poorer results in the analysis of the Short ACE-R in comparison to the validation analysis of the Mini-ACE in Chapter 5 which was performed on a more homogeneous sample. The similar diagnoses and patients represented in both the Mini-ACE development and validation samples is likely to result in less heterogeneity which can cause violations of IIO.

The brevity of the Short ACE-R reduces the timespan between learning the name and address and subsequent recognition relative to the full ACE-R. However the reduced time delay between learning and recognition is unlikely to have much of an effect on recognition scores. Particularly for patients with AD, where retention of novel information is limited to very short time spans, the remaining test items should cause sufficient interference (Benson, Slavin, Tran, Petrella & Doraiswamy, 2005; Fillenbaum et al., 1994). The time span between ‘name and address learning’ and ‘recognition’ would be slightly shorter in the Short ACE-R as opposed to the Mini-ACE due to the inclusion of ‘verbal fluency-animal’ which offers the patient one minute to respond making the administration of the Mini-ACE longer than the Short ACE-R.

A limitation of the Mini-ACE is that it was developed using data from patients from specialised or tertiary hospital clinics. The validation sample in this Chapter reflects the specialist nature of these research clinics with a preponderance of less common types of dementia. As the Short ACE-R development sample is more representative of the general

population with a greater proportion of AD patients and those with mixed aetiologies the results from this analysis in a larger sample can be considered more representative of the general population. To investigate this further the item properties and hierarchical structure of the ACE-III, Mini ACE and Short ACE-R should be assessed in a larger, more representative group of patients with dementia. This way both proposed brief scales can be assessed and it can also be determined whether the same items emerge as candidate items or whether an alternative item selection is suggested.

7.7 Conclusion

This Chapter endeavoured to derive a new brief screen for dementia from the ACE-R.

However the item properties and subsequent Mokken scaling analyses indicated that the scale was of limited use clinically as a screening tool. The use of the Mini-ACE is instead advocated as a brief screening tool in high functioning samples with two recommended cut-offs; $\leq 25/30$ ¹ has high sensitivity and specificity and is at least five times more likely to be the score of a patient with dementia than without and $\leq 21/30$ which is almost certainly diagnostic of dementia (Hsieh et al., 2015). Further trials of the Mini-ACE are required in different patients groups. Both efforts to develop shortened scales from the ACE-III (the development of the Mini-ACE in Chapter 5) and ACE-R (the development of the Short ACE-R in the present chapter) illustrate the value of examining item properties and hierarchical structure of items selected for inclusion in a brief condensed version. However the ultimate selection of items has to consider practical application and the purpose of the test. The new proposed scale must also then be validated in a clinical sample.

¹ Mini-ACE maximum score=30 due to the removal of one of the embedded 'orientation in time items' (orientation to season) which reduces the total score from 31 to 30 (Hsieh et al. 2015).

**Chapter 8: From ‘aisle’ to ‘labile’: a hierarchical NART scale revealed by
Mokken scaling**

Work presented in the following chapter is taken from the following paper:

McGrory, S., Austin, E.A., Shenkin, S.D., Starr, J.M and Deary, I.J. (*in press*). From ‘aisle’ to ‘labile’: a hierarchical NART scale revealed by Mokken scaling, *Psychological Assessment*

8.1 Introduction

Determining the degree of cognitive decline caused by dementia or a normal ageing process relies on establishing a valid estimate of prior ability level (Crawford, 1992). There are substantial individual differences in cognitive ability; therefore it is important to take a person’s prior/premorbid cognitive ability level into account to establish whether there has been a decline. Preferably, this would involve a comparison of current cognitive ability with an actual measure of prior cognitive ability. However, actual premorbid measures of ability are seldom available in clinical situations. This results in the dependence upon estimates of premorbid cognitive function.

A commonly used test for estimating peak premorbid cognitive ability is the National Adult Reading Test (NART) (Nelson, 1982; Nelson & Willison, 1991). This test examines pronunciation of 50 irregular English words of graded difficulty that violate the typical grapheme-phoneme and stress rules (e.g. *gauche*, *thyme*), i.e. guessing will not provide the correct pronunciation. The shortness of the words ensures that minimal demands are placed on the patient’s current mental capacity (Nelson & O’Connell, 1978). Therefore, successful

word reading is thought to depend on premorbid ability and not on current cognitive ability. The NART has been validated as an estimator of premorbid mental ability in mild to moderate dementia (Bright, Jaldow & Kopelman, 2002; Crawford, Parker & Besson, 1988; Sharpe & O'Carroll, 1991; McGurn et al., 2004) and also in normal cognitive ageing (Dykiert & Deary, 2013). After controlling for age 11 IQ, mean NART scores do not differ between those with and without mild-to-moderate dementia (McGurn et al., 2004).

The NART comprises words of graded difficulty starting with more commonly-used words, such as 'ache' and 'chord' and becoming more difficult as it progresses to less frequently-used words, such as 'syncope' and 'campanile'. While NART items may be considered as forming an informal hierarchy, as planned by the test's constructors, it is important to investigate item properties explicitly to determine whether the items conform to a formal hierarchy of *difficulty* and whether this hierarchy is the same for all respondents (i.e. is the ordering for people with higher levels of ability the same as for those with lower ability levels). The effect of ability level on item ordering was investigated by Deary, Watson, Booth and Gale (2013) who determined that the strength of hierarchies of item ordering of the Warwick-Edinburgh Mental Well-being Scale varied according to the cognitive ability of the sample. Establishing whether the NART items form an IIO hierarchy would simplify test administration and interpretation of responses. From a clinical perspective, hierarchical tests are attractive for their ease of use and scoring (Kempen, Myers & Powell, 1995). Confirming a hierarchy of NART item *difficulty* has meaningful clinical implications; continuing to test patients on words that they are predictably going to be unable to pronounce correctly may cause undue distress without adding any valuable information. Also, responses to individual items and not just total scores can provide insight into a respondent's level of ability based on the item's location in the hierarchy (Watson, Deary, & Austin, 2007). Hierarchical tests have proven valuable in the assessment of several constructs, for example, psychological distress

CHAPTER 8: DEVELOPMENT OF HIERARCHICAL MINI-NART

(Watson, Deary & Shipley, 2008), feeding difficulty in dementia (Watson, 1996) and activities of daily living (Fieo, Watson, Deary & Starr, 2009; Kempen & Suurmeijer, 1990).

The degree to which NART items form a hierarchy can be determined using Mokken scaling analysis. Mokken scaling analysis can be applied to examine clinically valuable properties of items within scales, including item *discrimination*. Considering item *discrimination* allows for the creation of scales with greater precision without having to increase the number of items. For example, Sabourin, Valois and Lussier (2005) used IRT methods to create a four item abbreviated form of the Dyadic Adjustment Scale, which was as effective as the original 32 item scale. Similarly a 10-item scale was derived from the 19-item Feelings Scale without the loss of measurement precision (Edelen & Reeve, 2007).

IRT methods have been applied to two measures of premorbid intelligence; a French language version of the NART; the *f*NART (Mackinnon, Ritchie & Mulligan, 1999) and the Adult Reading Test (ART) (Letz et al., 2003). Mackinnon et al. (1999) used a two-parameter logistic IRT model to examine the measurement properties of the 40-item *f*NART. The *discrimination* of the scale items varied considerably with several of the items contributing little to the assessment of premorbid intelligence. A refined 33-item *f*NART was revealed with the elimination of seven items with poor *discriminatory* power.

Letz et al. (2003) fit a one-parameter logistic (Rasch) model to the items of the Adult Reading Test (ART), adapted from the North American Adult Reading Test (NAART, Blair & Spreen, 1989). Rasch analysis provided an improved ordering of *difficulty* from the original subjective ranking, finding 'two' to be one of the least *difficult* items and 'demesne' to be the most *difficult* item. Results from this Rasch analysis formed the basis for the implementation of a computerised-adaptive ART whereby items are matched to respondents by *difficulty*. This prevents individuals being presented with items far beyond their ability

CHAPTER 8: DEVELOPMENT OF HIERARCHICAL MINI-NART

level helping to reduce frustration or anxiety and minimising the boredom or carelessness of those with higher ability when faced with very easy items.

The possibility of deriving a briefer scale from the NART from which to estimate premorbid IQ is not new. Beardsall and Brayne (1990) explored the idea of creating a shortened version of the NART. A regression equation was developed based on scores from the first 25 words of the NART to predict scores on the remaining 25 words (i.e. items 26 to 50). This method provided a reasonably accurate estimation of the full NART score with predicted NART and true NART scores correlating strongly ($r = .93, p < .001$). While the application of the Short NART left a proportion (23-31%) of the variance unaccounted for, the accuracy with which the Short NART predicted WAIS IQ was effectively equal to that of the full NART (Crawford, Parker, Allan, Jack & Morrison, 1991). The authors suggest the application of the Short NART with reasonable confidence where helpful or convenient in place of the full scale.

While these studies have analysed and refined the assessment of premorbid cognitive ability to our knowledge there has been no application of Mokken scale analysis to the NART. Therefore the aim of the present study was to examine the item properties and the hierarchical structure of the NART by assessing the fit of the items to Mokken's Monotone Homogeneity Model (MHM) and the non-intersection of item response functions (IRFs). Establishing the fit of the data to these models would allow the use of total scale scores (in the case of the MHM) and individual items (IIO) to assess estimated levels of premorbid cognitive ability. Additionally this analysis aims to determine the contribution of each item. Redundant items can be removed to form a new brief scale.

8.2 Method

8.2.1 Participants

The Lothian Birth Cohort 1936 (LBC1936) comprises 1091 community-dwelling older adults most of whom completed the Moray House Test No. 12 (MHT) (Scottish Council for Research in Education (SCRE), 1933) of verbal reasoning at a mean age of 11 as part of the Scottish Mental Survey of 1947 (Scottish Council for Research in Education, 1949; Deary, Whalley & Starr, 2009). The Scottish Mental Survey 1947 (SMS1947) measured the mental ability of almost all Scottish schoolchildren born in 1936 and attending school at age 11 years on June 4th 1947 using the MHT. The MHT is a well-validated measure of general intelligence comprising mostly verbal reasoning items with a maximum possible score of 76. Childhood MHT scores were highly correlated with the Stanford-Binet intelligence test, $r=.81$ in boys ($N=500$) and $r=.78$ in girls ($N=500$; SCRE, 1933). The Lothian Birth Cohort was established to study the determinants of individual differences in cognitive ageing from childhood to old age. Between 2004 and 2007 those residing in Edinburgh and the Lothians who may have taken part in the SMS 1947, who were then approximately age 70, were contacted and invited to participate in the LBC1936. The Community Health Index along with media advertisements was used to identify potential participants born in 1936. From this index 3810 potential participants were identified. Between June 2004 and November 2006 3686 of those identified were contacted. In total 2318 responses were received resulting in the recruitment of 1091 eligible participants to the cohort and were assessed at wave 1. All participants spoke English as their first language. Thorough and detailed demographic data, medical history and physical information data as well as measures of memory, reasoning, executive functioning, and processing speed were collected at each wave of assessment. The key strength of this cohort is the availability of a valid intelligence test score from childhood for participants currently in old age who have retaken the same test of intelligence in addition

CHAPTER 8: DEVELOPMENT OF HIERARCHICAL MINI-NART

to other cognitive and physical measures. The recruitment and testing of this cohort has been described in detail elsewhere (Deary et al., 2007; Deary, Gow, Pattie & Starr, 2012).

Participants in the LBC1936 returned for detailed cognitive and physical testing from age 70 (wave 1, N=1091), and item level responses to the NART were recorded at wave 3 (2012), at a mean age of about 76 years. Age 70 IQ was measured by the MHT (mean= 65.7, SD= 7.7) corrected for age in days at time of testing, and converted to an IQ score (mean IQ=102.42, SD=13.16). Social class was derived from the participants' reported highest occupational level as well as that of participants' fathers. Social class for the participants was calculated using the Office of Population Censuses and Surveys; Classification of Occupations, 1980. Social class of participants' fathers was calculated using the General Register Office's Census, 1951 Classification of Occupations. Both were classified as one of six categories from I (professional) to V (unskilled) with lower numbers designating higher social class. Married women also reported the occupation of their spouses which was used if higher.

Self-reported medical background was obtained for all participants at the cognitive and physical assessment. After excluding those who had a self-reported clinical history of dementia (N=8) data from all other participants returning at wave 3 with complete NART item level data were included for analysis (N=587, 51% male). Mini Mental State Examination (MMSE) (Folstein, Folstein & McHugh, 1975) scores indicated that 99.6% of this sample scored ≥ 23 . The characteristics of study participants are shown in Table 8.1.

Table 8.1 Baseline sample characteristics

	Mean	SD
Age	76.3	0.7
Sex		
Male (%)	51.1	
Female (%)	48.9	
Age 11 IQ	101.5	14.9
Age 70 IQ	102.4	13.2
Age 11 MHT	50.6	11.6
Age 70 MHT	65.7	7.7
MMSE	28.7	1.5
NART	35.3	7.7
Father's SES	2.9	0.9
Participant's SES	2.5	0.9
Education (years)	10.8	1.2

Note. SD=standard deviation. MMSE=Mini-Mental State Examination. NART=National Adult Reading Test. IQ calculated from MHT (=Moray House Test) score corrected for age in days at time of testing and converted to IQ scale. Father's SES (=socio-economic status) is participants' father's social class when the participants were 11 years old.

8.2.2 Measures

The administration of the NART requires respondents to read aloud 50 words which are irregular with regards to their grapheme-phoneme (letter-sound) correspondences (Coltheart et al., 1987). The responses to each of the 50 items of the NART are scored dichotomously; respondents are either able or unable to pronounce the word correctly. Higher scores (fewer errors) indicate higher premorbid cognitive ability. The NART has high internal consistency (0.90; Crawford et al., 1988), high test-retest reliability (0.98; Crawford, Parker, Stewart, Besson & Lacey, 1989) and good inter-rater reliability (0.88; O'Carroll, 1987). The percentage of respondents correctly pronouncing the NART items was used to indicate level of item *difficulty* with lower percentages indicating greater degree of *difficulty*.

8.2.3 Mokken scaling

Exploratory Mokken scaling analysis was applied to investigate whether the ordering of items by *difficulty* is the same for all respondents, making it invariantly ordered. The fit of the items to Mokken scaling properties was assessed by examining whether they conformed to the four assumptions; unidimensionality, local stochastic independence, monotonicity and non-intersection. Mokken scaling analysis was performed using the Mokken package in R (van der Ark, 2007). These assumptions were investigated using a hierarchical clustering algorithm, scalability coefficients, latent monotonicity, and the H^T coefficient.

8.2.4 Graphical analysis

The R package KernSmoothIRT (Mazza, Punzo, & McGuire, 2014) was used to graphically present item properties. The package applies kernel smoothing in the estimation of item response functions and related graphical analysis. It provides several plotting and analytical methods to consider properties of the items, subjects and test as a whole. The exploratory nature of the package makes it ideal to be used alongside Mokken analysis as it provides plots which can be helpful when examining the monotonicity and *discrimination* of items. For more detail on this package see Mazza et al. (2014).

8.2.5 Validation

The present study had access to childhood IQ scores which enabled the retrospective validity of the NART items as proxies for prior cognitive ability across the lifespan to be assessed. The correlation between NART items and prior and concurrent cognitive ability, both measured by converting MHT scores at age 11 and age 70 into IQ scores, was investigated. Regression and correlation analyses were performed using SPSS v. 19.0.

8.3 Results

Descriptive statistics for the sample variables are presented in Table 8.1. Mean (SD) total NART score for this sample was 35.3 (7.7), equivalent to an IQ of 112.3 (based on regression equations calculated by Nelson and Willison (1991)). The mean (SD) MHT score at age 11 for this sample of the LBC 1936 cohort was 50.6 (11.6) compared with a mean of 36.7 (16.1) for Scotland (N=70,805) (SCRE, 1949; Deary et al., 2012). Converted to an IQ score the mean IQ for this sample, 0.864 standard deviations above a mean of 100 (SD=15) is 113.

Items ordered from least to most *difficult* in Table 8.2 demonstrates several inconsistencies between this ordering by mean scores and the test order in this sample. For example, ‘capon’ and ‘drachm’ which are seventh and 33rd in the test administration order respectively are the 22nd and 50th item in the ordering by sample mean scores.

The Mokken automated item selection procedure partitioned 38 of the 50 items into one scale, three items into a second scale and determined the remaining nine items to be unscalable (see Table 8.3). The scalability coefficients of the 38 items of scale 1 were examined. All item-pair scalability coefficients (H_{ij} s) were non-negative and all item scalability coefficients were above 0.3 indicating that these 38 items belong in the same unidimensional Mokken scale. There were no significant violations of monotonicity. All 38 items of this abbreviated NART form a Mokken scale meeting MHM criteria ($H=0.47$, $SE=0.02$). The 38 abbreviated NART items ordered by *discrimination* are presented in Table 8.4.

These 38 items were examined for violations of non-intersection. Fifteen items violated IIO (hiatus, placebo, procreate, capon, façade, superfluous, deny, simile, banal, assignate, equivocal, puerperal, subtle, gouge, syncope) and were removed.

Table 8.2 NART items ordered by percentage of correct responses in LBC1936 (n=587) (from least to most *difficult*)

NART order	Item	Percentage correct	NART order	Item	Percentage correct
2	ACHE	99.3	32	ZEALOT	80.6
4	AISLE	99.1	28	BANAL	79.4
10	DEBT	99.0	15	CATACOMB	78.4
1	CHORD	99.0	16	GAOLED	76.8
6	PSALM	98.5	31	FACADE	75.1
18	HEIR	98.0	30	CELLIST	72.9
3	DEPOT	97.4	42	TOPIARY	72.6
9	NAUSEA	97.4	29	QUADRUPED	69.5
5	BOUQUET	96.9	36	ABSTEMIOUS	67.6
14	NAIVE	93.0	41	GAUCHE	63.2
23	PROCREATE	93.0	40	AVER	58.4
8	DENY	91.6	37	DETENTE	55.0
25	GOUGE	90.6	38	IDYLL	47.5
35	PLACEBO	89.9	19	RADIX	44.1
20	ASSIGNATE	89.8	34	AEON	42.4
11	COURTEOUS	89.4	39	PUERPERAL	40.7
22	SUBTLE	89.1	44	BEATIFY	37.3
12	RAREFY	88.6	43	LEVIATHAN	35.7
17	THYME	86.7	45	PRELATE	31.7
13	EQUIVOCAL	85.8	48	SYNCOPE	28.8
27	SIMILE	85.7	47	DEMESNE	22.0
7	CAPON	85.3	50	CAMPANILE	17.4
26	SUPERFLUOUS	84.7	46	SIDEREAL	17.2
21	HIATUS	84.7	49	LABILE	14.1
24	GIST	83.1	33	DRACHM	13.8

Note. NART=National Adult Reading Test. NART order = Item number of word order in current NART testing procedure/hierarchy (i.e. Item 1, chord, presented first). Percentage correct= percentage of respondents correctly pronouncing the items with higher percentages indicating lower *difficulty*.

Table 8.3 Partitioning of NART items by the Automated Item Selection Procedure (AISP)

	SCALE 1	SCALE 2	SCALE 0
DEPOT	SUPERFLUOUS	DRACHM	CHORD
AISLE	SIMILE	TOPIARY	ACHE
BOUQUET	BANAL	PRELATE	COURTEOUS
PSALM	QUADRUPED		RAREFY
CAPON	CELLIST		CATACOMB
DENY	FAÇADE		RADIX
NAUSEA	PLACEBO		ZEALOT
DEBT	ABSTEMIOUS		AEON
EQUIVOCAL	DÉTENTE		CAMPANILE
NAÏVE	IDYLL		
GOALED	PUERPERAL		
THYME	AVER		
HEIR	GAUCHE		
ASSIGNATE	LEVIATHAN		
HIATUS	BEATIFY		
SUBTLE	SIDEREAL		
PROCREATE	DEMESNE		
GIST	SYNCOPE		
GOUGE	LABILE		

Table 8.4 Items of the Abbreviated NART ordered by *discrimination* (H_i)

Item	Label	H_i	Item	Label	H_i
10	DEBT	0.694	22	SUBTLE	0.496
47	DEMESNE	0.673	41	GAUCHE	0.481
43	LEVIATHAN	0.604	40	AVER	0.453
46	SIDEREAL	0.601	24	GIST	0.436
4	AISLE	0.597	14	NAIVE	0.435
49	LABILE	0.582	5	BOUQUET	0.412
44	BEATIFY	0.558	20	ASSIGNATE	0.406
31	FACADE	0.556	3	DEPOT	0.405
37	DETENTE	0.543	16	GAOLED	0.400
36	ABSTEMIOUS	0.537	23	PROCREATE	0.398
18	HEIR	0.536	25	GOUGE	0.392
38	IDYLL	0.529	35	PLACEBO	0.377
26	SUPERFLUOUS	0.524	8	DENY	0.375
9	NAUSEA	0.517	13	EQUIVOCAL	0.374
48	SYNCOPE	0.513	17	THYME	0.365
30	CELLIST	0.502	6	PSALM	0.364
39	PUERPERAL	0.500	28	BANAL	0.334
27	SIMILE	0.500	21	HIATUS	0.318
29	QUADRUPED	0.497	7	CAPON	0.309

Note. Item=order of item presented in NART administration. Label=name of word to be read aloud. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*.

8.3.1 The Mini-NART

Removing the items that violated IIO resulted in a 23-item scale (the ‘Mini-NART’). These 23 items are presented in order of *difficulty* in Table 8.5. The total scale scalability coefficient for this subset was 0.534 (SE= 0.017), indicating a strong Mokken scale. H^T was 0.71, which indicates sufficient accuracy of item ordering within this scale (Ligtvoet, van der Ark, Te Marvelde & Sijtsma, 2010). Reliability was very high (MS=0.89).

The pattern of correlations between the NART and the Mini-NART and IQ measured at age 11 and age 70 are presented in Figure 8.1. The NART and the empirically derived Mini-NART positively correlated with age 11 IQ (NART: $r=.68$ $P=<0.001$, Mini-NART: $r=0.67$, $P=<0.001$). Both original and short versions of the NART correlated with age 70 IQ (NART: $r=.66$, $P<0.001$; Mini-NART: $r=.62$, $P=<0.001$).

To investigate the predictive accuracy of the total score from the 23-item Mini-NART, regression analyses were carried out. The Mini-NART accounted for 44.8% of the explained variability in age 11 IQ-tested 65 years previously-in this sample whereas the full version of the NART accounted for 46.5% of the variance. The 38-item abbreviated NART, conforming to the properties of the MHM, accounted for 48.3% of the variance. The regression equations (with 95% confidence interval (CI)) estimating an individual’s premorbid cognitive ability from performance on the Mini-NART and NART are presented below:

Mini-NART (23 item IIO scale):

Predicted age 11 IQ = 64.94 (2.345 x Mini-NART score), 95% CI [2.13 x Mini-NART score, 2.56 x Mini-NART score]

CHAPTER 8: DEVELOPMENT OF HIERARCHICAL MINI-NART

e.g. for Mini-NART score of 20, predicted age 11 IQ = $64.94 + (2.345 \times 20) = 111.84$, 95% CI [107.54, 116.14]

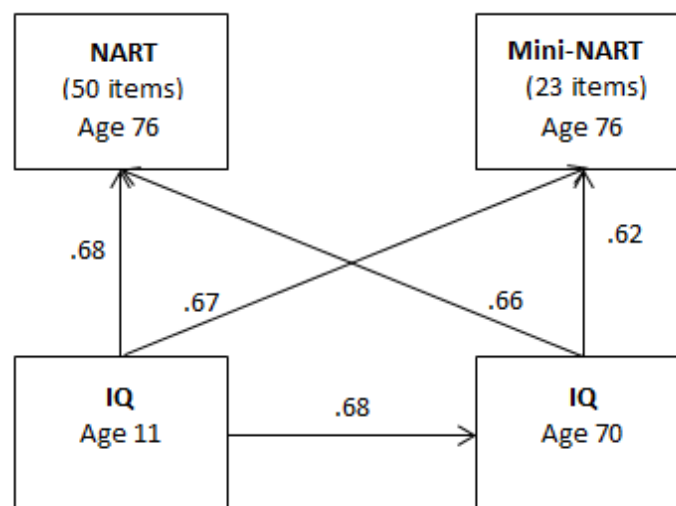
NART (original 50 item scale):

Predicted age 11 IQ = $55.97 + (1.306 \times \text{NART score})$, 95% CI [$1.19 \times \text{NART score}$, $1.42 \times \text{NART score}$]

e.g. for NART score of 45, predicted age 11 IQ $55.97 + (1.306 \times 45) = 114.74$, 95% CI [109.52, 119.87]

For ease of use Appendix C converts NART, abbreviated NART and Mini-NART scores to predicted IQ scores using these regression equations.

Figure 8.1 Correlations between age 11 IQ and the NART, Mini-NART, and age 70 IQ.



Note. IQ at both ages was assessed using the Moray House Test No. 12. NART=National Adult Reading Test. Mini-NART=Mini National Adult Reading Test.

Table 8.5 Item *difficulty* and *discrimination* of the Mini-NART

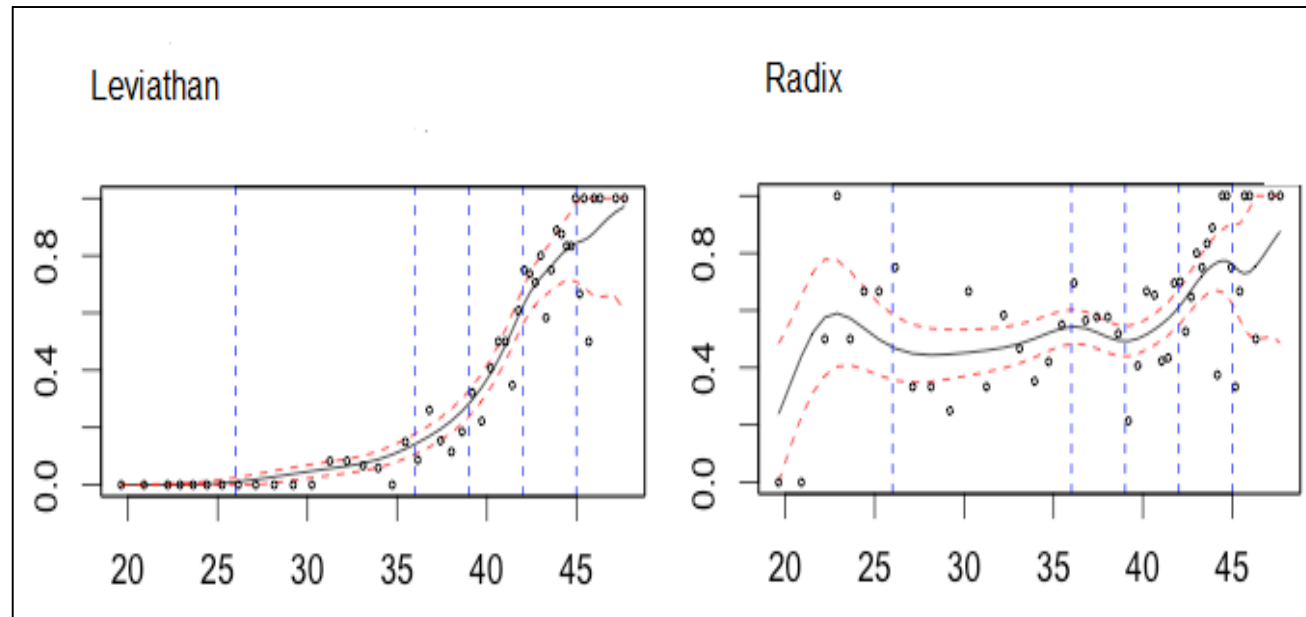
NART order	Item	H_i	% correct
4	AISLE	0.570	99.1
10	DEBT	0.592	99.0
6	PSLAM	0.409	98.5
18	HEIR	0.508	98.0
3	DEPOT	0.391	97.4
9	NAUSEA	0.483	97.4
5	BOUQUET	0.455	96.9
14	NAIVE	0.502	93.0
17	THYME	0.484	86.7
24	GIST	0.534	83.1
16	GAOLED	0.462	76.8
30	CELLIST	0.526	72.9
29	QUADRUPED	0.519	69.5
36	ABSTEMIOUS	0.541	67.6
41	GAUCHE	0.502	63.2
40	AVER	0.476	58.4
37	DETENTE	0.550	55.0
38	IDYLL	0.523	47.5
44	BEATIFY	0.561	37.3
43	LEVIATHAN	0.622	35.7
47	DEMESNE	0.701	22.0
46	SIDEREAL	0.606	17.2
49	LABILE	0.581	14.1
		$H=0.534$	

Note. NART=National Adult Reading Test. NART order= Item number of word order in current NART testing. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. H =scale scalability coefficient with higher values indicating greater scalability. % correct= percentage of respondents correctly pronouncing the items with higher percentages indicating lower *difficulty*.

8.3.2 Item discrimination

Looking at some items rejected by Mokken scaling it is clear that some NART items are not adequately distinguishing between respondents and are not contributing much to the accurate estimation of premorbid functioning. Figure 8.2 graphically presents the *discriminatory* power of two items of the NART: ‘leviathan’ (Mini-NART) and ‘radix’ (unscalable). These IRFs, produced by KernSmooth provide a representation of item *discrimination*. The slope here reflects the rate of change, designating the level of effectiveness at any point along the latent trait (De Jong & Molenaar, 1987). The poor *discrimination* value ($H_i=0.001$) of item 19 (‘radix’) is reflected in the relatively flat IRF. This means that large differences in ability are associated with very modest changes in the probability of correctly pronouncing with ‘radix’. Practically speaking, two people of different levels of ability are likely to achieve the same score on this item. This item adds little information to the overall estimate of premorbid cognitive ability as some respondents of different levels of ability have similar response profiles. The curve of item 43 (‘leviathan’) is very steep in the region of higher ability with small differences in ability at this level associated with substantial differences in the likelihood of correctly pronouncing the word.

Figure 8.2 Item response functions illustrating discriminatory power for two NART items: Item 43: 'leviathan', and Item 19: 'radix'.



Note. The x-axis represents the expected scale score. The y-axis represents the expected item score. Red dashed lines show the 95% confidence intervals. NART=National Adult Reading Test.

8.4 Discussion

The present study investigated the hierarchical nature of the NART by determining whether the data conformed to the assumptions of the MHM and IIO in 587 mostly healthy older adults with prior IQ measured at age 11. It demonstrated the utility of Mokken scaling and graphical analyses in exploring item level responses in the NART.

Two subscales within the NART were revealed (i) a 38 item abbreviated NART scale conforming to the MHM, and (ii) a 23 item Mini-NART with IIO. The items in the abbreviated NART can be stochastically ordered by degree of latent trait. However this ordering is not invariant across respondents of different levels of latent trait, i.e. the total score of this abbreviated NART, but not individual items, can be used by clinicians and researchers to obtain an estimation of a respondent's level of premorbid cognitive ability.

The Mini-NART, comprising only items strongly related to the latent trait with good *discrimination* values, conforms to a strong and invariantly ordered hierarchy. This adds value and clinical relevance to a scale as it implies a consistent ordering of items which is invariant for all values of the latent trait. Individual items within the Mini-NART can be used to approximate a respondent's level of premorbid cognitive ability. The score on a single item in the Mini-NART can represent a person's estimated prior cognitive ability, the most *difficult* item correctly responded to. This scale could be applied adaptively whereby only a section of the NART either in the higher or lower *difficulty* range of the scale needs to be applied, according to the ability of the individual patient. The test can be administered in order of ascending *difficulty* starting with 'aisle' or descending *difficulty* starting with 'labile'. For example, a participant who is able to correctly pronounce 'labile' or 'sidereal' would most likely be able to pronounce all other (less *difficult*) items in the scale. Likewise any

CHAPTER 8: DEVELOPMENT OF HIERARCHICAL MINI-NART

participant unable to correctly pronounce 'aisle' or 'debt' would most likely be unable to correctly pronounce any of the other (more *difficult*) words.

Administering IIO scales adaptively can help to reduce the time needed to test patients, reducing the burden placed on the patient helping to diminish the stress or frustration of the patient (van der Lee, Roorda, Beckerman, Lankhorst & Bouter, 2002). Although the NART in full is a relatively quick scale to administer the reading of progressively more *difficult* and infrequently encountered words aloud may still cause embarrassment and anxiety amongst those who are experiencing difficulty. Participants with early dementia or mild cognitive impairment with awareness of declining cognitive abilities are likely to be anxious facing a lengthy test battery. Shorter tests with less potential for distress and embarrassment may reduce the likelihood of participants withdrawing from testing, and may be particularly useful in clinical (medical) environments where time is limited. Adaptive testing or tailored assessment appears to be increasingly appealing in addressing the need for quick and reliable measurement. Ware et al. (2003) reported that the use of an adaptive form of the Headache Impact Survey performed better than the traditional version in terms of reducing respondent burden, measuring change over time and in test reliability and validity. Like the Rasch derived computerised-adaptive ART (Letz et al., 2003) the Mini-NART can be applied adaptively but importantly without the expense and practical implications of testing patients with a computerised test.

IRT methods can be used to ensure a scale is measuring what it is designed to measure (Langenbucher et al., 2004, Noerholm et al., 2004). With regard to the NART, 12 items were identified that did not conform to the unidimensional MHM, indicating that in this sample the NART in full includes items not measuring the same latent trait. Also Mokken scaling suggests that 'drachm', 'topiary' and 'prelate' form a separate cluster which may measure something other than premorbid cognitive ability. The inclusion of these items may

mean that the total NART score does not solely reflect premorbid cognitive ability. That the estimated premorbid IQ from the NART could be contaminated by 'noise' from other unidentified traits is a cause for concern. Rasch analysis of the ART which has several items in common with the NART identified 'aeon' and 'banal' as candidates for removal from misfit statistics (Letz et al., 2003). Neither of these items was retained in the Mini-NART, which adds validity to the removal of these items from the full NART.

By removing poor *discriminatory* items, the Mini-NART with similar predictive accuracy was identified. We have found that adding extra items to the Mini-NART does not increase the amount of variance of age 11 IQ explained in this sample. This Mini-NART, like the Short NART, offers predictive accuracy effectively equal to that of the full scale. However the Mini-NART avoids the complications of the Short NART testing process. Beardsall and Brayne (1990) suggest testing patients on the first half (Short NART) and applying a regression equation to predict the full score for patients scoring between 12 and 20 on this Short NART. If a patient scores less than 12 on the Short NART this score should be taken as the full NART score and for those scoring over 20 the full NART should be administered to determine their score. To observe these discontinuation rules a tally of errors must be kept during testing. Short NART total scores must then be converted to a NART error score before premorbid ability can be estimated. The Mini-NART requires no extra calculations and has the distinct advantage of being a hierarchical scale.

One limitation of the Mini-NART as a means of estimating premorbid cognitive ability is that with only 23 words it is not as finely graded as the full 50 item scale, or the 38 item abbreviated NART. With only 23 items it may not differentiate as efficiently between the higher levels of cognitive ability as its ceiling level of 23 items is predictive of an IQ score of 119. In this sample of 587 participants 59 have IQ scores greater than 119. However

using the full 50 item NART this ceiling is only extended by approximately two IQ points to 121. An estimated IQ based on a maximum score should be interpreted as a lower-limit estimate only with a Mini-NART score of 23 indicative of an IQ of 119 or higher.

The present analysis demonstrates the utility of IRT in examining item properties of established scales and how this insight can be used in the development of a shorter hierarchical scale. This study applied novel methods in a well-characterised sample with relatively large numbers. A particular strength of this study is the availability of a valid intelligence test score from age 11 for the sample, which ensures the scores are free from age-related decline. This permitted the validity of the Mini-NART to be assessed using the actual premorbid cognitive ability. Dykiert and Deary (2013) and Crawford, Deary, Starr and Whalley (2001) also utilised the prior ability of the LBC to examine the retrospective validity of the NART. Due to the rarity of actual premorbid ability data previous validation studies typically compared NART performance with measures of current abilities (Crawford et al., 1989; Nelson, 1982).

Some limitations of this study should be noted. The self-selected LBC1936 cohort is not fully representative of the population. Firstly, the cohort is geographically restricted. The LBC 1936 cohort is also somewhat restricted in range with regards to childhood cognitive ability. The individuals in this sample are of a higher than average ability level scoring almost 14 MHT points higher at age 11 than their peers across Scotland (Scottish Council for Research in Education, 1949; Deary et al., 2012). This is reflected in how few items there are with low percentage correct in the NART in this above-average ability sample. Performing the same analysis on a more representative sample with lower cognitive abilities with fewer participants approaching ceiling performance for many items would be a valuable extension to this analysis. Also this analysis was carried out using a sample of elderly participants without self-reported dementia. The self-reported history of dementia is subject to the

accuracy of recall. However with only 1% of participants scoring less than 24 points on the MMSE, suggesting possible dementia, the sample is mostly cognitively healthy. To examine the generalizability of these findings it is necessary to examine the accuracy of the Mini-NART in a cross-validation sample before applying the scale in clinical practice. Replication using participants with a range of abilities, and diagnoses of dementia and MCI is necessary to investigate the performance of the Mini-NART in pathological cognitive decline. Also the NART and Mini-NART account for less than 50% of the reliable variance in premorbid cognitive ability leaving a significant percentage unaccounted for. However this is a lower-bound estimate which does not account for restriction of range or measurement error.

The value of H^T here is very high and as such it is worth noting that in some cases elevated H^T values can be caused by violations of local stochastic independence (Watson, Wang & Thompson, 2014). Local stochastic independence is violated when items within a scale are linked (i.e. the response to one item is dependent on the response to another). In the case of the NART local stochastic independence is very unlikely to have been violated, as the responses are not dependent on each other.

One possible reason to explain why IIO did not hold for some items may reflect how people's knowledge of some of the more *difficult* and unusual words, some of which depend on specialist experience (e.g. medical terms like syncope, puerperal), is quite unpredictable which will have an effect on responses. This could also help to explain the inconsistencies between the item ordering by mean scores and the test administration order. The effect of regional variation in pronunciation is also likely to contribute this irregular response ordering. With regards to unscalable items it is possible that agreement between raters could be partly responsible. Crawford et al. (1989) found 'aeon' to have an agreement rate closer to chance than perfect agreement which could help to explain why this item did not follow the typical pattern of response one would expect.

8.5 Conclusion

Good scales with good psychometric properties, including IIO, are sought for accurate assessment in clinical practice and this paper demonstrates how Mokken scaling can help contribute to this goal. Mokken scaling analysis revealed that some NART items do not contribute to the measurement of premorbid cognitive ability in this sample and identified other items whose contribution is low. This analysis identified a useful, unidimensional and highly *discriminatory* scale within the NART; the Mini-NART, a hierarchical subset of 23 invariantly ordered items. While further research to support the validity of the Mini-NART, particularly in populations more representative of the general population is necessary the 23-item scale is presented as a promising alternative to the original NART for both clinicians and researchers. The Mini-NART could prove to be of clinical and practical benefit in the estimation of premorbid cognitive ability.

**Chapter 9: Lawton Instrumental Activities of Daily Living scale in dementia:
can item response theory make it more informative?**

Work presented in the following chapter is taken from the following paper:

McGrory, S., Shenkin, S. D., Austin, E. J., & Starr, J. M. (2014). Lawton IADL scale in dementia: can item response theory make it more informative? *Age and Ageing*, 43(4), 491-495.

9.1 Introduction

Functional impairment is a core feature of dementia. Diagnosis of dementia according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition and National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Dementias Association (NINCDS-ADRDA) Work Group requires a sufficient degree of cognitive impairment to cause an impairment or decline from prior level in functional ability (DSM-5; American Psychiatric Association, 2013; McKhann et al., 2011).

The loss of independence in personal care and lifestyle is among the most upsetting features of dementia for patients and their families (Cotrell & Schulz, 1993). In the earlier stages of dementia the awareness of a loss of complete independence can result in the patient feeling redundant and frustrated and burdensome due to having to rely on others to assist with activities and tasks (Cotrell & Schulz, 1993). The loss of independence in functional abilities such as feeding and moving around the home has significant consequences and indicates the requirement for a greater level of care and assistance in the home and ultimately the need for

institutionalisation. Hence, functional measures are essential for measuring disease progression. Treatments aimed at improving cognitive ability should cause an associated improvement in functional status, and any delays in the progression of cognitive deterioration should also be reflected in a slower rate of functional decline (Wattmo, Wallin & Minithon, 2012). In this way assessments of functional status can be considered as useful outcome parameters in drug trials where a change in functional ability can be used as a secondary outcome measure indicating that the effects of the drug on cognition are functionally relevant (Galasko et al., 1997; Wattmo, Wallin, Londos & Minthon, 2011).

A model of the process of functional disablement has been used to provide a conceptual framework for the development of methods of assessing functional decline (Nagi, 1964). This model provides a greater understanding of the interplay between cognitive, functional and social factors in dementia and their consequences and forms the basis for the development of tools that assess independent functioning in daily activities. The model presents three stages or processes of functional decline; *impairment*, *limitations* and *disability*. These stages represent a dynamic process of functional deterioration. With regards to dementia the initial stage of disablement commences with *impairment* due to disease pathology. This impairment then results in functional *limitations*. These *limitations* concern functional ability at a whole system level such as walking or stretching. The next and final step in this model is functional *disability*. This level of *disability* involves the interplay between disease, limitations and socio-cultural expectations (Nagi, 1964). *Disability* results from an interaction between impairments or limitations and a social context or environment. In this way the model has hierarchical implications; *difficulty* with a functional *limitation* such as stretching may or may not result in the performance of tasks such as housework, but *difficulty* with dressing is a *disability* due to its importance socially. Failure to perform such

disability activities can affect a person's health and wellbeing. Therefore disability concerns difficulties performing tasks necessary for maintaining independence.

Performance on these tasks and activities – Basic Activities of Daily Living (BADL) including activities such as dressing, bathing and eating and Instrumental Activities of Daily Living (IADL) such as handling finances, taking medication and doing housework - forms the basis for more assessments of functional ability with *disability* according to Nagi's model quantified by assessing functional performance on BADL and IADL. This chapter will focus on the assessment of IADL. IADL, requiring more complex neuropsychological organisation than BADL, are highly dependent on adequate cognitive capacity and are therefore most susceptible to the early effects of cognitive decline (Njegovan, Man-Son-Hing, Mitchell & Molnar, 2001). Assessing IADL can consequently be useful in detecting and diagnosing early dementia (Desai, Grossberg & Sheth, 2004).

Many IADL scales have been developed and they are conventionally scored by summing the responses to individual functional activity items on the scale to yield a total score. Despite the popularity of this method, there are issues that make it difficult to interpret (Reise & Henson, 2003). For example, the total-score method weights each item equally, which assumes that all items represent equal levels of severity. This is very rarely the case (Gibbons, Clark, Cavanaugh & Davis, 1985): e.g. doing laundry is more challenging than eating (Spector & Fleishman, 1998). Furthermore, the total-score method asserts that each item on the scale is equally related to the construct under examination, which again is rarely the case (Gibbons et al., 1985). These limitations can be overcome using item response theory (IRT) methods (Hambleton & Swaminathan, 1985; Fieo, Austin, Starr & Deary, 2011).

CHAPTER 9: MOKKEN SCALING ANALYSIS OF LAWTON IADL SCALE

According to IRT, the items on a scale are related to a latent construct; functional impairment in the case of the current study. IRT is based on the probability of a person achieving a particular score on a test given their standing on the latent construct (Reise & Haviland, 2005). This better reflects the underlying trait than traditional methods (Chan, Kasper, Brandt & Pezzin, 2012). IRT provides two useful measures; item *difficulty* and *discrimination*, which can provide valuable insight into the progression and assessment of functional decline. In the context of IRT, an item is considered *difficult* if a high degree of ability is required in order to respond positively. Only those with a high level of ability will be able to endorse the *difficult* items, whereas most will endorse or respond positively to the less 'difficult' items. From a clinical point of view *difficulty* can be thought of as severity, for example the degree of functional impairment required to cause challenges with handling finances. With regards to IADL items the more *difficult* a task is, the better the person's functional ability must be in order to be able to perform the task.

IRT methods also examine the *discrimination* of the scale items. *Discrimination* is the extent to which the item distinguishes participants with relatively low functional ability from those with relatively high levels of ability. An item with poor *discrimination* will distinguish poorly between mild and severe levels of functional impairment because the probability that the person will endorse the item is nearly the same across all levels of severity. An item with good *discrimination* distinguishes well between varying levels of functional ability because as the level of severity increases so too does the probability that a respondent will be unable to perform the task. For example, Fieo et al. (2010) determined 'Prepare a meal' had very weak *discriminatory* value and did not differentiate between people of different abilities, whereas 'Get on a bus' was the most *discriminatory* item differentiating between those with low functional ability and those with high functional ability. Determining item *discrimination* can identify key items on a scale and highlight weaker items or those whose function is

redundant (Reeve & Fayers, 2005). Establishing both item *discrimination* and item *difficulty* within a scale can help to examine which levels of *difficulty* have the highest level of *discrimination* and can identify if certain levels of *difficulty* are as accurately assessed.

These item parameters can be used to establish hierarchical scales of *difficulty*. A hierarchy of item *difficulty* details the expected order of functional impairment. Examining item *discrimination* can determine which levels of *difficulty* are most accurately assessed. Hierarchies are also valuable as they provide a means by which deviations in the rate of decline from the typical trajectory of loss can be identified (Fieo, Watson, Deary & Starr, 2009).

Hierarchies of IADL scales have been confirmed using IRT methods (Spector & Fleishman, 1998; Jette et al., 2002; Fieo et al., 2010). These hierarchies found that within the Townsend Functional Ability Scale (Townsend, 1962) the most *difficult* item was ‘cut toenails’ and the least *difficult* item was ‘tie a good knot in a piece of string’ (Fieo et al., 2010). ‘Active recreation’ and ‘Volunteer job’ were found more *difficult* than ‘Taking care of health’ and ‘Personal care needs’ as assessed by the Late-Life Function and Disability Instrument (Jette et al., 2002). The Lawton IADL (Lawton & Brody, 1969) scale is widely applied to measure functional status and decline in dementia and comprises eight items assessing the ability to maintain functional independence. The hierarchical pattern of decline of the eight Lawton IADL items analysed in a sample of elderly functionally disabled community dwelling participants started with the loss of independence with ‘going places outside of walking distance’ followed by ‘shopping’, ‘doing laundry’, ‘getting about inside the home’, ‘preparing meals’, ‘taking medicine’, ‘managing finances’, housework’ and finally leading to the loss of independence with ‘telephoning’ (Spector & Fleishman, 1998). These hierarchies of functional items were established in general populations.

CHAPTER 9: MOKKEN SCALING ANALYSIS OF LAWTON IADL SCALE

Lechowski et al. (2007) investigated the hierarchical structure of the Lawton IADL scale in a sample of community dwelling women diagnosed with Alzheimer's disease (N=471). This study did not apply IRT methods, instead it examined all possible the sets of losses, identified and counted the most common sequence of losses and the number of patients they represented. Using this technique this study found that for more than four fifths of the sample the sequence of IADL impairment was consistent. This most commonly occurring sequence first affected the ability to shop followed by taking medication, preparing food, travelling, managing finances, laundry, housekeeping with using the telephone the activity where independence was lost last. This homogeneous hierarchical pattern of impairment within an IADL scale outlining the functional impairment in female Alzheimer's disease patients could offer prognostic value to researchers and clinicians investigating functional impairment if these results were confirmed using IRT methods capable of establish IIO.

The Lawton IADL scale has not been analysed with IRT methods to investigate the pattern of functional impairment caused by dementia. This analysis in a sample comprising people with dementia could provide clinically useful information and can investigate whether this decline will differ from those identified in non-dementia populations and manifest in a dementia specific different ordering of impairment. Therefore, this study applied IRT methods to the Lawton IADL scale to establish a hierarchy of item *difficulties* and to assess the *discriminatory* power of the items in people with dementia.

9.2 Method

9.2.1 Participants

Data were obtained from the Scottish Dementia Research Interest Register, described in detail previously in Chapter 6. Participants were referred by a clinician, and had a diagnosis of dementia or related cognitive disorder. Two hundred and two participants for whom full itemised IADL data were available were included in the current study. Participants were assessed by clinical studies officers trained to ensure consistency.

9.2.2 Measures

The Lawton IADL scale (Lawton & Brody, 1969) (see Appendix D) assesses eight tasks providing information about functional skills necessary to live independently in the community, i.e. the ability to use the telephone, shop, prepare food, handle finances, do housework, take medications, do laundry and travel. The scale is scored out of eight with each of the eight activities scored as either 1 (can perform task independently) or 0 (not able to do) and item responses are summed to provide a total summary score ranging from 0, indicating poor functional ability and dependence, to 8, reflecting high functional ability and independence.

9.2.3 Statistical analysis

Non-parametric item response theory was used to confirm a hierarchy of item *difficulty* for the Lawton IADL Scale, and to establish the *discriminatory* power of each item in the scale. Data were analysed using the Mokken scaling analysis package in the public domain software ‘R’ (van der Ark, 2007).

The fit of the Lawton IADL items to the two probabilistic models in Mokken scaling; the MHM and IIO, formally referred to as double monotonicity, was assessed.

9.3 Results

The sample comprises 202 participants (105 male), mean age 76.39 years (SD = 7.90, range = 56–93), mean Mini Mental State Examination (MMSE) (Folstein, Folstein & McHugh, 1975) score 22.10 (range = 4–30, SD = 5.05, median = 23, interquartile range = 6), mean Lawton IADL score 4.2 (range=0-8, SD=2.1, median=4, interquartile range=3). A variety of aetiologies were included: Alzheimer's disease (AD) (133), mixed AD/VaD (35), VaD (17), frontotemporal dementia (2), dementia with Lewy bodies (4), Parkinson's disease dementia (4), mild cognitive impairment (6) and one with uncertain diagnosis.

A single scale with MS = 0.79 was obtained which indicates a reliable scale and satisfies the IRT assumption of a single unidimensional scale. The overall Loevinger's H value of 0.55 indicates a strong scale. Of the eight items included in the Lawton IADL scale 'Shopping' was the most *difficult* (mean score of 0.21) (see Table 9.1). 'Telephone use' was the least *difficult* (mean score of 0.92). Table 9.1 also shows the H_i coefficients reflecting item *discrimination* for each item.

Item scalability coefficients (H_i) were all positive and clearly exceeded the 0.3 threshold, signifying that the items meet MHM assumptions. Items can be ordered in terms of '*discrimination*' from high to low by these H_i s: 'Shopping' was the most '*discriminatory*' and 'Travelling' the least. The Lawton IADL scale showed IIO ($H^T=0.64$). This indicated that the ordering of the items is the same for all levels of the latent trait, i.e. the items are in the same order of *difficulty* regardless of the severity of functional impairment.

Table 9.1 Mokken Scaling Procedure applied to Lawton IADL scale data from a mixed dementia SDRIR sample.

Item	Mean	SD	S error	H_i
Shopping	0.21	0.41	0.03	0.71
Food Preparation	0.22	0.42	0.03	0.68
Medicine	0.26	0.44	0.03	0.53
Laundry	0.49	0.50	0.04	0.56
Finance	0.65	0.48	0.03	0.50
Travelling	0.69	0.46	0.03	0.41
Housework	0.77	0.42	0.03	0.48
Telephone use	0.92	0.27	0.02	0.62
Total H	4.20	2.12	0.15	0.55

Note. SD=Standard deviation. S error=Standard error. Range for each item is 0-1 (0 reflects dependence and 1 indicates independence). Total has a range of 0-8 with higher score reflecting higher functional ability. H_i =item scalability coefficient. Total H represents Loevinger's H coefficient for the entire scale.

9.4 Discussion

These data provide significant, novel information about the validity and practical worth of the Lawton IADL scale in this clinical sample of people with dementia. It is a strong unidimensional functional ability scale, with H of 0.55 and MS reliability statistic of 0.79. 'Shopping' and 'Food preparation' were found to be the most *difficult* items and therefore those lost earliest in the disease process. These items also demonstrated the highest levels of *discrimination*. This means that the loss of independence in shopping and preparing food occurs early and very quickly at this stage. 'Telephone use' was the least *difficult* item with problems performing this task indicating severe impairment. A patient reporting challenges with 'Telephone use' is very unlikely to be able to perform any other task in the scale. Likewise, it is likely that a patient reporting no problems with 'Shopping' or 'Food preparation' will have no limitations with other tasks. The items with the lowest *discrimination*; 'Travelling', 'Housework' are located within the less *difficult* range of the scale. This implies that the functional abilities lost later in the course of dementia progression

are measured less precisely than the initial functional abilities lost in the initial stages of dementia.

These findings have useful clinical implications. People requiring assistance with the most *difficult* item ‘Shopping’ should alert clinicians as, in the context of cognitive decline; it could herald the initial phase of functional impairment. Problems performing complex activities of daily living have been reported to precede dementia diagnosis by as much as 10 years (Peres et al., 2008). As the items of the Lawton IADL scale conform to a formal hierarchy the most *difficult* items such as ‘Shopping’ and ‘Food preparation’ can act as sensitive indicators of impending disability in the other activities (Finlayson, Mallinson & Barbosa, 2005).

Items with high *discrimination* are better able to detect differences in effects of interventions or drug therapies (Sijtsma, Emos, Bouwmesster, Nyklicek & Roorda, 2008). Ideally, a measure should comprise items of differing degrees of *difficulty* right across the spectrum of ability and demonstrate high levels of *discrimination*. This ensures that changes at every point along the ability spectrum will be detected resulting in more reliable and accurate measurement.

The inclusion of items such as ‘Shopping’ and ‘Food preparation’ which showed high *discrimination* may assist in the detection of small changes in milder stages of dementia as these abilities are lost rapidly at an early stage. ‘Telephone use’ *discriminates* well at the lower end of the hierarchy. The creation of more items such as this may help to introduce greater *discrimination* in the more advanced stages. IRT analyses can be applied to IADL/ADL scales making them more sensitive to identifying and monitoring changes in both mildly and severely impaired patients. Better assessment of the rate of decline could enhance prediction of future deterioration.

CHAPTER 9: MOKKEN SCALING ANALYSIS OF LAWTON IADL SCALE

The study was predominantly restricted to patients with mild–moderate dementia (mean MMSE score 22.1, SD = 5.05) with a range of aetiologies. Future research should investigate the loss-of-functional independence in more severe samples, and in specific dementia subtypes. For example, there is more rapid deterioration of functional abilities in patients with frontotemporal dementia compared with Alzheimer disease (Rascovsky et al., 2005). The majority of this sample (80%) was taking Cholinesterase Inhibitors, which are acknowledged to be effective in delaying or slowing the worsening of symptoms, although these effects are not large (Birks, 2006).

Limitations of the Lawton IADL scale should also be considered here. The dichotomous nature of scoring does not allow for different degrees of functional impairment to be considered implying that two people who receive the same score on the scale may have different levels of dependence on some items (i.e. an individual who is completely dependent on assistance with all shopping tasks will score 0 for this item as will an individual requiring assistance only with larger purchases). Also the scale does not take prior functional ability into account. There are likely to be significant differences between respondents level of functional activity prior to disease onset, which would have an effect on their responses regardless of functional decline due to dementia.

A particular limitation of the Lawton IADL scale is the inclusion of activities displaying gender differences in the assessment of functional decline (Lazaro Alquezar, Rubio Aranda, Sanchez Sanchez & Garcia Herrero, 2007). These gender differences could have influenced the item *difficulty* and *discrimination* parameters with the introduction of greater variability between men and women (Fleishman, Spector & Altman, 2002). This gender bias can result in the overestimation of functional dependency in men (Graf, 2009). For example, it is likely that some men would score zero for ‘food preparation’, as cooking and food preparation would have been done by their wife, even though they may have the

ability to do so. Therefore their score is not due to functional disability. This overestimation of functional dependence due to gender bias would also influence scores on ‘managing finances’ where many women would have no experience with this because their husbands’ would have taken responsibility for this task. This is particularly pertinent in older cohorts. The scale should be adapted to consider these issues and in the meantime stratified analyses of item *difficulty* and item *discrimination* should be carried out to determine if there are different patterns of functional decline for men and women.

IRT has benefits not only in the monitoring of patients, but establishing the sequence of decline, which can also help in characterising adaptations to disability and differences between subgroups. While additive summary scores can be helpful in summarising overall function, they can conceal as much information as they reveal, and IRT methods are a useful method to increase the information provided by simple functional scales. Furthermore, simultaneous analyses of cognitive and functional scales could enable the discovery of more precise associations between cognitive and functional outcomes.

Chapter 10: Patterns of decline in Instrumental Activities of Daily Living across different types of dementia: Extension of Mokken scaling analysis on the Lawton IADL scale in the SDRIR

10.1 Introduction

In Chapter 9 the hierarchical structure of the Lawton IADL scale (Lawton & Brody, 1969) was examined in a sample from the Scottish Dementia Research Interest Register (SDRIR). This sample consisted of a mixture of aetiologies including Alzheimer's disease, vascular dementia, mixed Alzheimer's disease/vascular disease, frontotemporal dementia, dementia with Lewy bodies (DLB), Parkinson's disease dementia (PDD) and mild cognitive impairment (MCI). In this sample the items of the Lawton IADL scale conformed to a strong hierarchical scale with items ordered by decreasing *difficulty*, with 'shopping' as the most *difficult* and 'using the telephone' as the least *difficult* item. However, the numbers in this sample were not sufficient for the item ordering in any diagnostic group to be analysed separately. Therefore the aim of this chapter is to perform a stratified Mokken analysis by diagnosis and gender using the same case register. At the time of the IADL analysis in Chapter 9 (2012) the SDRIR comprised 202 patients with complete IADL data. Returning to the register in 2014 allows for a more in-depth analysis due to the availability of an additional 623 patients with complete itemized IADL data.

As discussed in Chapter 1 performance of IADLs is necessary for maintaining independence with declining IADL performance resulting in functional dependence. In dementia poor functional ability is associated with greater health care cost, decreased quality of life, increased caregiver burden and decreased time to institutionalization (Andersen, Wittrup-Jensen, Lolk, Andersen & Kragh-Sorensen, 2004; Handels, Wolfs, Aalten, Verhey &

Severens, 2013; Vetter et al., 1999). Accurate and appropriate assessment of functional ability in dementia supports healthcare providers to provide adequate counselling and advice concerning safety and the need for institutionalization (Desai, Grossberg & Sheth, 2004).

Hierarchical scales can provide more information and insight into the pattern of functional decline. Identifying distinct patterns of functional decline for different dementia syndromes can not only help to differentiate between dementia pathologies but can also provide scope for the development of interventions specifically designed to promote independence and quality of life for both patients and carers in specific patient groups. Identifying differences in item ordering between male and female patients can help to establish targeted interventions aimed at maintaining independence. Establishing a formal hierarchy of functional decline in IADLs can reveal the earliest stages of impairment process and can be valuable in the identification of a key variable which may be important in terms of understanding or monitoring patients' functional status (Morris, Fries & Morris, 1999).

10.1.1 Determinants of functional ability in dementia

Identifying the factors underlying IADL performance is a crucial step in the identification of potential differences between patients and in the development of interventions. The most significant determinant of performance of IADLs in dementia is cognition (Galasko, 1998). While cognitive processes exert significant influence on functional ability in dementia other non-pathological aspects of ageing such as loss of strength, mobility, hearing and sight in addition to behavioural factors are also likely to affect functional ability. In the general population other diseases such as rheumatoid arthritis, osteoarthritis and fibromyalgia affect functional ability (Waehrens, Bliddal, Danneskiold-Samsøe, Lund & Fisher, 2012).

An exploratory factor analysis of the Lawton IADL scale using data from a non-institutionalised sample of older adults (N=1072) suggested a two-dimensional structure with

cognitive and physical domains (Ng, Niti, Chiam & Kua, 2006). The cognitive domain assesses the ability to use the telephone, take medication and manage finances. The physical domain measures performance in the other five tasks; doing laundry, housework, travelling, food preparation and shopping. Given that cognition is considered the most significant predictor of functional impairment in dementia it could be expected that the items in the cognitive domain would be lost first, followed by the abilities assessed by the items in the physical domain. On the other hand for some forms of dementia with more pronounced physical symptoms (e.g. Parkinson's disease dementia) the items more closely approximated with physicality may represent a greater challenge. Ng et al. (2006) did not assess this as only 7.4% of the sample analysed had been diagnosed with dementia.

Differences in the extent to which determinants of functional ability are present in different kinds of dementia could influence the extent of and pattern of functional decline. Among patients with vascular dementia, longitudinal changes in IADLs are most strongly associated with changes in executive functioning (Jefferson et al., 2006). This is consistent with similar findings associating executive dysfunction to IADL performance in Alzheimer's disease (Boyle et al., 2003; Cahn-Weiner, Ready & Malloy, 2003). These findings suggest that IADL performance for both Alzheimer's disease and vascular dementia appears to be driven by executive function. Executive deficits have also been associated with impaired functional ability in frontotemporal dementia (Mioshi et al., 2007). Executive functions, including attention, working memory and the planning and execution of complex goal-directed tasks, are necessary for performance of many IADL tasks such as paying household bills, taking medication and shopping (Cahn-Weiner, Malloy, Boyle, Marran & Salloway, 2000).

Executive dysfunction is more common and severe in patients with vascular dementia in comparison to patients with Alzheimer's disease (Mathias & Burke, 2009; Looi &

Sachdev, 1999; Roman & Royall, 1999). Impairment of motor skills is less common and may have less influence on IADL performance in Alzheimer's disease than dementia with Lewy bodies and vascular dementia where motor impairments are of particular concern (McKeith et al., 2005; Chen, Sultzer, Hinkin, Mahler & Cummings, 1998). With observed impairments in fine motor speed and dexterity (Roman & Royall, 1999) it is not surprising that patients with vascular dementia would exhibit difficulty performing tasks requiring precise motor control such as writing a cheque or performing household chores.

Impaired motor skills have been linked to functional dysfunction in vascular dementia. Poor IADL performance in vascular dementia could reflect physical or sensory-motor impairments secondary to stroke (Waite, Broe, Grayson & Creasey, 2000; Boyle, Cohen, Paul, Moser & Gordon, 2002).

Another possible determinant of functional ability in dementia is apathy (Mortimer, Ebbitt, Jun & Finch, 1992). Apathy has been associated with poor IADL scores in Alzheimer's disease and frontotemporal dementia (Boyle et al., 2003; Mioshi et al., 2007). In Alzheimer's disease 44% of the variance in IADL performance is accounted for by executive cognitive dysfunction (17%) and apathy (27%) (Boyle et al., 2003). Apathy, while found in all types of dementia (Clarke et al., 2008) is most prevalent in Alzheimer's disease and vascular dementia (Landes, Sperry & Strauss, 2005; Jonsson, Edman, Lind, Rolstad, Sjogren & Wallin, 2010). Apathy is also a prominent feature of frontotemporal dementia (Mioshi et al., 2007; Mendez, Lauterbach & Sampson, 2008). While apathy is common to dementia in general a recent review of the validity and reliability of apathy scales (Radakovic, Harley, Abrahams & Starr, 2014) noted the different apathy profiles which can be identified between Alzheimer's disease and frontotemporal dementia (Quaranta, Marra, Rossi, Gainotti & Masullo, 2012). These different profiles could cause variations in functional ability.

10.1.2 Differential item functioning

IRT methods can also provide an assessment of differential item functioning (DIF). DIF occurs when the item response function (IRF) for a given item differs between subgroups taken from the population of interest (Sijtsma & Molenaar, 2002). DIF in measurement theory concerns items that display differences in terms of how they are responded to by samples comprised of individuals with the same level of trait being measured (Hambleton, Swaminathan & Rogers, 1991). Item *difficulty* DIF reflects a disparity in how *difficult* an item is across different subgroups; an item with a lower degree of *difficulty*, making it more likely to be responded to or answered correctly at a lower level of ability, for one sample in comparison to another. DIF can arise due to demographic influences such as gender or age (Fleishman, Spector & Altman, 2002).

DIF can be introduced in the assessment of patients with dementia using IADL scales. There is likely to be bias due to potential gender differences and diagnostic differences. DIF can arise due to differences between different types of dementia. For example, the severity and rate of functional decline in patients with behavioural-variant frontotemporal dementia has been shown to exceed that of Alzheimer's disease (Mioshi, Kipps & Hodges, 2009). A comparison of functional and cognitive impairment in patients with bv-FTD and AD found differences between how the two patient groups approached functional activities with bv-FTD patients experiencing greater motivational and organisation problems than the patients diagnosed with AD (Lima-Silva et al., 2014). If such differences between diagnostic groups influence the *difficulty* of individual items within the IADL scale DIF can be demonstrated. Gender and age DIF has been identified for IADL items using data from a population comprising younger and older adults (N=5750) (Fleishman et al., 2002). The items with the greatest degree of bias due to age and gender were 'shopping' and 'finances'. Compared to men aged 70 and over younger men were more likely to demonstrate functional dependency

with managing their finances and both men and women in the youngest age groups (18-39 years) were less functionally impaired with shopping. DIF has been demonstrated in IADL performance across international surveys of ageing (Chan, Kasper, Brandt & Pezzin, 2012). In Lawton's scale development analysis in three items of the Lawton IADL scale failed to form a Guttman scale in a male sample. All eight items met Guttman scaling criteria in the female sample analysed leaving Lawton to speculate that the sex-linked content of the three items excluded from the male Guttman scale ('food preparation', 'laundry' and 'housework') was responsible for their failure to conform to the scale (Lawton & Brody, 1969).

An advantage of item response theory methods is its application in the assessment of item-bias or differential item functioning (DIF). An item may demonstrate DIF if the item has varying item properties in different subgroups. This indicates the influence of a variable other than the one being assessed (Thissen, Steinberg & Wainer, 1993). In Mokken scaling analysis DIF is assessed by checking the assumption of equal item ordering (Roorda, Houwink, Smits, Molenaar & Geurts, 2011). DIF is determined when different item orderings emerge within the subgroups analysed. This study aims to explore whether differences in item ordering, demonstrating DIF, across different types of dementia will be observed.

10.1.3 The present study

If the differences in these contributing factors to performance on IADLs between different forms of dementia are large enough it is possible that different, disease specific, patterns of functional decline may be observed. If the different patterns and symptoms of cognitive decline across different types of dementia affect the order of functional decline it can be expected that IIO will not hold in the full heterogeneous sample.

It could be expected that a patient's level of functional decline would be related to their diagnosis and potentially gender. The combination of responses from both men and

women with numerous diagnoses makes it less likely that the items will fail to conform to a hierarchical structure due to the greater heterogeneity, which would be reflected in violations of IIO in the IRT analyses.

However, Chapter 9 found that the items of the Lawton IADL scale formed an invariantly ordered hierarchy in a heterogeneous sample of dementia patients (N=202), which suggests that the pattern of functional decline is more robust than expected across dementia subtypes. However Chapter 9 was based on a relatively small sample with insufficient numbers for the exploration of the hierarchical structure between different types of dementia and gender. Therefore, this chapter aims to use a larger sample to examine item properties across relevant subgroups which provides a method of assessing differential item functioning which may affect comparisons by subgroup.

Therefore the aim of the present study is to investigate whether (i) this IIO hierarchy of Lawton IADL items is confirmed in a larger sample and (ii) whether this hierarchical structure demonstrates DIF by diagnosis, and (iii) whether items display DIF by gender.

10.2 Method

10.2.1 Participants

Data were obtained from the Scottish Dementia Research Interest Register (SDRIR). The database for this study was declared in March 2014. Nine hundred and sixteen (498 male, 418 female) patients with complete itemised IADL data were on the register. A variety of diagnoses were represented; late-onset AD (N=477); mixed AD VaD (N=138); VaD (N=99); young-onset AD (N=69); DLB (N=20); FTD (N=12); PDD (N=10); MCI (N=10) other dementia (N=37) and 44 with uncertain diagnosis.

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

Of these 916 participants 825 were included for analyses. As the aim of these analyses was to investigate IADL item properties and patterns of functional loss in dementia those with uncertain (N=44), people with MCI (N=10) and other dementia (N=37) were excluded. The remaining 825 register participants comprised the first dataset for analysis.

From this full dataset (N=825) four subsamples were formed; (i) patients diagnosed with late onset AD (N=477), (ii) patients diagnosed with mixed AD VaD (N=138) (iii) patients diagnosed with mixed AD VaD plus patients diagnosed with VaD (N=237) and (iv) patients diagnosed with non-Alzheimer's disease pathology; VaD, DLB, FTD, PDD (N=142). In order to increase numbers in each subgroup there is some degree of overlap between samples (i.e. patients with mixed AD and VaD are common to both groups ii and iii and patients diagnosed with VaD are common to both groups iii and iv. The full sample was also divided by gender. These subgroups along with the full mixed sample were analysed to determine if the pattern of functional decline differed between groups of people with a specific diagnosis or gender.

10.2.2 Measures

The Lawton IADL scale measures the ability to perform tasks necessary to maintain independence ('telephone use', 'shopping', 'food preparation', 'housework', 'laundry', 'travelling', 'taking medication' and 'handling finances'). Responses to each item are coded dichotomously as 0 (unable or partially able) or 1 (able) and are summed to provide a total summary score ranging from 0, indicating poor functional ability and dependence, to 8, reflecting high functional ability and independence. Cognitive functioning was assessed by the ACE-R which is scored out of 100 with higher scores indicating better cognitive ability.

10.2.3 Mokken scaling analysis

All Mokken scaling analyses were performed using the ‘Mokken’ package in R. The fit of the eight scale items to the monotone homogeneity model and IIO was examined in each of the seven samples.

10.3 Results

Descriptive statistics for cognitive and functional ability from a sample of 825 register patients (441 male, 384 female) plus the four diagnostic subgroups and gender specific samples are presented in Table 10.1.

Table 10.1 Characteristics of the SDRIR samples analysed

	N	Sex (% male)	Mean age	Mean ACER	Mean Lawton IADL
Complete sample	825	441 (53.4%)	77.6	60.4	3.9
Male	441		77.0	62.0	3.5
Female	384		78.1	58.5	4.4
Late onset AD	477	229 (48%)	79.4	60.5	4.0
Mixed AD VaD	138	78 (56.5%)	78.8	62.2	3.5
Mixed AD VaD plus VaD	237	142 (60%)	78.2	61.3	3.5
Non-AD	142	96 (67%)	77.0	58.1	3.4

Note. AD=Alzheimer’s disease, VaD=vascular dementia, ACE-R=Addenbrookes Cognitive Examination-Revised, IADL=Instrumental Activities of Daily Living.

Mean Lawton IADL item scores for each sample are presented in Table 10.2 in order of mean score for the complete sample, from most to least *difficult*, with lower mean scores indicating poor functional ability.

Table 10.2 Mean IADL item scores for SDRIR sample plus four diagnostic subgroups

Complete SDRIR sample		Late onset AD		Mixed AD VaD		Mixed AD VaD + VaD		Non-AD	
Item	Mean	Item	Mean	Item	Mean	Item	Mean	Item	Mean
Shopping	0.18	Shopping	0.20	Shopping	0.17	Shopping	0.16	Shopping	0.12
Food prep	0.19	Food prep	0.21	Food prep	0.14	Food prep	0.15	Food prep	0.12
Medicine	0.25	Medicine	0.26	Medicine	0.19	Medicine	0.24	Medicine	0.25
Laundry	0.44	Laundry	0.47	Laundry	0.42	Laundry	0.38	Laundry	0.31
Finance	0.62	Finance	0.63	Finance	0.57	Finance	0.56	Finance	0.57
Travelling	0.64	Travelling	0.66	Travelling	0.57	Travelling	0.58	Travelling	0.60
Housework	0.75	Housework	0.86	Housework	0.69	Housework	0.66	Housework	0.63
Telephone	0.85	Telephone	0.86	Telephone	0.83	Telephone	0.85	Telephone	0.85

Note. IADL=Instrumental Activities of Daily Living. AD=Alzheimer’s disease. VaD=vascular dementia. Item mean range for each item is 0-1 (0 indicates impaired ability and 1 indicates no impairment).

(i) Full SDRIR sample (N=825)

Complete itemised IADL data for the complete SDRIR sample were analysed. This mixed sample comprised patients with several dementia diagnoses (late onset AD n=477, Mixed AD VaD n=138, VaD n=99, early onset AD n=69, DLB n=20, FTD n=12, PDD n=10). Mean scores for the sample are presented in Table 10.2.

Mokken scaling of the full sample determined that items formed a unidimensional Mokken scale. All items formed a single item cluster using the automated item selection procedure (AISP). All item-pair scalability coefficients (H_{ij}) were nonnegative and item scalability coefficients (H_i) were greater than the lower bound threshold (0.3). There were no exclusions due to violations of monotonicity. The scale scalability coefficient was 0.55.

One item ('food preparation') was removed in the assessment of IIO. The remaining 7 items (see Table 10.3) formed a reliable (MS=0.76) strong hierarchical scale ($H=0.53$) with IIO ($H^T=0.58$).

These findings are compared with Chapter 9's results in Table 10.4. While there are differences in mean item scores and scalability coefficients the hierarchical ordering of items by *difficulty* and *discrimination* between the analyses of the complete sample from the current study with the sample analysed in Chapter 9 demonstrates a consistent pattern of decline.

Table 10.3 IIO hierarchy items from complete sample (N=825) listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Shopping	0.18	Shopping	0.62
Medicine	0.25	Telephone	0.57
Laundry	0.44	Laundry	0.53
Finance	0.62	Medicine	0.51
Travelling	0.64	Finance	0.51
Housework	0.75	Housework	0.50
Telephone	0.85	Travelling	0.49

Note. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*.

Table 10.4 Comparison of hierarchies from complete sample from the present study (N=825) and Chapter 9 (N=202). Items are listed from most to least *difficult* and most to least *discriminatory*

Item	Chapter 9			Item	Chapter 10		
	Mean	Item	H_i		Mean	Item	H_i
Shopping	0.21	Shopping	0.71	Shopping	0.18	Shopping	0.62
Food Prep	0.22	Food prep	0.68	Medicine	0.25	Telephone	0.57
Medicine	0.26	Telephone	0.62	Laundry	0.44	Laundry	0.53
Laundry	0.49	Laundry	0.56	Finance	0.62	Medicine	0.51
Finance	0.65	Medicine	0.53	Travelling	0.64	Finance	0.51
Travelling	0.69	Finance	0.50	Housework	0.75	Housework	0.50
Housework	0.77	Housework	0.48	Telephone	0.85	Travelling	0.49
Telephone	0.92	Travelling	0.41				

Note. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*.

10.3.1 Differential item functioning assessment by diagnosis

(ii) Late onset AD (N=477)

Complete itemised IADL data for SDRIR participants diagnosed with late onset AD were analysed. Mean scores for the sample are presented in Table 10.2. All items formed a single item cluster using AISP. All item-pair scalability coefficients (H_{ij}) were non-negative and

item scalability coefficients (H_i) were greater than the lower bound threshold (0.3). There were no exclusions due to violations of monotonicity. The scale scalability coefficient was 0.54.

One item ('shopping) was removed in the assessment of IIO. The remaining 7 items (see Table 10.5) formed a reliable ($MS=0.76$) strong hierarchical scale ($H=0.53$) with IIO ($H^T=0.58$).

Table 10.5 IIO hierarchy items from IADL in late onset AD listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	Hi
Food prep	0.21	Food prep	0.66
Medicine	0.26	Telephone	0.56
Laundry	0.47	Housework	0.54
Finance	0.63	Laundry	0.53
Travelling	0.66	Finance	0.53
Housework	0.78	Medicine	0.50
Telephone	0.86	Travelling	0.47

Note. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*.

(iii) Mixed AD VaD (N=138)

Complete itemised IADL data for SDRIR participants diagnosed with mixed AD VaD were analysed. Mean scores for the sample are presented in Table 10.2. Mokken scaling of this SDRIR sample determined that the eight IADL items formed a strong Mokken scale ($H=0.57$). Each item was partitioned into a single cluster using AISP and all item scalability coefficients were above the 0.3 cut-off level with no nonnegative item-pair scalability coefficients. There were no violations of monotonicity. There were no exclusions in the assessment of non-intersection. The eight items formed a reliable ($MS=0.80$) strong

hierarchical scale ($H=0.57$) with IIO ($H^T=0.62$). The invariantly ordered IADL items are presented in Table 10.6 in order of *difficulty* and *discrimination*.

Table 10.6 IIO hierarchy items from IADL in patients with Mixed AD VaD (N= 38) listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	Hi
Food Prep	0.14	Telephone	0.71
Shopping	0.17	Food prep	0.63
Medicine	0.19	Shopping	0.63
Laundry	0.42	Laundry	0.55
Finance	0.57	Finance	0.55
Travelling	0.57	Housework	0.52
Housework	0.69	Travelling	0.49
Telephone	0.83	Medicine	0.47

Note. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*.

(iv) Mixed AD VaD and VaD (N=237)

Complete itemised IADL data for SDRIR participants diagnosed with mixed AD VaD plus VaD were analysed. Mean scores for the sample are presented in Table 10.2. Mokken scaling of this SDRIR sample determined that the eight IADL items formed a strong Mokken scale ($H=0.56$). Each item was partitioned into a single cluster using AISP and all item scalability coefficients were above the 0.3 cut-off level with no nonnegative item-pair scalability coefficients. There were no violations of monotonicity. There were no exclusions in the assessment of non-intersection. The eight items formed a reliable ($MS=0.80$) strong hierarchical scale ($H=0.56$) with IIO ($H^T=0.62$). The invariantly ordered IADL items are presented in Table 10.7 in order of *difficulty* and *discrimination*.

Table 10.7 IIO hierarchy items from IADL Mixed AD VaD and VaD (N=237) listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	Hi
Food Prep	0.15	Telephone	0.67
Shopping	0.16	Shopping	0.65
Medicine	0.24	Food prep	0.62
Laundry	0.38	Laundry	0.60
Finance	0.56	Finance	0.55
Travelling	0.58	House	0.53
Housework	0.66	Travelling	0.52
Telephone	0.85	Medicine	0.47

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

(v) Non-AD pathology (N=142)

Data from all patient groups excluding those with a diagnosis of late onset AD, young onset AD or mixed AD/VaD were isolated for analysis. This sample comprised patients with various non-Alzheimer's pathology; VaD (N=99), DLB (N=21), FTD (N=12), PDD (N=10). Mean scores for the sample are presented in Table 10.2.

The eight IADL items met the assumptions of the monotone homogeneity model. In more detail these items formed a single unitary item cluster using the automated item selection procedure. The item pair scalability coefficients were non-negative and the values of all item scalability coefficients exceeded 0.3. There were no violations of monotonicity. The scale met the requirements of a strong Mokken scale ($H=0.50$).

The assessment of the fit of the eight items to the assumptions pertaining to IIO in this population found no violations of IIO. The items of the Lawton IADL scale formed a reliable ($MS=0.77$), strong ($H=0.50$) Mokken scale with $H^T=0.60$. The invariantly ordered IADL items are presented in Table 10.8 in order of *difficulty* and *discrimination*. The IIO

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

hierarchies of IADL items from each of the three Mokken scaling analyses are presented in Table 10.9.

Table 10.8 IIO hierarchy items from IADL non-AD listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	Hi
Food prep	0.12	Shopping	0.65
Shopping	0.12	Food prep	0.65
Medicine	0.25	Laundry	0.57
Laundry	0.31	Travelling	0.46
Finance	0.57	Telephone	0.45
Travelling	0.60	Medicine	0.45
Housework	0.63	Finance	0.44
Telephone	0.85	Housework	0.43

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

Table 10.9 IADL IIO hierarchies: comparison across diagnostic groups

Complete SDRIR sample		Late onset AD		Mixed AD VaD		Mixed AD VaD +VaD		Non-AD	
Item	Mean	Item	Mean	Item	Mean	Item	Mean	Item	Mean
Shopping	0.18	Food prep	0.21	Food Prep	0.14	Food Prep	0.15	Food prep	0.12
Medicine	0.25	Medicine	0.26	Shopping	0.17	Shopping	0.16	Shopping	0.12
Laundry	0.44	Laundry	0.47	Medicine	0.19	Medicine	0.24	Medicine	0.25
Finance	0.62	Finance	0.63	Laundry	0.42	Laundry	0.38	Laundry	0.31
Travelling	0.64	Travelling	0.66	Finance	0.57	Finance	0.56	Finance	0.57
Housework	0.75	Housework	0.78	Travelling	0.57	Travelling	0.58	Travelling	0.60
Telephone	0.85	Telephone	0.86	Housework	0.69	Housework	0.66	Housework	0.63
				Telephone	0.83	Telephone	0.85	Telephone	0.85
$H^T=0.58$		$H^T=0.58$		$H^T=0.62$		$H^T=0.62$		$H^T=0.60$	

Note. AD=Alzheimer’s disease. VaD=vascular dementia. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. H^T =H Trans is a measure of the accuracy of item ordering within a scale with higher numbers reflecting greater accuracy

10.3.2 Differential item functioning assessment by gender

(i) Male (N=441)

Complete itemised male IADL data for the complete SDRIR sample were analysed. This mixed sample comprised patients with several dementia diagnoses (late onset AD, N=229; mixed AD VaD, N=78; VaD, N=64; early onset AD, N=38; DLB, N=14; FTD, N=9; PDD N=9). Mean scores for the sample are presented in Table 10.2.

All items formed a single item cluster using AISP. All item-pair scalability coefficients (H_{ij}) were nonnegative and item scalability coefficients (H_i) were greater than the lower bound threshold (0.3). There were no exclusions due to violations of monotonicity. The scale scalability coefficient was 0.53. There were no violations of IIO. The eight scale items formed a reliable (MS=0.77), strong Mokken scale (H=0.53) with IIO ($H^T=0.77$) (see Table 10.10).

Table 10.10 IIO hierarchy items from full sample-male-listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	Hi
Food preparation	0.13	Food preparation	0.63
Shopping	0.15	Shopping	0.60
Laundry	0.20	Telephone	0.57
Medicine	0.27	Laundry	0.57
Finance	0.59	Travelling	0.53
Housework	0.65	Finance	0.51
Traveling	0.68	Medicine	0.49
Telephone	0.82	Housework	0.43

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

(ii) Female (N=384)

Complete itemised female IADL data for the complete SDRIR sample were analysed. This mixed sample comprised patients with several dementia diagnoses (late onset AD n=248, Mixed AD VaD n=60, VaD n=35, early onset AD n=31, DLB n=6, FTD n=3, PDD n=1). Mean scores for the sample are presented in Table 10.2.

All items formed a single item cluster using AISP. All item-pair scalability coefficients (H_{ij}) were nonnegative and item scalability coefficients (H_i) were greater than the lower bound threshold (0.3). There were no exclusions due to violations of monotonicity. The scale scalability coefficient was 0.64. One item ('shopping') was removed due to a violation of IIO. The remaining seven items formed a very reliable (MS=0.82), strong Mokken scale ($H=0.65$) with IIO ($H^T=0.72$) (see Table 10.11).

Table 10.12 presents the hierarchical scales for the male and female samples. Differential item functioning by gender is graphically presented in Figure 10.1.

Table 10.11 IIO hierarchy items from full sample-female- listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Medicine	0.22	Food preparation	0.71
Food preparation	0.27	Medicine	0.69
Travelling	0.60	Housework	0.68
Finance	0.65	Laundry	0.67
Laundry	0.71	Finance	0.62
Housework	0.86	Travelling	0.62
Telephone	0.88	Telephone	0.56

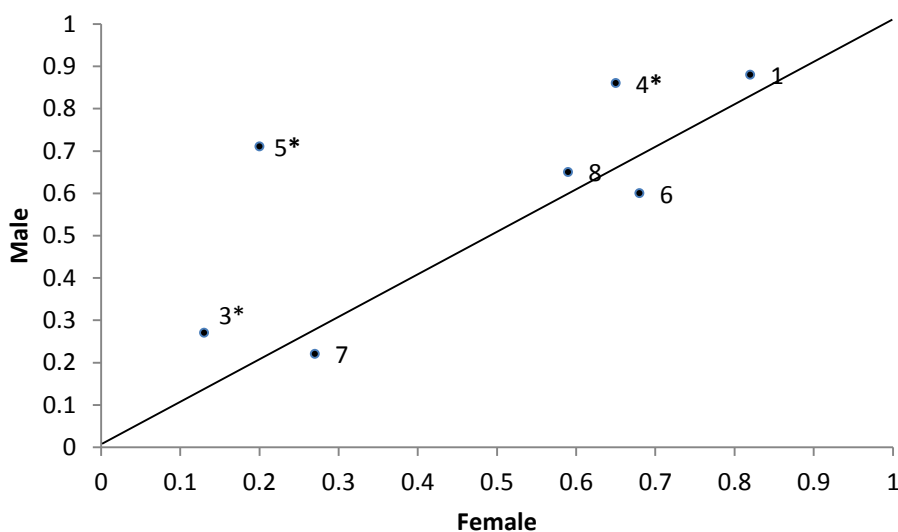
Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

Table 10.12 IADL IIO hierarchies: comparison across gender

Male		Female	
Item	Mean	Item	Mean
Food preparation	0.13	Medicine	0.22
Shopping	0.15	Food preparation	0.27
Laundry	0.20	Travelling	0.60
Medicine	0.27	Finance	0.65
Finance	0.59	Laundry	0.71
Housework	0.65	Housework	0.86
Travelling	0.68	Telephone	0.88
Telephone	0.82		
$H^T=0.77$		$H^T=0.72$	

Note. Mean item scores (range from 0-1) with higher scores indicating lower difficulty. $H^T=H$ Trans is a measure of the accuracy of item ordering within a scale with higher numbers reflecting greater accuracy.

Figure 10.1 Mean item scores for male and female participants.



Note. 1=telephone, 3=food preparation, 4=housework, 5=laundry, 6=travelling, 7=medicine, 8=finance. Mean item scores for females are on the x axis and for male, on the y axis. Low mean item scores designate the items indicating severe functional impairment, and high mean item scores designate the items indicating minor functional impairment. The numbers near the data points refer to the item numbers. An identity line is drawn through the origin with a slope of 1. * = Items with major deviations from the identity line demonstrating differential item functioning (food preparation, housework, laundry)

10.4 Discussion

This analysis revealed a consistent IIO hierarchy across the sample as a whole and the four diagnostic subsets. In the analyses of the full dataset and the late onset AD group one item, ‘food preparation’ was excluded due to an IIO violation. In these samples ‘Shopping’ is the most *difficult* task. In the mixed AD VaD group ‘Food preparation’ is the most *difficult* item. Using the telephone is the least *difficult* task across all groups. Item ordering differed substantially between male and female participants with most items (‘travelling’, ‘medicine’, ‘finance’, ‘laundry’, ‘housework’, ‘food preparation’) demonstrating gender DIF.

That the item ordering is consistent across these groups suggesting that the pattern of decline in IADLs is not influenced by the different dementia variations represented here. The item ordering here is consistent with the analysis from Chapter 9 with the exception of the mixed group here where ‘food preparation’ is more *difficult* than ‘shopping’. The opposite pattern is observed in the hierarchy from Chapter 9 where ‘shopping’ was found to be more *difficult* than ‘food preparation’. This generally consistent pattern between the sample analysed here - which consisted of 825 people with a range of dementia diagnoses - and the sample analysed in Chapter 9 – which consisted of 202 people mostly with Alzheimer’s disease (N=133) diagnoses - suggests that the pattern of functional decline may be generally consistent across different dementia profiles. However it is important to note that this claim is made based on the results of one sample and importantly the sample from Chapter 9 is included in the current chapter’s larger sample. These considerations mean the results here must be interpreted with caution. That the item ordering holds up despite the diagnostic heterogeneity suggests that while the determinants of functional decline may differ between dementia types they manifest in the same pattern of decline. While it cannot be determined if the mechanisms underlying loss of functional ability are common across these samples the invariant hierarchical pattern indicates that the cognitive decline in dementia appears to result

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

in the same pattern of loss as measured by the Lawton IADL scale. As noted earlier, however the Lawton scale is a crude measure of functional status which implies that again results must be interpreted with caution. Bearing these limitations in mind the results here suggest while different diseases may cause different patterns of cognitive decline, through several mechanisms, these processes affect the ability to function independently in the same way.

Spector and Fleishman (1998) also confirmed an invariant item ordering of Lawton IADL items in their Rasch analysis in a sample of functionally disabled individuals aged 65 and over. They found a similar pattern of decline with 'shopping' being the most *difficult* and 'telephone' being the least *difficult* item in this sample. While there are some alterations in the ordering between these extremes this research again supports that as the most *difficult* item, shopping can be used to identify differences in functional ability at the lower end of the range and 'telephone', as the least *difficult* item, can help to identify changes in severe functional disability. That the same items emerge as the most and least *difficult* in a representative sample of the disabled elderly in the community suggests that not only do these items assess the upper and lower ranges of functional decline in dementia but that they also assess the same levels of ability in a sample of elderly without dementia. This suggests that 'shopping' and 'telephone' abilities can act as indices of initial and severe functional decline in both normal ageing and dementia. It is unlikely that there are no differences between people with dementia compared to those ageing normally. The consistencies reported here are most likely artefacts of the scale's limited number of items at either end of the *difficulty* spectrum.

Due to the close association between IADLs and cognition (Galasko, 1998) and the application of this scale to detect functional deterioration due to the cognitive decline of dementia it could be expected that the items lost first would be those more closely approximated with cognitive complexity. However, the order of decline revealed from these

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

analyses does not adhere to the pattern of decline expected from the two-dimensional structure of the scale (Ng et al., 2006) (i.e. cognitive domain items; using the telephone, taking medication and managing finances lost first). The current results show significant amount of overlap in terms of *difficulty* between the items of the cognitive domain and those in the physical domain. For example, ‘telephone’, one of the ‘cognitive’ items is the least *difficult* item and ‘shopping’-a ‘physical’ item is the most *difficult* and therefore the first ability with which patients lose independence. This lack of clear delineation in ordering is evident across all of the analyses carried out here and indicates that while the two factor structure may *discriminate* between items with high cognitive components from those with more physical components the ordering by *difficulty* does not appear to follow the expected trajectory whereby cognitively demanding tasks such as using the telephone, taking medications and managing finances are lost first. It should be noted however that ‘shopping’ has a significant cognitive component and it requires a large range of physical and cognitive abilities. Furthermore this task is easily and conveniently taken over by other family members. Therefore the relative *difficulty* with this item could be a reflection of the ease with which the responsibility for the task can be reallocated. Again this refers to the limitations of assessing the ordering of functional decline in dementia with the use of a relatively crude scale.

The most *discriminatory* item in the full sample and late onset Alzheimer’s disease was ‘shopping’ whereas for the mixed group ‘telephone’ has a higher *discriminatory* value. The least *discriminatory* item in the full and late onset Alzheimer’s disease sample was ‘travelling’ and in the mixed group ‘medicine’ demonstrated the worst *discrimination*. These findings are generally consistent with the previous SDRIR analysis where ‘shopping’ was the most *discriminatory* item and ‘travelling’ was the least *discriminatory*. Although the current results demonstrate some slight differences in the both ordering by *difficulty* and

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

discrimination between the mixed hierarchy and the other two hierarchies a general trend across is that the items with the highest *discrimination* are those at either end of the *difficulty* range.

The insight into item *difficulty* and *discrimination* replicates findings from Chapter 9; that ‘shopping’ and ‘telephone’ are key items within the scale. The loss of independence in shopping and preparing food occurs very quickly at an early stage and likewise while the ultimate loss of ability to use the telephone occurs late in the course of dementia the ability is lost abruptly. As the scale conforms to a formal and invariantly ordered hierarchy whereby items ordered on the basis of mean scores are responded to in the same order for all respondents regardless of their level of functional ability the hierarchical structure can be applied to quickly infer a patient’s level of functional ability (for example, a patient reporting loss of independence with housework would most likely have already lost independence with travelling, finances, laundry, medications and shopping). This emphasises the clinical applications of Mokken scaling. As the items from an IIO hierarchy the results here outline the expected sequence of functional decline in dementia. While the Lawton IADL items appear to provide little information for differential diagnosis these items, particularly ‘shopping’ will have a significant contribution towards identifying specific levels of disability; the initial (shopping) and late (telephone) stages of functional and, by association, cognitive decline.

Some DIF was found for gender meaning different IIO hierarchies were established for men and women. While the ordering differed for many items the most significant differences in terms of mean scores and differences in placement on the hierarchy were for ‘laundry’, ‘housework’ and ‘food preparation’. These differences are clearly demonstrated in Figure 10.1. For each of these items the mean item score is lower for male participants. These lower scores could reflect the earlier point of overestimation of male functional impairment

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

due to tasks like 'laundry; and 'food preparation' representing gender-typed activities that men had historically less involvement with making them more likely to look for and receive assistance when they do attempt to do it (Fleishman, Spector & Altman, 2002). The items with the largest mean item differences are 'laundry' (0.51) and 'food preparation' (0.14).

The hierarchical pattern differs for men and women in this sample. For example, whereas loss of ability to take medication independently precedes the loss of ability to prepare food among women the reverse pattern is observed in men, and while the loss of independence in doing the laundry precedes the loss of financial independence for men, again the opposite pattern is seen in women. Using the telephone is the least *difficult* for men and women and is therefore most resistant to the effects of functional impairment and is unaffected by gender.

It is expected that some differences in IADLs between the sexes would be found. Traditionally, women have been more involved in domestic activities, which helps to account for the differences in performance for these tasks. For example, for some men, their wives or female partners may have taken a more active role in preparing meals and performing household chores. Therefore these men may have never taken an active role in these tasks which would therefore lead to the appearance of earlier loss and higher *difficulty* for these items. From this analysis it is evident that the Lawton IADL items function differently with different patterns of loss for men and women. Men generally performed worse despite having a higher mean ACE-R score reflecting better cognitive ability than the women of the sample. The results indicate that men are on an escalated course of functional decline. Whether this is due to more severe functional impairments occurring at an earlier stage or due to cultural or societal issues relating to the methods of assessing functional capacity should be explored further.

CHAPTER 10: HIERARCHICAL PATTERNS OF FUNCTIONAL DECLINE

Generally items were less *discriminatory* for men than women. For example, the ability to do housework was substantially less *discriminatory* for men than for women in this sample. These results indicate that the Lawton IADL items have a stronger association with functional decline for women compared to men. The poorer association of items to functional decline in men could be attributed to the unfamiliarity of some men to some of the tasks.

Gender DIF raises some concerns about the validity of combining male and female groups in the analysis of functional decline in dementia. While items with DIF can be retained in scales without the loss of measurement quality (Roznowski & Reith, 1999), scores from the Lawton IADL scale should be interpreted cautiously when comparing the functional performance of men and women. While a more extensive analysis of DIF is necessary these results demonstrate the potential of using Mokken scaling to examine how items may function differently in pertinent subgroups.

H^T for the full sample was 0.58 which is sufficient for establishing accuracy of item ordering within a scale. However, that DIF with respect to gender was present illustrates that while the IIO fit the data adequately it is important to examine relevant subgroups as well as looking at the overall sample to ensure that the item ordering holds up and if not to establish and interpret DIF. H^T values were greater once the sample was segregated by sex (male; $H^T=0.77$, female; $H^T=0.72$) which demonstrates the greater degree of accuracy of item ordering in these samples.

The combination of both male and female participants in the analysis of the full SDRIR sample may have had an influence on the IIO violation for 'food preparation'. DIF was demonstrated for 'food preparation' in this sample which would cause variability in the responses in the full sample, which would increase the probability of IIO violations. The IIO violations occurred only in two samples; the full dataset and late onset AD, also the two

largest samples with the most equal gender split. The smaller samples where 'food preparation' was retained have a greater proportion of male participants, which would shift the mean towards the male trend of poorer performance on this item. Lawton (1969) also reported difficulty regarding the fit of 'food preparation' to an ordered scale in his Guttman analysis by gender. All eight scale items met Guttman scaling criteria for the female sample whereas three items ('food preparation', 'laundry' and 'housework') did not meet the scaling criteria for the male sample.

The SDRIR sample analysed here consists largely of patients with Alzheimer's disease which limits the scope for identifying distinct patterns of loss. The only pure diagnostic groups with sufficient numbers for separate analysis were late onset Alzheimer's disease and mixed Alzheimer's disease and vascular dementia. That Alzheimer's disease was present in each of these samples restricts the ability to identify unique functional loss profiles. Furthermore it is possible that there could be some potential overlap between the late onset Alzheimer's and mixed groups. Autopsy studies from dementia clinics have found vascular pathology in 24% to 28% of patients diagnosed with Alzheimer's disease (Massoud et al., 1999; Gearing et al., 1995). This illustrates the difficulty in obtaining a complete discrimination of non-Alzheimer's disease dementia from Alzheimer's disease in life. It is possible that a proportion of the 477 patients diagnosed with late onset Alzheimer's disease could have coexisting pathologies, which could have contributed to the similar patterns identified between these groups.

It is possible that a more comprehensive analysis of several different diagnostic groups could identify different patterns of functional loss. Differences in the pattern of functional decline might have been revealed if there had been sufficient numbers for a strict separation of different types of dementia. A more in-depth analysis including a wider range of diagnostic groups with sufficient numbers to carry out a more comprehensive stratified

analysis could further elucidate the order of functional decline. The separate analysis of patients with bv-FTD where significant impairments in functional ability are commonly observed would be an interesting future study. A large sample with sufficient numbers for a stratified analysis (N~250) would be required. The functional decline in bvFTD may follow a different pattern than the ones observed here.

Variations in functional loss between the groups analysed here could have been identified using a different measure of assessing functional impairment. The Lawton IADL relies on self or informant reports of functional ability on eight tasks and chores rather than a demonstration of these tasks. This can lead to over or under-estimation of functional decline. While the Lawton IADL scale is the most commonly used IADL scale in dementia (Sikkes, De Lange-de Klerk, Pijnenburg & Scheltens, 2009) few studies have tested the psychometric properties of the scale. A more comprehensive measure, assessing a more varied range of abilities and including practical demonstrations might allow greater opportunity to discern differences between patient groups. For example, an analysis of the Assessment of Motor and Process Skills (AMPS, Fisher & Jones, 1999), an observational assessment of functional status, found no DIF between men and women (Merritt & Fisher, 2003). Administration of the AMPS involves a prior interview with each participant to ascertain which of the AMPS activities matches the participants' everyday functional routine. Tasks are selected from a choice of 50. Examples of AMSP tasks include making a salad, cleaning a bathroom, and weeding. One reason why there were no differences in performance by men and women is that all tasks measured are practiced and familiar to all being tested.

10.5 Conclusion

Mokken scaling analyses of instrumental activities of daily living scales in dementia provide valuable insight into the patterns of loss across diagnoses and gender. This study

demonstrated the invariant pattern of functional decline between a heterogeneous sample and four diagnostic subgroups. These findings support a consistent hierarchical pattern of impairment in instrumental activities of daily living across types of dementia. Although longitudinal data is required to ascertain a prior level of ability and to determine when functional abilities are actually lost the item parameters in this cross-sectional study suggest that the initial stages of functional decline can be observed in the loss of independence in shopping and food preparation, both of which appear to be lost quickly (due to the high *discrimination* levels associated with both items) at an early stage (reflected in the high *difficulty* of these items) across the types of dementia represented here. However without a measure of change it cannot be determined that the loss of these abilities is due to onset of dementia. Dependence using the telephone appears to represent the final stage of functional impairment assessed by the Lawton IADL scale for all groups and genders. These items at the extremes of the breadth of assessment show consistent high *discrimination*. If the results of this chapter are confirmed in similar populations it could be hypothesised that decline in IADL in individual's diagnosed with dementia, generally follow the same sequence of impairment as assessed by the Lawton IADL scale. This common hierarchy could be valuable in identifying current functional status in diverse patient populations and in predicting future impairments.

Gender differences were demonstrated in the pattern of decline in a mixed dementia sample. The DIF revealed between men and women indicates that whereas valid comparisons can be made between the diagnostic groups some caution should be used when comparing performance of male and female participants. Violations of IIO can also serve to indicate some degree of DIF, which may be responsible for the failure of items to conform to a consistent hierarchical pattern.

**Chapter 11: Hierarchical patterns of functional loss in Activities in Daily Living
and Instrumental Activities in Daily Living**

11.1 Introduction

While assessments of instrumental activities of daily living (IADL) are important in the assessment of early dementia many IADL show a floor effect as the severity of dementia increases because of the tendency to lose IADL abilities early in the course of dementia (Galasko et al., 2007). Basic activities of daily living (BADL) are useful in the advanced stages of disease as they are less closely underpinned by cognitive abilities and relate to day-to-day core necessary abilities. BADLs can provide valuable guidance in identifying when additional levels of support or long-term care placement are required (Desai, Grossberg & Sheth, 2004). The advantage of a measure assessing both BADL and IADL is that a wider range of levels of functional impairment can be assessed in one scale and increases test sensitivity (Desai, Grossberg & Sheth, 2004; Spector & Fleishman, 1998). The BADL-IADL measure can extend insight on the distribution of functional limitations in dementia. This information can be applied to improve identify initial decline and permit efficient allocation of resources and health care provisions as the level of dependency increases. The development of a hierarchical BADL-IADL scale extends the insight further as test performance can be used to develop a more accurate prognosis with regards to expected future decline and associated health and care needs.

In Chapters 9 and 10 the hierarchical properties of the Lawton IADL scale were assessed. This Chapter will focus on analyses of a BADL scale, the Physical Self-Maintenance Scale (PSMS) (Lawton & Brody, 1969), and will ultimately seek to combine items from the PSMS and Lawton IADL scale into one hierarchical BADL-IADL scale. The

CHAPTER 11: HIERARCHICAL PATTERNS OF DECLINE IN ADL AND IADL

PSMS assesses six basic activities of daily living (BADL) eating, dressing, grooming, physical ambulation, going to the toilet and bathing and is commonly used to assess functional impairment in dementia (Desai et al., 2004). In comparison to IADL tasks less is known about the cognitive determinants of the tasks commonly measured by BADL scales. IADL tasks involve complex and cognitively demanding activities whereas BADL scales include more fundamental and less complex tasks relating to personal care and wellbeing. The ability to perform these tasks independently is lost later than IADL (Barberger-Gateau, Fabrigoule, Helmer, Rouch & Dartigues, 1999). There is however a degree of overlap in terms of the range of abilities that these items assess (Spector & Fleishman, 1998).

Many BADL scales and measures have been devised and include self-report, informant reported and performance-based measures. Within each type there are two further forms of BADL scales; generic and disease specific measures. Generic scales, such as the Katz Index of Activities of Daily Living (Katz ADL; Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963) and the PSMS have wider applicability and are generally well validated and reliable. These scales permit the comparison of functional status across disorders and diagnoses. However, within dementia studies are required to determine item bias within these generic scales (Desai, Grossberg & Sheth, 2004). As these scales are designed to be widely used in many populations they may lack sensitivity to functional impairment caused by cognitive impairments. Disease-specific measures on the other hand, such as the Bristol Activities of Daily Living scale (Bucks, Ashworth, Wilcock & Siegfried, 1996) permit the assessment of functional loss due to cognitive impairments rather than a mixture of determinants of functional impairments such as physical impairment or psychiatric causes. Being more specific to dementia allows for more sensitive measurement (Spector, 1997).

Hierarchical BADL scales facilitate the identification of risk factors for further functional deterioration and their associated care needs. On an individual level the ability to

anticipate the disablement process in dementia could help both patients and carers make provisions, plan and anticipate the future level of assistance required. This also has wide reaching implications for society with hierarchical patterns of functional decline valuable in planning health care requirements and institutional assistance and public health interventions (Delva, et al., 2013). Investigating the item properties of BADL scales within different types of dementia samples and in relation to the properties of IADL items could help to further elucidate the cognitive underpinnings of these more basic activities of daily living. Determining distinctive IIO hierarchies of the BADL impairment process for different types of dementia would add insight into the patterns of functional decline associated with different types of dementia which could contribute towards differential diagnosis. This information could also be valuable in the development of dementia type specific scales.

Previous studies have investigated the hierarchical structure of BADLs by identifying the pattern of loss in dementia. Lechowski et al. (2010) identified a pattern of loss of functional decline assessed by the six items of the PSMS in a sample of 687 community-dwelling patients with mild-moderate Alzheimer's disease. This study determined a hierarchical pattern of decline from the binary item scores. This sequence of impairment started with an initial loss of the ability to walk out of the home, followed by grooming, bathing, dressing, going to the toilet and culminating in the inability to eat. Njegovan, Man-Son-Hing, Mitchell and Molnar (2001) found a natural hierarchy of functional loss in a community-dwelling elderly cohort (N=5874). This hierarchical decline differed from that identified by Lechowski et al. (2010). Here the decline began with bathing, which was followed by walking, going to the toilet, transferring, getting dressed, grooming and eating. This analysis was performed on the Older Americans Resources and Services (OARS, Fillenbaum, 1988) questionnaire which differs from the PSMS scale (for example, the OARS assesses as extra task; the ability to get in and out of bed and while the other items assessed

are similar there are three response levels of each item; can do without help, needs some help, completely unable) which could partly explain the different pattern exposed. Neither of the studies used IRT analyses to confirm IIO in the hierarchies established which could contribute to the different patterns identified. However the differences could also be attributed to the different samples analysed; the former hierarchy pertains to the decline in Alzheimer's disease while the latter was demonstrated in a heterogeneous population with regards to the disease affecting cognitive ability with low numbers of participants diagnosed with dementia (13.8%).

Delva et al. (2013) confirmed that four items of the Katz ADL scale formed a Guttman hierarchy in a French dementia cohort (N=838) with moderate dementia. Transferring and feeding followed primary difficulties in bathing and getting dressed. The Katz ADL scale examines functional performance in six activities; bathing, dressing, toileting, transferring, continence and feeding. Continence was not included in the analysis as the authors felt the item was better described as impairment rather than a disability and toileting was excluded, as it is closely associated with transferring. This Guttman scale of the four Katz ADL items analysed was established in item pairs with loss of independence occurring initially in bathing and/or dressing followed by transferring and/or eating with complete disability in both bathing and dressing, culminating in loss of independence in both transferring and eating. This hierarchical structure means that a patient unable to independently transfer or eat would always have already lost independence in bathing and getting dressed.

11.1.1 Combined hierarchical structure of BADL-IADL items

Previous research has explored the possibility of forming a hierarchical, cumulative scale to measure both BADLs and IADLs by applying IRT analyses (Kempen, Suurmeijer, 1990;

Spector & Fleishman, 1998; Kempen, Myers & Powell, 1995). Understanding the hierarchical relationship between items and sets of items can help to ensure that all items within the scale display sufficient levels of *discrimination* and that all items contribute to the assessment of functional ability by eliminating any redundant or idiosyncratic items.

As the activities assessed by IADLs are more complex and cognitively demanding these items may consistently be lost prior to the activities measured in BADL scales (Lawton & Brody, 1969). However an element of overlap between the two classifications of functional activities has been identified. Spector and Fleishman (1998) demonstrated a blurring between the BADL-IADL boundary in terms of item *difficulty*. For example, their results found the *difficulty* level for two IADLs ('getting around outside' and 'food preparation') lay between that of two BADLs ('bathing' and 'dressing'). This hierarchical pattern of decline suggests that bathing was a more *difficult* item than 'getting around outside' and 'food preparation' suggesting that this hierarchical nature between IADL and BADLs is less consistent than previously thought. Overlap in terms of *difficulty* was also reported by Spector, Katz, Murphy & Fulton (1987), Suurmeijer et al. (1994) and Kempen et al. (1995). These results indicate that while BADL and IADL items may form a hierarchical scale the pattern does not necessarily demonstrate a clear cut separation of basic and instrumental ADLs. However these studies were based on BADL and IADL responses from elderly community-dwelling populations without dementia. Further research is required to determine if the BADL IADL combined scale can be extended to the pattern of functional impairment in dementia.

In this chapter I aim to assess the hierarchical structure of BADL activities as assessed by the PSMS in a sample of people with dementia. Differential item functioning by diagnosis and gender will also be assessed in a stratified analysis. The item properties of both BADL (PSMS) and IADL items (Lawton IADL) will be examined in relation to each other to

determine the hierarchical structure across both scales. Results will be presented individually in the following order (i) hierarchical analysis of the PSMS items, (ii) differential item functioning of the PSMS items by diagnosis and gender, and (iii) the combined hierarchical structure of both PSMS and Lawton IADL scale items. These individual analyses will then be summarised and compared.

11.2 Method

11.2.1 Participants

Complete PSMS data for three samples from the SDRIR were analysed: (i) complete sample (N=873; late onset AD, N=502; mixed AD VaD, N=146; VaD, N=99; early onset AD, N=77; DLB, N=24; FTD, N=15; PDD, N=10); (ii) late onset AD (N=502); (iii) mixed AD & VaD (N=146). An analysis was also performed on the whole sample stratified by sex. To determine the combined hierarchical structure of the items of the Lawton IADL and PSMS a sample of patients with complete itemised IADL and BADL data (N=822; late onset AD, N=475; mixed AD VaD, N=138; VaD, N=98; early onset AD, N=69; DLB, N=20; FTD, N=12; PDD, N=10) was isolated for a combined analysis. This sample was drawn from the complete sample above (N=873). From this sample the 822 participants with fully itemised data for both PSMS and Lawton IADL scales were included in this additional analysis. Demographic and cognitive information for each group is presented in Table 11.1

Table 11.1 Characteristics of the SDRIR samples analysed

	N	Sex (% male)	Mean age	Mean ACER	Mean PSMS
Complete sample	873	465 (53.3%)	77.4	60.3	4.4
Male	465		76.8	62.0	4.3
Female	408		78.0	58.4	4.4
Late onset AD	502	262 (52.2%)	79.3	60.6	4.5
Mixed AD VaD	146	81 (55.5%)	78.8	62.5	4.4
Combined BADL and IADL data	822	440 (53.5%)	77.5	60.4	3.9

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised, PSMS=Physical Self-Maintenance Scale, AD=Alzheimer's disease, VaD=vascular dementia, BADL= Basic Activities of Daily Living, IADL=Instrumental Activities of Daily Living.

11.2.2 Measures

The Physical Self-maintenance Scale (PSMS, Lawton & Brody, 1969) was devised to assess basic activities of daily living. This informant-based scale was adapted from an already existing functional scale (Lowenthal, 1964). The PSMS retained the original six basic functioning items measuring behaviours of toileting, feeding, dressing, grooming, physical ambulation and bathing, broadening the content of some to ensure the item was applicable to both community-dwelling individuals and residential care patients. The scoring was adapted so that each item was assessed out of five levels of ability (for example the response to 'feeding' is one of five levels of ability: (i) eats without assistance, (ii) eats with minor assistance at meal times and/or with special preparation of food, or help in cleaning up after, (iii) feeds self with moderate assistance and is untidy, (iv) requires extensive assistance for all meals and (v) does not feed self at all and resists efforts of others to feed him). From these levels the ability closest to the individual's current ability is selected by the best available source (the patient, family or other informant). The scale is scored dichotomously; only one of the five levels available is associated with a score of 1, all other levels where some assistance is required result in a score of 0. See Appendix D for full scale. Guttman scaling

criteria are met for the PSMS items (Lawton & Brody, 1969). The PSMS has shown good inter-rater reliability and can be used in clinical and research settings (Hokoishi et al., 2001). As the activities assessed measure the lower-end of self-care ability the PSMS is more useful in the advanced stages of dementia. PSMS items may not be sensitive to changes in functional ability at baseline when levels of functional independence are higher.

Mean PSMS item scores for each sample are presented in Table 11.2. From the five levels of ability for each item PSMS items are scored dichotomously (0 or 1) with lower mean scores indicating poor functional ability. Items are summed to provide a total score out of six.

11.2.3 Mokken scaling analysis

All Mokken scaling analyses were carried out using the ‘mokken’ package in R. The fit of the six scale items to the monotone homogeneity model and IIO was examined in each of the six samples.

11.3 Results

Mean PSMS item scores for each sample are presented in Table 11.2 in order of mean score for the complete sample (N=873), from most to least *difficult*, with lower mean scores indicating poor functional ability. Mean PSMS and Lawton IADL scores for the combined analysis is presented in Table 11.3. Items from each scale are presented together in order of decreasing *difficulty*.

Table 11.2 Mean PSMS item scores for SDRIR sample plus four subgroups

Complete sample		Late onset AD		Mixed AD VaD		Male		Female	
Item	Mean	Item	Mean	Item	Mean	Items	Mean	Items	Mean
Physical ambulation	0.60	Physical ambulation	0.63	Physical ambulation	0.71	Physical ambulation	0.62	Physical ambulation	0.57
Dressing	0.67	Dressing	0.70	Dressing	0.66	Dressing	0.60	Dressing	0.73
Grooming	0.68	Grooming	0.71	Grooming	0.70	Grooming	0.65	Grooming	0.71
Bathing	0.71	Bathing	0.72	Bathing	0.71	Bathing	0.74	Bathing	0.67
Toilet	0.82	Toilet	0.85	Toilet	0.79	Toilet	0.82	Toilet	0.82
Feeding	0.91	Feeding	0.92	Feeding	0.92	Feeding	0.89	Feeding	0.92

Note. PSMS=Physical Self-Maintenance Scale, AD=Alzheimer’s disease. VaD=vascular dementia. Item mean range for each item is 0-1 (0 indicates no impairment and 1 indicates impairment).

Table 11.3 Mean PSMS and Lawton IADL scores for combined BADL and IADL sample (N=822). Items presented from most to least *difficult*, with lower mean scores reflecting poor functional ability.

Scale	Item	Mean
Lawton IADL	Shopping	0.18
Lawton IADL	Food Preparation	0.19
Lawton IADL	Medication	0.24
Lawton IADL	Laundry	0.44
PSMS	Physical ambulation	0.59
Lawton IADL	Finances	0.62
Lawton IADL	Travelling	0.64
PSMS	Dressing	0.66
PSMS	Grooming	0.67
PSMS	Bathing	0.70
Lawton IADL	Housework	0.75
PSMS	Toilet	0.83
Lawton IADL	Telephone	0.85
PSMS	Feeding	0.91

Note. PSMS=Physical Self Maintenance Scale, IADL=Instrumental Activities of Daily Living.

11.3.1 Mokken scaling analyses of diagnostic groups

(i) Data as a whole (N=873)

Complete itemised PSMS data for the complete SDRIR sample were analysed. This mixed sample comprised patients with several dementia diagnoses (late onset AD n=502, Mixed AD VaD n=146, VaD n=99, early onset AD n=77, DLB n=24, FTD n=13, PDD n=12). Mean scores for the sample are presented in Table 11.2.

An assessment of unidimensionality of the PSMS scale in this sample showed that all items formed a singular cluster using the automated item selection procedure (AISP) function. All item-pair scalability coefficients were non-negative and item scalability coefficients were greater than the 0.3 lower bound threshold level. There were no exclusions from the assessment of monotonicity. The six items formed a strong Mokken scale ($H=0.59$).

The six items failed to form an IIO hierarchy with violations in item ordering for ‘physical’ and ‘bathing’. The removal of these items resulted in confirming IIO in the remaining items however reducing the scale to just four items (‘toilet’, ‘feeding’, ‘dressing’, ‘grooming’) may cause a reduction in the reliability of the sum scores.

(ii) Late onset AD (N=502)

Complete itemised PSMS data for SDRIR participants diagnosed with late onset AD were analysed. Mean scores for the sample are presented in Table 11.2. The six items met the assumptions of the monotone homogeneity model. More specifically, all items formed a singular cluster using the automated item selection procedure and item-pair scalability coefficients were nonnegative with all item scalability coefficients greater than the 0.3 lower bound threshold level. There were no exclusions from the assessment of monotonicity. Therefore the six items formed a strong Mokken scale ($H=0.59$).

The items surpassed the minimum criteria to confirm invariant item ordering. There were no exclusions in due to violations of item ordering and H^T was greater than 0.3. The six items (Table 11.4) formed a reliable (MS=0.84) and strong hierarchical Mokken scale ($H=0.59$) with $H^T=0.33$.

Table 11.4 IIO hierarchy items from late onset AD sample listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Physical	0.63	Feeding	0.69
Dressing	0.70	Dressing	0.60
Grooming	0.71	Toilet	0.60
Bathing	0.72	Bathing	0.58
Toilet	0.85	Grooming	0.58
Feeding	0.92	Physical	0.55

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

(iii) Mixed AD VaD (N=146)

Complete itemised PSMS data for SDRIR participants diagnosed with mixed AD VaD were analysed. Mean scores for the sample are presented in Table 11.2. Again for this sample items met the necessary requirements for establishing fit to the monotone homogeneity model. All items formed one cluster using the automated item selection procedure. All item scalability coefficients were greater than 0.3 and item-pair scalability coefficients were nonnegative. There were no exclusions due to monotonicity. The six item scale formed a strong Mokken scale ($H=0.54$).

One item ('toilet') was removed in the assessment of IIO. The remaining 5 items (Table 11.5) formed a reliable ($MS=0.77$) and strong hierarchical Mokken scale ($H=0.54$) with $H^T=0.41$.

Table 11.5 IIO hierarchy items from Mixed AD VaD sample listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Physical	0.55	Bathing	0.57
Dressing	0.66	Feeding	0.57
Grooming	0.70	Grooming	0.55
Bathing	0.71	Dressing	0.54
Feeding	0.92	Physical	0.48

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

11.3.2 Mokken scaling analyses by gender

(i) Gender analysis: Male (N=465)

Complete itemised PSMS data for male SDRIR participants (N=465) were analysed. Mean scores for the sample are presented in Table 11.2.

The six items met the assumptions of the monotone homogeneity model. More specifically, all items formed a singular cluster using the automated item selection procedure and item-pair scalability coefficients were nonnegative with all item scalability coefficients greater than the 0.3 lower bound threshold level. There were no exclusions from the assessment of monotonicity. Therefore the six items formed a strong Mokken scale ($H=0.60$).

Five of the six items surpassed the minimum criteria to confirm invariant item ordering. 'Physical ambulation' was excluded due to violations of item ordering. The five items (Table 11.6) formed a reliable ($MS=0.84$) and strong hierarchical Mokken scale ($H=0.66$) with $H^T=0.42$.

Table 11.6 IIO hierarchy items from sample of male SDRIR participants listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Dressing	0.60	Dressing	0.74
Grooming	0.65	Grooming	0.72
Bathing	0.74	Toilet	0.62
Toilet	0.82	Feeding	0.60
Feeding	0.89	Bathing	0.59

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

(ii) Gender analysis: Female (N=408)

Complete itemised PSMS data for female SDRIR participants (N=408) were analysed. Mean scores for the sample are presented in Table 11.2. The six items met the assumptions of the monotone homogeneity model. More specifically, all items formed a singular cluster using the automated item selection procedure and item-pair scalability coefficients were nonnegative with all item scalability coefficients greater than the 0.3 lower bound threshold

level. There were no exclusions from the assessment of monotonicity. Therefore the six items formed a strong Mokken scale ($H=0.62$).

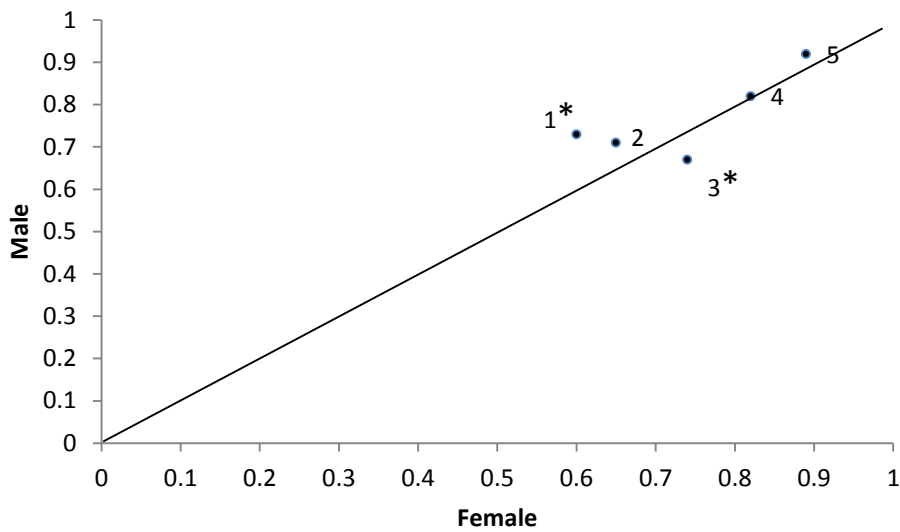
The items surpassed the minimum criteria to confirm invariant item ordering. There were no exclusions in due to violations of item ordering and H^T was greater than 0.3. The six items (Table 11.7) formed a reliable ($MS=0.84$) and strong hierarchical Mokken scale ($H=0.62$) with $H^T=0.39$. The IIO hierarchies for SDRIR PSMS samples are presented in Table 11.8. Figure 11.1 demonstrates the gender DIF for PSMS items.

Table 11.7 IIO hierarchy items from sample of female SDRIR participants listed from most to least *difficult* and most to least *discriminatory*

Item	Mean	Item	H_i
Physical	0.57	Feeding	0.69
Bathing	0.67	Bathing	0.66
Grooming	0.71	Dressing	0.65
Dressing	0.73	Physical	0.61
Toilet	0.82	Grooming	0.60
Feeding	0.92	Toilet	0.57

Note. H_i =item scalability coefficient (item *discrimination*) with higher values indicating greater *discrimination*. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

Figure 11.1 Mean item scores for male and female participants.



Note. 1=dressing, 2=grooming, 3=bathing, 4=toilet, 5=feeding. Mean item scores for females are on the x axis and for male, on the y axis. Low mean item scores designate the items indicating severe functional impairment, and high mean item scores designate the items indicating minor functional impairment. The numbers near the data points refer to the item numbers. An identity line is drawn through the origin with a slope of 1. * = items with major deviations from the identity line demonstrating DIF

Table 11.8 PSMS IIO hierarchical comparisons across groups; items listed from most to least *difficult*

Late onset AD		Mixed AD VaD		Male		Female	
Item	Mean	Item	Mean	Item	Mean	Item	Mean
Physical ambulation	0.63	Physical ambulation	0.55	Dressing	0.60	Physical	0.57
Dressing	0.70	Dressing	0.66	Grooming	0.65	Bathing	0.67
Grooming	0.71	Grooming	0.70	Bathing	0.74	Grooming	0.71
Bathing	0.72	Bathing	0.71	Toilet	0.82	Dressing	0.73
Toilet	0.85	Feeding	0.92	Feeding	0.89	Toilet	0.82
Feeding	0.92					Feeding	0.92

Note. AD=Alzheimer’s disease, VaD=vascular dementia, mean item scores (range from 0-1) with higher scores indicating lower *difficulty*. Items in bold show differential item functioning

11.3.3 Combined BADL-IADL analysis

To assess the combined hierarchical structure of BADL and IADLs complete itemised PSMS and Lawton IADL data for SDRIR participants (N=822) were analysed. This heterogeneous sample included a variety of dementia types (late onset AD, n=475, mixed AD/VaD, n=138, VaD, n=98, young onset AD, n=69, DLB, n=20, FTD, n=12, PDD, n=10). Cognitive and demographic information is presented in Table 11.1 and mean PSMS and Lawton IADL item scores for the sample are presented in Table 11.3.

The 14 items met the assumptions of the monotone homogeneity model. More specifically, all items formed a singular cluster using the automated item selection procedure and item-pair scalability coefficients were nonnegative with all item scalability coefficients greater than the 0.3 lower bound threshold level. There were no exclusions from the assessment of monotonicity. Therefore the 14 items formed a strong Mokken scale ($H=0.53$).

Six items, five IADL items, one PSMS, failed to meet the minimum criteria to confirm invariant item ordering. 'Dressing', 'housework', 'travelling', 'finance', 'telephone' and 'food preparation' were excluded due to violations of item ordering. The eight remaining items (Table 11.9) formed a reliable ($MS=0.79$) and strong hierarchical Mokken scale ($H=0.57$) with $H^T=0.64$.

Results from this combined analysis demonstrate that the items from both PSMS and Lawton IADL scales meeting IIO criteria conform to the typical ordering expected where Lawton IADL items have lower mean scores demonstrating that patients in this sample perform worse on these IADLs ('laundry', 'medicine' and 'shopping') than the PSMS BADL

items ('feeding', 'toilet', 'bathing', 'grooming' and 'physical'). These results support the hypothesised pattern of functional decline in dementia where independence is first lost with IADL followed by BADL.

Table 11.9 Combined BADL IADL IIO hierarchy items listed from most to least *difficult* and most to least *discriminatory*

Scale	Item	Mean	Scale	Item	H_i
Lawton IADL	Shopping	0.18	PSMS	Feeding	0.65
Lawton IADL	Medicine	0.24	PSMS	Toilet	0.62
Lawton IADL	Laundry	0.44	PSMS	Bathing	0.61
PSMS	Physical ambulation	0.59	Lawton IADL	Shopping	0.60
PSMS	Grooming	0.67	PSMS	Grooming	0.60
PSMS	Bathing	0.70	Lawton IADL	Medicine	0.56
PSMS	Toilet	0.83	PSMS	Physical	0.52
PSMS	Feeding	0.91	Lawton IADL	Laundry	0.46

Note. BADL=Basic Activities of Daily Living, IADL=Instrumental Activities of Daily Living, H_i =item scalability coefficient. Mean item scores (range from 0-1) with higher scores indicating lower *difficulty*.

11.4 Discussion

This chapter sought to determine whether the activities of daily living measured by the PSMS followed a common hierarchy of functional decline and whether this hierarchy was consistent across different types of dementia and gender. A further aim of this chapter was to examine item properties of both PSMS and Lawton IADL items in relation to each other and to determine the hierarchical structure of this combined scale. Establishing a formal and invariantly ordered hierarchy is valuable to clinicians and researchers. It provides prognostic value and due to its hierarchical nature items within such scales can be used to gain a quick insight into a patient's functional status. Additional value can be gained from investigating whether the hierarchical patterns are invariant across subgroups of patients groups.

The initial analysis of the full SDRIR PSMS data failed to retain a sufficient number of items in an invariantly ordered hierarchy. All six items met the assumptions of the

monotone homogeneity model. This means that while the summed score of the PSMS can be used to quickly indicate a patient's functional status the individual items are not invariantly ordered across the sample's respondents, precluding the application of specific items in isolation to gauge a patient's functional ability level.

In the comparison of the IIO pattern from the late onset AD and mixed AD VaD samples the item ordering did not show any variation. This consistent pattern of decline starts with the loss of independence in physical ambulation, which is followed by dressing, grooming, bathing and finally feeding disabilities. This pattern bears some similarity to the results from Lechoswki et al. (2010) with the same items at either ends of the *difficulty* range. However there are inconsistencies in the ordering of the items in-between these extremes. Importantly, the hierarchy of decline established by Lechowski et al. (2010) is not invariantly ordered, as it was not established formally using item response theory method, which allow for the assessment of IIO. Many studies of the hierarchical structure of BADL scales have failed to supply sufficient or accurate evidence supporting these hierarchies (Sijtsma, Meijer & van der Ark, 2011). It is often incorrectly assumed that any IRT procedure can establish IIO (Ligtvoet et al., 2010). Therefore it is not surprising that differences would be reported in studies failing to correctly test for IIO.

Differential item functioning (DIF) by gender was demonstrated. This reveals that for some items of the PSMS the likelihood of equal responses from men and women is not equal. These results indicate that some caution should be used when comparing scores from men and women. While the pattern of loss for women differed the pattern of decline found by Lechowski et al. (2010) for men is the same as the present findings. This pattern of loss for men in items common to both studies was: initial inability to dress, then to groom, bathe, to go to the toilet and finally to eat.

The DIF revealed for BADL items was not as substantial as the DIF demonstrated for IADL items in Chapter 10. The accuracy of item ordering in the gender specific hierarchies did not increase in the BADL gender stratified analysis suggesting as in the case with the gender specific analyses in Chapter 10 suggesting the gender DIF may have less influence on BADLs than IADLs. However the values of H^T were much lower across all analyses of BADLs than IADLs indicating that the IIO hierarchies are of somewhat low accuracy. These findings indicate that while a consistent pattern of BADL loss emerges it may be less consistent than the deterioration of IADL skills. Furthermore, as Mokken scaling analysis is stochastic in nature there are at least a small number of respondents for whom the pattern of functional impairment differs from the hierarchical pattern specified in the IIO ordering. This is due to error and should be considered when using the response to one IIO item to form the basis of a patient's functional status, particularly where low values of H trans are concerned. Inaccurate assessments can be made by assuming that all patients who do not report functional limitations in a relatively *difficult* task will be able to perform the other less *difficult* tasks assessed. However, establishing a hierarchical pattern common to most within the sample can be valuable in identifying those patients with atypical patterns of responses.

The poorer precision in item ordering could be due to the cognitive ability of the sample. The study was predominantly limited to patients with mild-moderate dementia. Patients lose independence in IADL functioning whereas BADLs show preserved functioning until late in the course of dementia (Spector et al., 1987). The sample and range of ability may be less suited to the sensitive and accurate assessment of the order of BADL decline. Further analysis in a sample with greater dementia severity may show stronger IIO with fewer violations. Additionally, the sample size of the mixed AD VaD analysis in this Chapter is relatively small (N=142). Replication studies of larger samples are required with the present results from this analysis interpreted with caution. Future studies investigating the

hierarchical functional decline in a wider range of dementia types may identify differences in this pattern. In a comparison with progressive nonfluent aphasia, semantic dementia and Alzheimer's disease Mioshi et al. (2007) identified a unique pattern of BADL decline for bv-FTD where poor initiation, followed by poor planning led to poor execution of tasks.

The analysis of the hierarchical structure of the combined IADL and BADL items resulted in a formal IIO hierarchy of eight of the original items (five BADLs, three IADLs). The pattern of loss across both kinds of functional activities followed the expected item ordering with IADLs lost first followed by BADLs. Previous studies found some variation in item *difficulty* between BADLs and IADLs. For example, in a combination of 16 BADL and IADL items the ability to use the telephone was the third most *difficult* with several BADL items less *difficult* and several IADL items were less *difficult* than 'bathing' (Spector & Fleishman, 1998). While the items analysed here forming the IIO hierarchy demonstrated a clear separation in terms of *difficulty* when the mean scores from all items are considered there is considerable overlap; the ability to use the telephone is second least *difficult* to the ability to feed and maintaining independence in housework is also among the least *difficult* items. The most *difficult* BADL-physical ambulation-is more *difficult* than four IADL items (telephone, housework, travelling and finance). This finding supports previous results suggesting that telephone use is less *difficult* than many BADLs. This item could be added to the usual group of BADL items to assess severe levels of impairment (Spector & Fleishman, 1998). However it is of note that the items crossing the BADL IADL boundary ('physical ambulation', 'finances', 'travelling', 'housework', 'telephone') do not meet IIO criteria which indicates a greater degree of variability in item responses to these items which may warrant further investigation. The variability in these responses may be driven by gender DIF in the case of 'finances' and 'housework'. Further stratified analyses by age, disease severity, diagnosis and gender could reveal a more inclusive combined BADL/IADL scale.

11.5 Conclusion

An assessment of the invariant item ordering of BADL items in the PSMS showed a similar pattern of impairment in different diagnostic groups. While this study did not uncover specific items that could assist in differential diagnoses or in the development of disease specific scales it is possible that a sample representing a wider range of disease severity, particularly in the more severe ranges, with sufficient numbers for a purer stratified analysis by diagnosis could reveal more information. For example, analyses of data from patients diagnosed with bv-FTD could demonstrate a deviation from the general hierarchy of decline found here.

As with the results of the previous chapters the cross-sectional analyses applied here restrict the interpretations of these results. Longitudinal studies examining the evolving functional disablement process and associated performance on BADL and IADL scales are required to confirm the hierarchical patterns of decline demonstrated from Mokken scaling analyses.

Chapter 12: Discussion

The aim of this thesis was to test the suitability of IRT analysis in dementia assessment methods. In doing so this research attempted to identify applications of IRT methods to dementia assessment measures and applied one of these methods—the nonparametric Mokken scaling analysis—to examine the psychometric properties of commonly used measures of cognitive and functional assessment in dementia.

This thesis explored the item parameters and hierarchical properties of scales assessing current (the ACE-R and ACE-III) and premorbid cognitive ability (the NART). The use of Mokken scaling methods as a data driven scale development technique was also examined and applied to derive new brief scales for the assessment of current cognitive ability in dementia as well as for the estimation of premorbid intelligence. The estimation of patients' level of cognitive ability prior to disease onset is crucial to ascertain the level of cognitive decline caused by dementia. As functional outcomes also form a crucial component of the dementia diagnosis and assessment I also assessed the item properties of two widely used functional impairment scales, the Lawton IADL and PSMS scales.

12.1 Systematic Review

To appreciate the literature and the previous applications of item response theory analyses to scales assessing cognitive impairment in dementia better, a systematic review was undertaken. This review provided an insight into the current state of research and applications of IRT analyses in dementia.

Due to the limited number of included studies the review was relatively ineffective in establishing a common hierarchy of cognitive decline in dementia. Of the four studies included in the review only three confirmed a hierarchy of decline and of these two

hierarchies were based on the MMSE (Ashford et al., 1989; Gibbons et al., 2002) while the third confirmed a hierarchy of decline for a new ‘Global Function’ scale (Mungas & Reed, 2000). If a common hierarchy was established, this could be valuable in research and importantly in clinical practice as it would enable clinicians to establish, with relative ease, a patient’s degree of disease severity, and ultimately his or her risk for future cognitive decline.

The systematic review was on a narrow topic. Most of the studies identified examined the cognitive decline in Alzheimer’s disease. There is very little literature on Mokken scaling in dementia particularly on a wider range of dementia syndromes. This gap in the literature highlights the need for more studies applying these methods to commonly used tests in dementia and in a range of dementias. This will be helpful because it may lead to the identification of unique and distinctive clinical profiles of cognitive decline. Furthermore these studies could reveal key items with good item parameters for the accurate measurement of various types of dementia and those with poor contribution to measurement of cognitive decline. Therefore this thesis performed a series of studies to extend this literature and add to the hierarchical pattern of cognitive decline in dementia.

12.2 Hierarchical ordering by difficulty across ACE analyses

The pattern of cognitive decline as assessed by the ACE-R and ACE-III was examined through Mokken scaling analyses. The results of these analyses in terms of the hierarchical patterns of item *difficulty* revealed provided some valuable theoretical and clinically practical insights.

Table 12.1 presents eight hierarchies of item *difficulty* from the analyses of the ACE-R and ACE-III scales performed in this thesis. These samples have several important differences in terms of diagnostic representation, age, disease severity and size of sample. Therefore items consistently meeting the strict criteria for inclusion in an invariantly ordered

CHAPTER 12: DISCUSSION

item hierarchy across these different samples can be considered as psychometrically strong items for the assessment of dementia. Three items; ‘memory retrograde’, ‘recognition’ and ‘write a sentence’, each appear in seven of the eight hierarchies. Where all three of these items appear in a hierarchy together they generally follow the same ordering of *difficulty* with respondents across samples finding ‘memory retrograde’ the most *difficult*, followed by ‘recognition’ and ‘write a sentence’ as the least *difficult*. An exception to this ordering is the pattern in the Mini-ACE development sample in Chapter 5 where ‘recognition’ was the least *difficult* item. However this hierarchy is based on an analysis of ACE-III in which ‘write sentence’ is ‘write two sentences’. Intuitively it would seem that writing two sentences would be more *difficult* than writing one sentence. This change in item content is most likely responsible for the alteration in item *difficulty* for this sample.

Therefore while these items appear to have good psychometric properties and are closely related to the assessment of cognitive decline the ordering is too general to be of any use for differential diagnosis. These items are consistently in the same order despite the differences in type of dementia and severity of the disease.

Inspection of the hierarchical patterns across samples did identify some items, which may have applications in supporting diagnosis. For example, where present, ‘name and address recall’ is generally the most *difficult* item as is the case for the hierarchies for the AD type sample, the Mini-ACE development sample (Chapter 5), and the late-onset AD sample (Chapter 6). The only exception to this is in the case of the other frontotemporal lobe degenerative disorders sample in Chapter 4 where ‘name and address recall’ is the third most *difficult* item in the hierarchy after ‘reading’ and ‘verbal fluency-animal’. ‘Three item recall’ is also less *difficult* for this sample than the other samples. This sample comprises two patients groups; semantic dementia and progressive nonfluent aphasia. The pattern of decline from reading, fluency to memory for this sample therefore supports the characterisation of the

CHAPTER 12: DISCUSSION

impairments associated with semantic dementia as the loss of conceptual knowledge in the context of relatively intact recent episodic memory (Adlam et al., 2009; Mion et al., 2010) and PNFA as an expressive language disorder with word retrieval difficulties (Carthery-Goulart, Knibb, Patterson & Hodges, 2012).

While we may consider the items partitioned into hierarchical Mokken scales as strong items these analyses also reveal several weak items by the same standards (i.e. items failing to conform to any of the hierarchical Mokken scales). Neither ‘follow written command-close eyes’, ‘repetition of single multi-syllabic words’ nor ‘draw overlapping infinity loops’ met criteria for inclusion in an IIO hierarchy in any of these analyses. Additionally, ‘syntactical comprehension’ and ‘draw a cube’ appear only once, in the AD type and the mixed AD VaD samples respectively and ‘count dot arrays’ appears only twice in the full mixed sample and the late onset AD samples in Chapter 6. However ‘draw overlapping infinity loops’, as an ACE-III item, is only relevant to the Mini-ACE development sample analysis. Therefore the failure of ‘draw overlapping infinity loops’ to conform to one hierarchy is not sufficient to determine the quality of the item.

‘Syntactical comprehension’ and ‘follow written command-close eyes’ have previously been found to demonstrate poor sensitivity to the assessment of cognitive impairment (Brugnolo et al., 2009). This was the reasoning behind the alterations and removal of ACE-R items in the development of the ACE-III. ‘Syntactical comprehension’ in the ACE-R assesses the ability of the subject to follow a three stage command (‘take the paper in your right hand, fold the paper in half, and put the paper on the floor’). Due to the poor sensitivity of this item the *difficulty* level was increased in the ACE-III version of ‘syntactical comprehension’ by assessing the ability to follow three single-step commands of a greater syntactical complexity (‘place the paper on top of the pencil’, ‘pick up the pencil but

not the paper' and 'pass me the pencil after touching the paper'). The assessment of the ability to 'follow a written command-close eyes' was removed outright.

12.3 Patterns of item difficulty between samples of patients with Alzheimer's disease

Comparing the hierarchical patterns of the analysis of the 'AD type' sample in Chapter 4 and the late-onset AD sample in Chapter 6 reveals a generally consistent pattern of cognitive decline. There are only five items common to both hierarchical scales however which limits the degree of comparability between the two patterns. Examining these five items demonstrates a parallel pattern of decline which starts with impairment in episodic memory ('name and address recall') and is followed by deficits in retrograde memory, memory recognition, temporal orientation with the ability to write a sentence the least *difficult* of these five common items within the two hierarchies. These commonalities suggest that this trajectory of impairment may be common to all patients diagnosed with AD. Further analyses of larger samples matched for disease severity and age need to be carried out to investigate this further.

There were insufficient numbers of patients diagnosed with early-onset AD in Chapter 6 for separate analysis. Therefore these patients were added to the late-onset AD sample to create a combined early and late-onset AD sample. This provided another sample of AD, which could lend further support to the theory of a disease specific pattern of decline, should the IIO hierarchy reveal a similar pattern. This extra sample also enabled the examination of the pattern of decline in early-onset AD. Although the proportion of early-onset AD patients within the sample was low (12.6%) it is possible that should potential differences in the sequence of cognitive decline between early and late onset be large enough, the IIO hierarchy may reveal valuable insights.

CHAPTER 12: DISCUSSION

Table 12.1 Comparison of all ACE-R and ACE-III IIO hierarchical Mokken scales. Items listed from most to least *difficult*.

Alzheimer's type	Chapter 4		Chapter 5		Chapter 6		
	Predominantly frontal dementia	Other frontotemporal lobe degenerative disorders	Mini-ACE development sample	Full SDRIR sample	Late-onset AD	Late and early-onset AD	Mixed AD VaD
Name and address recall	3 item recall	Reading	Name and address recall	Recognition	Name and address recall	Memory retrograde	3 item recall
Retrograde	Retrograde	Fluency-animal	Fluency-animal	Draw intersecting pentagons	Memory retrograde	Recognition	Fluency-animal
Draw a clock	Semantic comprehension	Name and address recall	Fluency-letter	Semantic comprehension	Recognition	Draw intersecting pentagons	Memory retrograde
Recognition	Recognition	Fluency-letter	3 item recall	Naming 2	Draw intersecting pentagons	Draw a clock	Draw a cube
Orientation in time	Naming (10 items)	Retrograde	Memory retrograde	Count dot arrays	Name and address learning	Name and address learning	Fluency-letter
Serial sevens	Orientation in time	3 item recall	Write two sentences	Write a sentence	Orientation in time	Repetition 3	Orientation in time
Syntactical comprehension	Draw a clock	Recognition	Serial sevens	Reading	Repetition 3	Serial sevens	Draw intersecting pentagons
Orientation in geography	Serial sevens	Name and address learning	Draw a clock	Orientation in geography	Semantic comprehension	Naming 2	Draw a clock
Write a sentence	Orientation in geography	Write a sentence	Semantic comprehension	Repetition 2	Naming 2	Write a sentence	Name and address learning
Naming (pencil and watch)	Name and address learning	Draw a clock	Naming*	Identify fragmented letters	Count dot arrays	Reading	Serial sevens
3 item registration	Count dot arrays	Orientation in geography	Repetition-3*	Naming 1	Reading		Naming 2
	Naming (pencil and watch)	Serial sevens	Recognition		Write a sentence		Write a sentence
		Orientation in time	Orientation in time		Repetition 2		Orientation in geography
		3 item registration	Orientation in geography		Identify fragmented letters		
			Name and address learning				
			Repetition-2*				
			Identify fragmented letters				

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised, ACE-III=Addenbrooke's Cognitive Examination-III, IIO=invariant item ordering, SDRIR=Scottish Dementia Research Interest Register, AD=Alzheimer's disease, VaD=Vascular dementia, Mini-ACE=Mini-Addenbrooke's Cognitive Examination. *Items differ in ACE-III: Repetition 2 in ACE-III=all that glitters is not gold, Repetition 3 in ACE-III=A stitch in time saves nine. Naming in ACE-III assess ability to name 12 pictures.

CHAPTER 12: DISCUSSION

The hierarchies from both analyses consisted of eight items common to the two samples. All eight items follow the same ordering of *difficulty* for both the late onset AD and combined early and late onset AD according to the mean item scores. The common items and common pattern of item *difficulty* from greatest *difficulty* with ‘memory retrograde’ followed by ‘recognition’, ‘draw intersecting pentagons’, ‘name and address learning’, ‘repetition 3’, ‘naming 2’ ‘reading’ and least *difficulty* with ‘write a sentence’ is most likely due to the preponderance of the same data in both samples. It is very likely, of course, that analysing two samples with 87% overlap would produce similar results. However, what is interesting to note is the number of items retained in each hierarchy; 14 in the more homogenous sample of late onset AD and 10 items in the larger mixed early and late onset AD sample. This suggests some greater extent of variation of item responses, which could be the result of different cognitive profiles between the early and late onset groups. This hypothesis needs to be confirmed by the analysis of sufficiently large sample of patients for separate analysis.

Comparing the patterns of item *difficulty* between all three samples of patients with AD; AD type sample (n=131) in Chapter 4, the late onset AD sample (n=471) and early and late onset AD sample (n=539) in Chapter 6 demonstrates the inconsistent partitioning of items into Mokken scales within similar diagnostic samples. Table 12.2 presents the hierarchies of ACE-R item *difficulty* for the three AD based samples.

Across the three analyses the only three items of the ACE-R common to all three AD hierarchies; ‘memory retrograde’, ‘recognition’ and ‘write a sentence’ are the items previously identified as common to the eight main analyses of the ACE-R and ACE-III. Focusing on these item hierarchical overlaps the three items are in a consistent order of *difficulty* for each of the three samples; the patients in these samples experience difficulties with ‘memory retrograde’ initially, followed by ‘recognition’ and at a further level of severity will experience difficulty with ‘write a sentence’. However, this three item hierarchical

CHAPTER 12: DISCUSSION

pattern is also present in all other hierarchies of ACE items across all samples reduces the clinical applicability of these item ordering insights to differentiate between different patient groups.

Having examined and exhausted the use of the scant similarities between all three of the AD samples the comparison of items common to two of the three hierarchies for discrepancies of these item orderings establishes a couple of reversals in item *difficulty* between samples. On the one hand, comparing the five items common to the AD type hierarchy and the combined early and late AD hierarchy demonstrates a slight shift in the item *difficulty* ordering for ‘draw a clock’ and ‘recognition’; ‘draw a clock’ is slightly more *difficult* than ‘recognition’ for the AD type sample analysed in Chapter 4, whereas on the other hand this pattern is reversed for the combined early and late onset AD sample from Chapter 6. However the difference in mean scores of the items in the AD type sample are most likely too insignificant to be of any clinical significance.

Therefore the main difference between these hierarchies is not in the ordering of *difficulty* but in the items retained in the hierarchies. This raises a larger issue concerning the reliability of these results. Between the three samples, two of which comprised most of the same data (the late onset AD sample in full is present within the slightly larger combined early and late-onset AD in the analyses of Chapter 6) the application of the same analytic methods resulted in different items being partitioned into the formal hierarchical Mokken scales. If the aim of these analyses was to identify a common and invariant ordering of items by *difficulty* for different dementia groups these mixed hierarchies cast some doubt over the application of Mokken scaling for this purpose.

However, the analyses depend on the samples analysed and as mentioned before the samples analysed in Chapter 4 and Chapter 6 differ considerably. Therefore while the groups

may be diagnostically similar the other differences such as disease severity, age, location etc. could be driving the variation in the results. Additionally the sample size of the AD type group in Chapter 4 was restrictively small which will be discussed later. These factors are likely to influence the results.

12.4 Patterns of poor item discrimination

Examining the pattern of items excluded from analyses due to low scalability coefficients across all of the eight analyses of the full versions of the ACE-R plus the five sensitivity analyses of the ACE-III from Chapter 5 reveals some interesting consistencies. Table 12.3 presents all items excluded for low H_i values for comparison across studies. In nine of the 13 analyses ‘count dot arrays’ and in eight of the 13 analyses ‘repetition of single multi-syllabic words’ are removed due to poor *discriminatory* values. The consistent findings of poor item *discrimination* across several analyses and different clinical samples for ‘count dot arrays’ and ‘repeat: ‘hippopotamus’; ‘eccentricity’; ‘unintelligible’; ‘statistician’ (ACE-R), repeat: ‘caterpillar’; ‘eccentricity’; ‘unintelligible’; ‘statistician’ (ACE-III) indicate that these items bear weak association to the assessment of cognitive impairment in dementia. With regards to ‘count dot arrays’ the low *difficulty* of this item may have influenced the poor *discrimination* in these analyses. With performance often at or near ceiling this item could lack sensitivity to cognitive impairment. In these samples most respondents perform well causing little variation in responses. This near perfect score across samples could lessen the extent of the relationship between item performance and cognitive impairment in dementia. Assessing performance of this item in a sample of patients with more severe levels of dementia may result in a greater level of *discrimination* for this item. However, examining the performance of other low *difficulty* items reveals that this possibility is less likely.

Table 12.2 ACE-R IIO hierarchies from most to least *difficult*: comparison across Alzheimer's disease sample

Chapter 4 Alzheimer's type		Chapter 6 Late onset AD		Chapter 6 Early and late onset AD	
Item	Mean	Item	Mean	Item	Mean
Name and address recall	0.22	Name and address recall	0.06	Memory retrograde	0.42
Memory retrograde	0.50	Memory retrograde	0.39	Recognition	0.50
Draw a clock	0.68	Recognition	0.49	Draw intersecting pentagons	0.57
Recognition	0.69	Draw intersecting pentagons	0.61	Draw a clock	0.64
Orientation in time	0.70	Name and address learning	0.66	Name and address learning	0.65
Serial sevens	0.77	Orientation in time	0.68	Repetition 3	0.68
Syntactical comprehension	0.78	Repetition 3	0.69	Serial sevens	0.76
Orientation in geography	0.79	Semantic comprehension	0.70	Naming 2	0.77
Write a sentence	0.84	Naming 2	0.77	Write a sentence	0.87
Naming (pencil and watch)	0.87	Count dot arrays	0.88	Reading	0.87
3 item registration	0.91	Reading	0.88		
		Write a sentence	0.89		
		Repetition 2	0.91		
		Identify fragmented letters	0.93		

Note. ACE-R=Addenbrooke's Cognitive Examination-Revised, IIO= Invariant item ordering, AD=Alzheimer's disease, Naming 2: name 10 pictures, Repetition 2: repeat 'Above, beyond and below', Repetition 3: repeat 'No ifs, ands or buts'. Items common across all samples listed in bold.

CHAPTER 12: DISCUSSION

‘Three item registration’ one of the least *difficult* items (and less *difficult* than ‘count dot arrays’) was not excluded from any analysis due to poor *discrimination*. Additionally, ‘naming 1’, again less *difficult* than ‘count dot arrays’, was only identified as poorly *discriminatory* in one analysis. However, ‘naming 1’ is not included in the ACE-III which forms the basis for all analyses in Chapter 5. This item is subsumed into ‘naming’ which assesses the subjects ability to identify 12 pictures by name rather than just two (pencil and watch) assessed by ‘naming 1’ in the ACE-R. This change increases the *difficulty* of the item. Therefore it is possible that ‘naming 1’, as a low *difficulty* item, could have demonstrated poor *discrimination* in the analyses of Chapter 5 had it been included. Without such evidence of low *discrimination* as a result of low *difficulty* in these analyses, alternative reasons for the poor *discrimination* of ‘count dot arrays’ along with the other poorly *discriminatory* items will be considered.

The poor *discrimination* could be the result of a weak association between the items and the latent trait (i.e. the ability to count dots or to repeat three words is not strongly associated with cognitive impairment in dementia). Other factors such as poor vision, hearing, fatigue or boredom could be responsible for the less precise association of some items to the assessment of cognitive impairment in dementia. Among the items commonly identified through these analyses as poorly associated to cognitive impairment are three items that rely on sight (‘count dot arrays’, ‘follow written command-close eyes’ and ‘draw a cube’) the other two items associated with poor *discrimination* rely on hearing (‘repetition of single multi-syllabic words’ and ‘repetition: no ifs, ands or buts’). Should fatigue and boredom be influencing the responses to items one would expect the items administered toward the end of the measure be the most affected. Examining the pattern of item *discrimination* in relation to item position in the test administration demonstrates that ‘count dot arrays’, the item most often excluded, is the 23rd item in the ACE-R administration. Other

CHAPTER 12: DISCUSSION

commonly excluded items; ‘repeat: ‘hippopotamus’, ‘eccentricity’, ‘unintelligible’, ‘statistician’ and ‘repeat: no ifs, ands or buts’ are the 13th and 15th items in the test administration order. The position of these items suggests that participants’ attention and concentration may lag in the middle and towards the end of testing. However, the ‘name and address recall’ and ‘name and address recognition’, the last items in the scale, were each excluded only once. It is also interesting to note that items close to the end of the scale represent some of the most (‘name and address recall’) and least (‘count dot arrays’, ‘identify fragmented letters’) *difficult* items in the scale. It would be interesting to examine the effect of item position variation on item responses.

‘Follow written command-close eyes’ was excluded from Mokken scales due to poor *discrimination* in five of the six ACE-R analyses in this thesis. In the revision of the ACE-R and subsequent development of the ACE-III (Hsieh et al., 2013) this item was removed from the scale as it was found to lack sensitivity to cognitive impairment. These analyses corroborate the poor association of this item to the assessment of cognition in dementia. This demonstrates how the results of Mokken scaling analyses can be applied practically.

‘Repetition of single multi-syllabic words’ performed better in the analyses of the ACE-III (Chapter 5) than in the analyses of the ACE-R (Chapters 4 and 6). This item underwent a slight alteration in the updated version with the repetition of ‘hippopotamus’ replaced by the repetition of ‘caterpillar’ in the ACE-III.

Another difference between the ACE-R and ACE-III is ‘Repetition 3’ which in the ACE-R asked subjects to repeat the phrase ‘no ifs, ands or buts’, whereas in the ACE-III the phrase is ‘a stitch in time saves nine’. This change was motivated by the observation that healthy adults were performing poorly on this item (Valcour, Masaki & Blanchette, 2002). Poor performance on this item in healthy subjects could reflect inattention or poor hearing.

CHAPTER 12: DISCUSSION

‘Repeat ‘no ifs, ands or buts’’ was identified as a poorly *discriminatory* item in five of the seven ACE-R analyses whereas ‘repeat: ‘a stitch in time saves nine’’ demonstrated poor *discrimination* in only one of the six ACE-III analyses. This pattern suggests that this item adds greater contribution to the assessment of verbal repetition in the ACE-III.

The *discriminatory* performance of items across these analyses reflects the sensitivity of the scale items to cognitive impairment. Brugnolo et al. (2009) reported similar findings from an assessment of the factorial structure of the MMSE. A factor analysis, based on principal component analysis (PCA) with Varimax rotation identified two components, the first of which contained MMSE items ‘delayed recall’, ‘serial sevens’, ‘orientation in geography’ and ‘orientation in time’ and accounted for 65% of the variance. The second component explained 20% of the total variance and included ‘verbal repetition’, ‘obeying a command’ and ‘obeying a 3 stage command’. This second component did not function as a reliable index of cognitive impairment along the MMSE score range between 29 and 10. Therefore this component does not contribute to the assessment of mild to moderate AD. Most of the participants in the ACE-R and ACE-III analyses had mild-moderate dementia. The results of this PCA therefore support the poor *discrimination* of the items identified here in the assessment of mild-moderate dementia. Alternatively the poor *discrimination* of items could be the result of the size of the samples analysed here. This possibility will be discussed later.

Table 12.3 Number of ACE-R/ACE-III items removed due to low scalability coefficients across samples

	Chapter 4			Chapter 5					Chapter 6				Total	
	AD type	Predominantly frontal dementia	Other frontotemporal lobe degenerative disorders	Mini-ACE development sample	Data plus controls (n=147)	Data minus bvFTD (n=122)	Data minus AD (n=113)	Data minus LPA (n=134)	Data minus PNFA (n=134)	Mixed sample	Late onset AD	Early and late onset AD		Mixed AD VaD
	N=131	N=119	N=100	N=117	N=147	N=122	N=113	N=134	N=134	N=808	N=471	N=539	N=137	
Count dot arrays	X		X	X	X	X	X	X	X				X	9
Repetition 1		X	X	X				X		X	X	X	X	8
Repetition 3	X	X	X						X	X			X	6
Follow written command-close eyes	X		X								X	X	X	5
Draw a cube	X	X	X	X										4
Reading		X		X									X	3
Draw pentagons	X		X											2
Draw loops				X				X						2
Syntactical comprehension			X										X	2
Semantic comprehension			X											1
Recognition													X	1
Name and address recall													X	1
Naming 1													X	1
Naming 2			X											1
Total	5	4	9	5	1	1	1	3	2	2	2	2	9	

Note. AD=Alzheimer's disease, bv-FTD=behavioural variant frontotemporal dementia, LPA=Logopenic progressive aphasia, VaD=Vascular dementia, Mini-ACE=Mini Addenbrooke's Cognitive Examination, Repetition 1=repetition of single multi-syllabic words, Repetition 2='Above, beyond and below' (ACE-R), 'All that glitters is not gold' (ACE-III), Repetition 3='No ifs, ands or buts' (ACE-R), 'A stitch in time saves nine' (Mini-ACE). Naming 1': name pencil and watch, Naming 2: name 10 pictures.

12.5 Using item discrimination to provide insight into the cognitive processes underlying item performance

Determining the degree of *discrimination* can offer an insight into the processes underlying item performance. An item with high *discrimination* is highly related to the construct being assessed whereas those items with poorer *discrimination* are not as closely underpinned by the latent construct. Analysing the *discrimination* of a wide item pool within different clinical subgroups could be valuable in detecting which items display the greater *discrimination*, i.e. sensitivity to the specific cognitive impairment associated with each of the diseases represented in the subgroups. These items could be useful in the development of disease specific scales or identifying items within a general scale that have greater sensitivity to specific diagnoses. For example, items related to memory, such as ‘name and address recall’ may display the highest levels of *discrimination* in an AD sample whereas the items with the greatest *discrimination* in an FTD sample may be those related to frontal executive function such as verbal fluency. Further breakdown of the fluency items may be detected with greater *discrimination* for letter based verbal fluency, which is considered as an assessment of frontal lobe executive function (Elfgren, Ryding & Passant, 1996) in FTD samples and greater *discrimination* for category based verbal fluency which relies on semantic memory (Martin, Wiggs, Lalonde & Mack, 1994) in a semantic dementia sample. While for the most part this level of analysis was not possible due to the limited sample sizes analysed in this thesis, examining item *discrimination* in the SD PNFA sample from Chapter 4 there is an interesting difference in H_i values between both fluency items; fluency-animal is more *discriminatory* ($H_i=0.53$) than fluency-letter ($H_i=0.35$) in this sample.

CHAPTER 12: DISCUSSION

Looking at item *discriminatory* values of items conforming to a hierarchical Mokken scale for the AD sample ($n=471$) from Chapter 6 the item with the highest *discrimination* and hence association to the latent trait is ‘name and address recall’ ($H_i=0.57$). The item with the highest *discrimination* value of the items conforming to the hierarchical scale in the analysis of mixed AD VaD is ‘fluency-animal’ ($H_i=0.49$). The most *discriminatory* item from the ‘predominantly frontal dementia’ sample in Chapter 4 is ‘draw a clock’ ($H_i=0.60$). These differences in highly *discriminatory* items across different samples and patient groups could provide additional insights into which items within the scale are closely associated to the measurement of the latent trait of cognitive impairment.

Comparing the *discrimination* of the five items common to both the hierarchy from the analysis of the ‘AD type’ group in Chapter 4 with those in the late onset AD sample in Chapter 6 reveals a consistent pattern of item *discrimination*. These five items are all sufficiently *discriminatory* having met the criteria for inclusion in Mokken scales. These five items are also common to the hierarchical scale in the analysis of the ‘other frontotemporal lobe degenerative disorders’ sample in Chapter 4. However the items do not follow the same pattern of item *discrimination* as the AD samples. The consistency of the ordering of item *discrimination* from ‘name and address learning’ as the item with the highest *discrimination* to ‘orientation in time’ with the lowest *discrimination* within the hierarchies for each sample supports the use of *discrimination* as a potentially useful item parameter in the comparison of different causes of cognitive decline.

The assessment and comparison of item *discrimination* between samples demonstrates how in two samples of patients with AD ‘name and address recall’ is a highly *discriminatory* item showing a strong association to cognitive impairment and capable of differentiating between patients of different levels of ability. Two other memory items and

one of the orientation items are also amongst these common hierarchical items providing good *discriminatory* power.

12.6 Limitations of ACE analyses

Over the course of performing these analyses on data from the ACE-R and ACE-III some methodological issues and limitations were identified. These will be discussed here before a more general discussion of limitations of the methods and results of the analyses and thesis as a whole.

12.6.1 Importance and implications of samples used

The samples on which analyses of the ACE-R and ACE-III are based on differ considerably in several ways. This has significant implications for the comparison and collation of results across samples and chapters. The cognitive assessment data in Chapter 6 were collected from a Scottish dementia case register (SDRIR). The data analysed in Chapter 4 were collected by a specialised tertiary research clinic in Sydney, Australia (Frontier Research Group) and the data analysed in Chapter 5 were collected from two specialised research clinics in the UK (the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom) and the same research clinic in Sydney (Frontier Research Group). The differences between these samples and potential impact of these differences on the interpretation and comparison of results will be discussed.

12.6.2.1 Different diagnoses of samples

Due to the research interest of the specialised research clinics in Sydney, Oxford and Cambridge the samples collected from these clinics comprise a greater proportion of patients

CHAPTER 12: DISCUSSION

with less common diagnoses such as bv-FTD, semantic dementia, primary nonfluent aphasia and logopenic progressive aphasia. The addition of these rarer forms of dementia is a valuable addition to these studies and extends the scope of the analyses. However this greater variation of diagnoses complicates the comparison of the results of the analyses of these samples with the data from the SDRIR participants. The SDRIR participants are generally more representative of the general population with a majority of patients with AD on the register. While one aim of these analyses was to determine whether there were disease specific patterns of decline the samples of the less common patient groups were too small for this level of stratified analysis. These diagnostic differences raise important questions for the new brief versions of the ACE created in this thesis as one—the Mini-ACE—was developed from the analysis of data from Sydney whereas the other—the Short ACE-R—was developed from the analysis of SDRIR data.

12.6.2.2 Age difference

Data from the Frontier Research Group, Sydney, Australia, the Memory Clinic of Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom and the Oxford Cognitive Disorders Clinic, Oxford, United Kingdom comprised samples with a younger mean age than the data collected by the SDRIR. The younger mean age of the samples from the tertiary research groups is a result of preponderance of frontotemporal dementia which occurs much more commonly in younger populations (Jefferies & Agrawal, 2009). Disease onset prior to the age of 65 years is typically referred to as early-onset dementia. A study of the clinical characteristics of early-onset dementia reported that patients diagnosed with early-onset AD followed a more rapid rate of progression than late-onset groups (Jacobs et al., 1994). Different patterns of performance on the MMSE have been found for early and late-onset AD. Patients diagnosed with early-onset AD performed significantly worse on attention items within the MMSE than those diagnosed with late-onset AD whereas the early-onset patients

CHAPTER 12: DISCUSSION

performed significantly better on memory and naming assessments at baseline (Jacobs et al., 1994). Comparison of clinical features of late and early onset AD found a greater prevalence of non-memory presentations in early-onset patients than those with late-onset (Koedam et al., 2010). These non-memory cognitive impairments included visuospatial impairments, which were followed by impairments within the language domain.

The different patterns of impairment could have significant implications for the generalisability of the results of analyses of these relatively young samples in this thesis.

In Chapter 6 I analysed a mixed clinical group of late and early onset AD. While comparison of the hierarchical pattern of decline of the mixed sample and the purely late onset AD did not suggest that the patterns of item performance differed between late and early-onset AD there is some indication of differences within these patient groups. Fewer items conformed to an IIO hierarchy in the combined late and early-onset AD sample despite it being a larger sample than the late-onset only sample. While this is a relatively speculative claim and further analysis of two distinct and sufficiently large samples of each is necessary, the number of violations of Mokken scaling assumptions suggests a greater degree of response variation. In the analysis of late-onset AD in isolation 14 items were retained in an IIO hierarchy of *difficulty* whereas in the combined analysis of late and early-onset patients only 10 items were retained. Additionally, within the patients groups of the SDRIR examined in Chapter 6 the late onset-AD patients ($n = 68$) had the lowest mean ACE-R score (60.9, $SD=21.4$). These findings suggest that not only do the patterns of cognitive decline potentially differ but also the severity of cognitive impairment.

These possible differences between patients diagnosed with early-onset and late-onset AD suggest that the low mean age (mean age= 65.4 years, $SD=8.5$) of the samples collected from the specialised research clinics could limit the generalizability of the results of the

CHAPTER 12: DISCUSSION

Mokken scaling analyses of the samples in Chapter 4 and 5 as it is more indicative of early onset dementia.

These age differences have implications for the generalizability of and comparison of the two new scales derived from Mokken scaling analyses of ACE-R and ACE-III data. The Mini-ACE was developed and validated using a sample of patients with a mean age more typical of early-onset dementia (mean age= 65.4 years). This raises some concerns regarding the generalizability the IIO hierarchy that formed the basis for item selection for the Mini-ACE, particularly in light of the evidence for different cognitive profiles for early and late-onset AD.

The Short ACE-R was developed in a typically late-onset aged sample (mean age= 77.5, SD=7.8) but was validated in a younger sample (mean age=65.4, SD=8.5). The difference in Short ACE-R development and validation samples could help to explain the poorer relative performance of the Short ACE-R in the validation analyses than that of the Mini-ACE, which was validated in a similar sample to the one in which it was developed from. This variance in age across analyses could limit the reliability of the Short ACE-R.

12.6.2.3 Testing conditions and location

The samples used to develop and validate the Mini-ACE and in the analyses of the ACE-R in Chapter 4 were assessed at tertiary referral clinics. The SDRIR samples used to develop the Short ACE-R and in the analyses in Chapter 6 were tested in their own homes. These differences potentially introduce some further variance between the results and could affect the comparison of results across these analyses. For example, many patients attending the specialised research clinics for assessment would have travelled some distance to attend the clinics meaning they would be out with their normal geographical range. The stress and change of environment could influence a patient's test performance in general and could have

specific effects of the performance of certain items. A patient's temporal and geographic orientation is likely to be affected by significant changes in their location (i.e. the items may be more or less *difficult* for participants depending on the familiarity of their location at assessment). For example, a participant who has been reminded several times that day that they are travelling to a clinic in a certain location is likely to be more aware of the time, date and location. However the novelty of the surroundings may also make it more likely for participants to lose their bearings and become disorientated in time and location.

The SDRIR participants were tested in their own homes. Assessment in the more familiar surroundings of home where the time of the day is less likely to be reinforced may make it harder for these participants to be aware of the time and date due to the everyday routine. While being in familiar surroundings may make it easier for a patient to lose track of time and date the familiarity of the home may increase orientation to location.

It would be interesting to compare performance on the orientation items of the same patients assessed on two occasions in different locations. Comparing performance of those assessed at a specialist clinic with those tested in their own home could help to confirm whether the location and novelty of surroundings has an effect on test performance. It is more difficult to determine the effect of testing location between the SDRIR and specialised clinic samples due to the other differences between the samples; diagnosis, disease severity and age.

12.6.2 Formation of clinical groups for analysis

The analyses in Chapter 4 were performed to investigate different sequences of cognitive decline in different dementia syndromes including some of the less common forms of dementia such as semantic dementia and progressive non-fluent aphasia. In order to determine whether the item ordering differed between these different disorders it was

CHAPTER 12: DISCUSSION

necessary to perform dementia specific stratified analysis. However due to the limited number of participants overall and of patients in some of these more rare patient groups the desired level of analysis could not be performed. Therefore clinical groups had to be formed to provide sufficient numbers for analysis.

There were difficulties in the formation of these groups as discussed in Chapter 4 but the ultimate decision was made to combine patients diagnosed with AD with those diagnosed with LPA to form an ‘Alzheimer’s type’ group, to combine patients diagnosed with bv-FTD and patients diagnosed with FTD-MND to form the ‘frontal dementia’ group and to form the ‘other frontotemporal lobe degenerative disorders’ group by combining patients diagnosed with SD and PNFA.

These groups were formed on the basis of theoretical or structural similarities between the diagnostic groups. Based on the commonality of pathology it was hypothesised that the cognitive decline observed in AD might resemble that observed in LPA. Clinicopathologic studies have most frequently associated LPA with the pathology of AD (Mesulam et al., 2008). LPA is classified as an atypical focal language variant of AD (Karantzoulis & Galvin, 2011). The core deficit of LPA seems to relate to the phonological loop function which is involved in auditory-verbal short term memory (Gorno-Tempini et al., 2008). It is suggested that impairments to this phonological loop account for the pattern of language impairment; the concurrence of naming and sentence repetition impairments in the absence of impairments in single word comprehension and spared grammar (Bonner, Ash & Grossman, 2010). The clinical symptoms of LPA are similar to the language deficits often described in AD (Grossman, Mega, Cummings, Joynt & Griggs, 2004).

The impairments associated with AD affect memory and orientation more than any of the other cognitive domains initially (Dubois et al., 2007). The pattern of cognitive decline

CHAPTER 12: DISCUSSION

continues to cause deficits in attention, visuospatial skills and language (Alladi et al., 2007). This pattern of decline corresponds to the early pathology in the medial temporal lobe which then spreads to other neocortical association regions (Alladi et al., 2007).

Patients with LPA have greater episodic memory impairments than patients with other primary progressive aphasia variant (Bonner et al., 2010) and Flanagan, Tu, Ahmed, Hodges and Hornberger (2013) reported patients with LPA and AD were similarly impaired on assessments of memory and orientation. Therefore as the clinical features of LPA are commonly associated with AD patients from these two clinical groups were combined for analysis.

Frontotemporal degeneration resulting in focal atrophy of the frontal and or the anterior temporal lobes presents in one of two major variations; behavioural (bv-FTD) and a language variant (Lillo & Hodges, 2009). Within the language variant there are two further divisions based on the pattern of underlying atrophy and associated clinical characteristics; semantic dementia and progressive nonfluent aphasia. The progressive impairment of semantic memory was described as semantic dementia by Snowden, Goulding and Neary (1989). Thorough characterisation of semantic dementia followed (Hodges, Patterson, Oxbury & Funnell, 1992). Soon afterwards another different variation of progressive language impairment was described as progressive nonfluent aphasia (Grossman et al., 1996). A consensus meeting established criteria for both semantic dementia and progressive nonfluent aphasia in relation to degeneration of the frontotemporal lobe (Neary et al., 1998). This led to the classification of cases of primary progressive aphasia as semantic dementia (“fluent”) or progressive nonfluent aphasia (“nonfluent”). These two variations of the language variant of frontotemporal degeneration; fluent and nonfluent progressive aphasia formed one of the clinical groupings in these analyses.

CHAPTER 12: DISCUSSION

Patients diagnosed with the fluent semantic variation demonstrate a signification loss in semantic memory, which affects comprehension of words, objects and faces in the presence of fluent speech. Impaired confrontation naming and single word comprehension are hallmarks of this disorder (Gorno-Tempini et al., 2011). The ability to repeat words and phrases is spared, as are episodic memory and visuospatial skills (Lillo & Hodges, 2009).

In comparison patients diagnosed with the PNFA demonstrate a gradual and progressive loss of expressive language skills (Lillo & Hodges, 2009). This manifests in agrammatical and nonfluent speech where phonological errors commonly occur. Unlike SD, the ability to repeat phrases and multisyllabic words is impaired (Gorno-Tempini et al., 2011) and the dissociation between the two variants is also evident in the sparing of single word comprehension and object recognition (Lillo & Hodges, 2009).

While both SD and PNFA are presentations of frontotemporal lobe degeneration the distinctive clinical features used to delineate SD and PNFA make the analysis of the ‘other frontotemporal lobe degenerative disorders’ group more complex. For example patients with SD will have difficulty with the ‘naming’ items of the ACE-R whereas patients with PNFA are less likely to have difficulty with these items. Therefore attempting to find a consistent pattern of *difficulty* of both SD and PNFA patients will be more problematic. Both naming items of the ACE-R were excluded from the IIO hierarchy of *difficulty* for this group. ‘Naming 2’ (naming 10 objects) demonstrated poor *discriminatory* power indicating that it was weakly associated with the cognitive impairment demonstrated by these groups. It is noteworthy that ‘Naming 2’ did not demonstrate poor *discrimination* in any other analysis in this thesis. ‘Naming 1’ (naming pencil and watch) was excluded from the formal hierarchy of items due to a violation of IIO. It would be interesting to examine the performance of these items in a sample of SD patients where these items are likely to demonstrate a high level of *difficulty* and better association to the underlying cognitive impairment.

CHAPTER 12: DISCUSSION

For the formation of the 'Frontal dementia' group patients diagnosed with bv-FTD and FTD-MND were combined for analysis. Results from genetic and pathologic studies support the idea of a continuum between FTD and MND (Valdmanis et al., 2007). Both neurodegenerative disorders have overlapping common clinical and neuropathological characteristics (Lillo & Hodges, 2009). Clinical and neurophysiological evidence of MND are found in approximately 10% of patients diagnosed with FTD (Lillo, Garcin, Hornberger, Bak & Hodges, 2010; Lomen-Hoerth, Anderson & Miller, 2002). Patients diagnosed with MND also demonstrate changes in behaviour and or language that are sometimes sufficiently severe to warrant an FTD diagnosis (Lillo, Mioshi, Zoing, Kiernan & Hodges, 2010). While these two groups appear homogenous there has been difficulty identifying a specific cognitive profile for early bv-FTD (Piguet, Hornberger, Mioshi & Hodges, 2011). While cognitive impairments in executive function and episodic memory generally emerge from through cognitive examination these deficits become less specific as the disease and atrophy spreads to the anterior temporal regions (Piguet, Hornberger, Mioshi & Hodges, 2011).

Variable pathology underlying several types of dementia has been reported. For example, AD pathology was found in nine out of 20 patients diagnosed with PNFA (Kertesz, McMonagle, Blair, Davidson & Munoz, 2005). Grossman et al. (2008) also reported this overlap where AD pathology was noted in three of nine cases of PNFA followed longitudinally. Again pathological overlap was discovered amongst 23 patients with PNFA; seven cases where AD pathology was found (Alladi et al., 2007; Galton, Patterson, Xuereb & Hodges, 2000) and four cases where 'motor neurone disease inclusion dementia' was reported (Knibb, Xuereb, Patterson & Hodges, 2006). A case of PNFA was also reported which demonstrated the pathological characteristics of dementia with Lewy bodies (Kertesz, McMonagle, Blair, Davidson & Munoz, 2005).

These reports of variation in pathology suggest that the attempts to identify patterns of cognitive decline in different types of dementia using Mokken scaling methods are restricted by diagnostic classification and accuracy limitations. In Chapter 4 the interpretation of the findings are restricted not only by the necessity to form non distinct clinical groups but also by the potential for variation within the seemingly distinct diagnostic classifications.

12.6.3 Diagnostic circularity concerns

As both the ACE-R and ACE-III formed part of the neuropsychological test battery there are concerns regarding the potential for circularity bias. Relying on clinical rather than neuropathologic diagnoses of dementia can introduce an element of diagnostic circularity. Patients from the Sydney, Oxford and Cambridge samples (Chapter 4, Chapter 5) were assessed by multidisciplinary teams who based their diagnoses on current diagnostic criteria (Mathew, Bak & Hodges, 2012, McKhann et al., 2011, Rascovsky et al., 2011, Gorno-Tempini et al., 2011) taking extensive clinical assessments, comprehensive neuropsychological assessment, and evidence of atrophy on structural MRI brain scans into account. The diagnoses of the participants on the SDRIR (Chapter 6,7,9:11) were established by the classification by an old-age psychiatrist and physician following independent assessment comprising ICD-10 diagnosis, severity using the Clinical Dementia Rating Scale (CDR) (Morris, 1993), co-morbidities, medication, communication difficulties, cognition (using the ACE-R), function (Lawton Instrumental Activities of Daily Living and Personal Self-maintenance scales; Lawton and Brody, 1969) and behavioural and psychological symptoms of dementia (Neuropsychiatric Inventory with Carer Distress Scale; Kaufer et al., 1998).

As the scales under assessment in these analyses were used to support clinical diagnoses there is a concern that the patterns identified here in the clinical groupings may be

overestimated. For example, if clinicians used severe episodic memory impairments to support a diagnosis of AD the patterns revealed in Mokken scaling analysis would reflect this making it likely that ‘name and address recall’ for example would be identified as the most *difficult* item for these patients. The different patterns of item performance identified through Mokken scaling analyses could have been what contributed to the respective diagnoses in the first place. For example in Chapter 6—that the patients with mixed AD VaD performed worse on the visuospatial task of ‘drawing intersecting pentagons’ than the patients with AD whereas the patients with AD performed worse on an assessment of ‘orientation in time’ could have contributed to the respective diagnoses.

To avoid this potential influence of diagnostic circularity on the patterns identified here further research where diagnoses are made completely independently of ACE-R, ACE-III or MMSE (owing to the considerable overlap in terms of item content between the MMSE and ACE-R) is necessary.

12.6.4 Polytomous item score equating

As a nonparametric model Mokken scaling relies on ranking by mean scores as a means of establishing item *difficulty*. This presents a problem when, as in the case of the ACE, the items differ in the range of scores. As a method of equating the scores to enable comparison and ranking the mean score for each item was divided by the maximum number of points available for each item. For example, the mean score for ‘memory retrograde’ was divided by four, the mean score for ‘draw a cube’ was divided by two and ‘name and address recall’ was divided by seven. This process provided mean values between zero and one for all items, which could then be ordered by relative *difficulty*.

While for the most part this was a straightforward solution and created comparable scores for all ACE items ranging between zero and one, the design of some items presented

CHAPTER 12: DISCUSSION

some cause for concern. Whereas raw scores are used for all other scale items verbal fluency was scored using a scaled scoring system derived using a Gaussian distribution of the raw scores from control data (Mathuranath et al., 2000). This scoring system assigns a score from zero to seven based on the number of words generated (see Appendix A). Using a mean value to describe the *difficulty* of this item is somewhat problematic. For example, the mean equated score for verbal fluency letter in Chapter 6 is 0.54, which in its original un-equated form is 3.78. Applying this mean value to the scoring system for this item means on average most people in this sample generated approximately 8-10 words. Therefore ranking this item amongst the other items implies that generating 8-10 words is less *difficult* than 'draw a cube' and more *difficult* than 'orientation in geography'. This disregards the difficulty involved in generating any other number of words.

In Chapter 4 the mean equated score for 'verbal fluency-letters' for the 'AD type' group is 0.48, which in its original un-equated form of 3.4 corresponds to the generation of between six and seven words. The mean equated score for this item for the 'other frontotemporal lobe degenerative disorders' group is 0.29, which in its original un-equated form of 2.0 corresponds to the generation of between four and five words. 'Verbal fluency-animals' is also less *difficult* than the AD type sample than for the 'Other frontotemporal lobe degenerative disorders' sample but again the 'AD type' and 'Other frontotemporal lobe degenerative disorders' differ in the mean scores for 'verbal fluency' which means the *difficulty* will reflect two different things making this comparison redundant. Therefore comparisons across these samples by *difficulty* do not equate for these items. This discrepancy illustrates how the apparent *difficulty* corresponds to different outcomes depending on the sample analysed. Unless the number of animals or words is held constant across samples the comparisons are not valid.

However these inconsistencies did not affect the interpretation of any comparisons across samples in these analyses. In Chapter 4 the verbal fluency items were only retained in the IIO hierarchy of ‘other frontotemporal lobe degenerative disorders’ sample and in Chapter 6 the fluency items were only retained in the IIO hierarchy ‘mixed AD VaD’ sample.

12.6.5 Testlets

Several of the ACE-R items can be referred to as testlets (Wainer & Kiely, 1987) whereby the sum of the items within the testlet can be considered as a polytomous item score. For example, ‘identify fragmented letters’, contains four individual responses; ‘identify letter K’, ‘identify letter M’, ‘identify letter A’ and ‘identify letter T’.

Full and complete itemisation is required to fully examine the items within the ACE. For example the recording of scores for ‘orientation in time’ should include the dichotomous response to ‘orientation to day’, ‘orientation to date’, ‘orientation to month’, ‘orientation to year’ and ‘orientation to season’. This more precise level of scoring would allow further delineation of item properties. For example, it could be determined how *difficult* correctly identifying the year is in comparison to knowing which day of the week it is. The Ashford et al. study (1989), included in the systematic review in Chapter 3, examined the MMSE items at this level and ascertained that a correct response to the date was the most *difficult* and a correct response to the season was the least *difficult*. Given this variance the item ordering by *difficulty* determined by the Mokken scaling analyses in this thesis may not be as practically significant. For example, if for patients with AD ‘orientation to date’ is the most *difficult* and ‘orientation to season’ is the least *difficult* the overall performance on ‘orientation in time’ will be an average of performance on the embedded items. Additionally, useful information is lost without the consideration of these embedded items. While different profiles of temporal and geographical disorientation for different patient groups were reported in Chapter 4 there

CHAPTER 12: DISCUSSION

may also be different patterns of impairment within the orientation items for different patient groups. If these differences exist they will not be detected using Mokken scaling methods unless the embedded items of testlets are scored and reported individually.

Ashford et al. (1989) revealed the specific ordering for the individual orientation items in an item response theory analysis of patients with Alzheimer's disease. For this sample the most *difficult* of the embedded orientation items was 'orientation to date', followed by 'orientation to day', 'orientation to country', 'orientation to month', 'orientation to year', 'orientation to season', 'orientation to place', 'orientation to floor', 'orientation to city' and finally 'orientation to state'. It would be interesting to determine whether the ability to analyse the AD samples in this thesis at this level would replicate these findings and whether there would be a different pattern of item ordering for the other diagnostic samples.

However, for other items it is not as practically significant to determine the item ordering within the testlet. For example, in the assessment of 'memory retrograde' it is not especially important to discern whether naming the president of the USA is a more *difficult* item than naming the current Prime Minister. In this instance four questions are asked so as to provide a reasonable assessment of a person's retrograde memory, which is the summed score out of a possible four.

While it would be an interesting and valuable extension of this research to determine item properties of all embedded items more extensively, another approach would be to apply an alternative model which is capable of estimating item response characteristics of polytomous item with different numbers of response categories.

12.6.6 Alternative method: Graded Response Model

The graded response model (Samejima, 1996) extends the two-parameter IRT model to enable the characterisation of item responses as ordered categorical responses (Hays, Morales & Reise, 2000). Importantly given the ACE-R and ACE-III data examined in this thesis, under the graded response model the items are not required to have the same number of response categories. Within this model each item is described by a slope or *discrimination* parameter and between category threshold parameters that demonstrate the level of latent trait required to have a .50 probability of responding above the particular threshold level.

Category response curves illustrate the probability of a response within a particular category of response as a function of the latent trait. With regards to the ACE-R this level of analysis allows the quantification of the trait level required at each response category (e.g. scoring 1 out of 5) to have a .50 probability of scoring in a higher category within the same item (e.g. scoring 2 or more out of 5).

The graded response model can yield additional interesting item features to the measurement on condition that the model fits the data well (Sijtsma, Emons, Bouwmeester, Nyklicek & Roorda, 2008). However this method could not be applied to the data in this thesis as it is a parametric method. Mokken's MHM is a nonparametric version of Samejima's (1969, 1972) homogenous case of the graded response model (Sijtsma & Molenaar, 2002). The nonparametric Mokken scaling is very well suited for data analysis of cognitive outcomes in a dementia population. Analysis of control data could be performed using the parametric method but this would be unlikely to contribute any meaningful information as most subjects would score at ceiling for all items. Additionally this analysis, while interesting from a psychometric point of view, would not contribute to the understanding of the patterns of cognitive decline in dementia. Therefore as Mokken scaling

is the most appropriate analytic method for samples analysed here it is apparent that there is a need for the further delineation of the testlets of the ACE-R and ACE-III.

12.7 Assessment of functional assessment scales

As functional impairment is a diagnostic criterion for dementia (DSM V, American Psychiatric Association, 2013) the outcome of the assessment of functional ability plays an important role in the diagnosis of dementia. In Chapters 9, 10 and 11 the item properties and hierarchical structure of two of the most commonly used measures of functional impairment were examined using Mokken scaling analyses. The motivation of these analyses was to determine the pattern of functional decline in dementia and to investigate whether there were disease specific patterns of impairment.

12.7.1 Hierarchical ordering by difficulty

The pattern of item *difficulty* for both the Lawton IADL and PSMS scales was generally consistent across diagnostic samples and analyses. The order of decline in function as measured by the Lawton IADL scale started with difficulty with shopping followed by preparing food, taking medicine, doing laundry, handling finances, travelling, housework with the loss of independence in using the telephone occurring last in this sequence. There are some minor variations in this item ordering across all hierarchies; doing laundry is more difficult than managing finances for the non-AD sample analysed in Chapter 10.

This generally consistent pattern of decline suggests that regardless of the type of dementia and different cognitive processes underlying these conditions the loss of functional independence occurs in a uniform sequence. However, the samples analysed here are largely comprised of patients with AD with the exception of the non-AD analysis in Chapter 10. This diagnostic similarity makes it difficult to fully determine the variation in functional

assessment in dementia. Analyses of a wider range of diagnostic groups, matched for age and disease severity is necessary. Additionally the limited range of items in these scales provides less opportunity for differences in patterns of functional loss.

While the majority of Lawton IADL and PSMS scale items were retained in the formal hierarchies ‘shopping’ and ‘food preparation’ and ‘physical ambulation’ and ‘toilet’ were excluded from the Lawton IADL scale and PSMS scale respectively due to violations of IIO. The exclusions were made due to IIO violations. No functional scale item was excluded due to poor *discrimination*. The items of the Lawton IADL and PSMS scales therefore demonstrate sufficient association to the latent trait of functional impairment and are capable of differentiating between patients of different levels of functional dysfunction. The results of Mokken scaling of the functional scales in these analyses do not suggest that any item should be removed as in the case of the identification of items within the ACE, which should be removed or altered. However these analyses brought flaws in the scales as a whole to attention.

12.8 Limitations of analyses of functional scales

All analyses of functional scales were performed on data collected from the SDRIR. The analysis on the sample in Chapter 9 was performed in 2012 whereas the analyses in Chapters 10 and 11 were performed in 2014. Due to the use of data from the same register in all analyses of functional scales it is likely that the data analysed in Chapter 9 are also present in the subsequent analyses of the SDRIR data. This overlap of a potential 202 participants increases the similarity of the samples and of the results. This limits the extent to which comparisons can be made between the results of the analysis in Chapter 9 with the results of other analyses. The item ordering of the Lawton IADL items is consistent between the results of Chapter 9 and the largest sample in Chapter 10, which does support the similarity of the

samples. However, the similarity of the results is also likely to reflect the similarity of the patients referred and consenting to take part in the SDRIR.

12.8.1 Scoring bias of scales

While the items of the Lawton IADL scale are scored dichotomously they vary in terms of the range of response categories (0-3: 'laundry', 'medications' and 'finances' 0-4: 'telephone', 'shopping', 'food preparation', 0-5: 'housework' and 'travel') and the different score options associated with these responses. For example, within the four possible response categories for 'telephone' three are associated with a score of one and only one response category ('does not use telephone at all') is associated with a score of zero whereas within the four response categories for 'shopping' this pattern is reversed with only one of the four associated with a score of one ('takes care of all shopping needs independently') and the other three resulting in a score of zero for the item). With regards to the score outcome 'taking care of all shopping needs independently' is equal to 'answers telephone, but does not dial' as both options result in a score of one. These scoring divergences have a clear impact on the expected scores; for 'housework', based on the response categories of this item and their associated scores, the likelihood of scoring zero for this item is 20%, for 'telephone', for 'shopping' and 'food preparation' the probability of a score of zero is 25%, 'laundry' and 'finances' have a 33.3% likelihood of scoring zero, 'travel' has a 40% likelihood of scoring zero and the response categories and associated scores result in a 66.6% probability of a score of zero for 'medications'. The scale design therefore forms its own natural and informal hierarchy based on the scores assigned to each of the response categories. Therefore due to the scale design alone a hierarchy of expected item scores emerges starting with 'medications' as the most *difficult*, followed by 'travel', 'finances', 'laundry', 'food preparation', 'shopping' with 'telephone' as the least *difficult*.

CHAPTER 12: DISCUSSION

Looking at the results of the analyses of this scale in Chapter 9 and 10 we can compare the hierarchies formed from Mokken scaling analyses with this informal hierarchy of *difficulty* based on the likelihood of scoring a zero. This comparison shows that it is possible that the relatively high *difficulty* of ‘medications’ could be influenced by the high probability of scoring a zero due to the item response options (66.6%). Only those who are ‘responsible for taking medication in correct dosages at correct time’ receive a score of one for this item which demonstrates the level of ability required to score well on this item in comparison to a score of one for ‘housework’ for an individual who ‘needs help with all home maintenance tasks’.

That ‘shopping’ and ‘food preparation’ were consistently identified as the most *difficult* items in the hierarchy despite the relatively low percentage likelihood (25%) of receiving a score of zero emphasises the importance of these items in characterising the initial signs of functional impairment. The items assessing the mid-range of *difficulty* by mean scores; ‘laundry’ and ‘finances’ are also in the mid-range of *difficulty* by response options.

To determine the effect of these score options on the results further analysis is required. The various response options should be exchanged for a simple and dichotomous ‘requires assistance/does not require assistance’ classification system. Alternatively or simultaneously the scale as it stands could be examined using Samejima’s graded response model, which accommodates the different number of response categories. This model could be applied to determine the level of latent trait required to have a .50 probability of being able to ‘launder small items’ and compare it to the trait level required to have a .50 probability of being able to perform ‘personal laundry completely’. This analysis would provide additional insight into the decline in functional ability than the dichotomous classification analysis using Mokken scaling. However as in the case of the ACE, the data would be required to conform to parametric assumptions.

12.8.2 Influence of scales analysed

The findings of this thesis are also influenced by the measures examined. For example, in the assessment of gender differential item functioning in Chapter 10 men were generally found to have poorer functional ability as reflected by their relatively worse performance on the Lawton IADL scale despite having a higher mean ACE-R score reflecting better cognitive ability than the women of the sample. These results indicate that men are on an escalated course of functional decline. However this needs to be explored further to determine whether this is the case or whether it is actually an artefact of the measurement scale. With the established gender differences in the assessment of functional impairment with the Lawton IADL scale it is necessary to discern whether the worse performance of men is due to more severe functional impairments occurring at an earlier stage or due to cultural or societal issues relating to the methods of assessing functional capacity. This could be achieved by assessing functional performance in a more comprehensive manner. Observational measures whereby ability is measured through practical demonstrations may provide greater scope for detecting differences between patient groups. Merrit and Fisher (2003) did not observe any gender related differential item functioning using the Assessment of Motor and Process Skills (AMPS). This observational assessment of functional status involves a prior interview with each participant to ascertain which of the 50 AMPS activities (e.g. making a salad, cleaning a bathroom, and weeding) matches the participants' everyday functional routine. Here it can be argued that the similar performance by men and women is because all tasks assessed are practiced and familiar to all being tested. This practice circumvents the possibility of individuals receiving a low score on an item due to unfamiliarity with the task. This highlights a limitation regarding the use of scales such as the Lawton IADL and PSMS scale to measure functional ability. However the amount of time required for a more comprehensive assessment, such as the AMPS in comparison to the relative ease with which

the Lawton IADL scale can be administered mean that the simple scales are still widely used in busy clinical settings where longer assessments would put the available resources under pressure. Therefore it is important to draw attention to the potential for DIF using traditional measures such as the Lawton IADL. IRT methods can detect DIF and could be applied to develop a new scale without these gender differences.

12.9 Development and practical significance of hierarchical scales using Mokken scaling analysis

I applied item response theory modelling to the development and evaluation of new assessment scales in dementia. Three new scales were derived from these applications of Mokken scaling analysis; the Mini-ACE, the Short ACE-R and the Mini NART. Using IRT methods to derive new measures from existing scales allows for the removal of items that do not contribute to the underlying trait. Excluding these items with poor *discrimination* and scalability reduces the effect of irrelevant ‘noise’ on measurement accuracy.

The development of scales using Mokken scaling analysis can support the creation of hierarchical scales. Scales meeting hierarchical criteria are practically valuable as well as psychometrically significant. Hierarchical scales allow clinicians working in busy clinical environments to reduce the time of cognitive testing without the loss of measurement precision. Adapting testing to individual patients, as permitted with IIO hierarchical scales, additionally reduces the burden placed on patients by administering items that are too *difficult* for them. Testing patients on items that are too easy for them leads to an unnecessary addition to testing time. Importantly, reducing the burden in this way will most likely lead to reduction in measurement error caused by poor concentration and inattention and fatigue due to the administration of inappropriate items.

The items of the Mini-ACE formed a formal hierarchical Mokken scale in the validation analyses. Therefore the scale can be used in the manner described above. While these are not essential features of a brief scale such as the Mini-ACE the creation of hierarchical scales ensures good psychometric properties, i.e., that all items are highly *discriminatory*.

Unlike the Mini-ACE the items of the Short ACE-R did not conform to a formal hierarchical scale. Its items however did exceed the *discriminatory* requirements of all items in a Mokken scale which demonstrates that like the Mini-ACE its items were capable of differentiating between different levels of cognitive impairment.

12.9.1 Comparison of Mini-ACE and Short ACE-R

While both the Mini-ACE and the Short ACE-R were formed from analyses of the same design and methodology the items selected for inclusion in the scales differ. The Mini-ACE was derived from the analysis of the ACE-III data collected in Sydney, Australia in Chapter 5, whereas the Short-ACE-R was developed from the analysis of ACE-R data collected in Scotland in Chapter 7.

The pool for item selection for both scales was determined by the results of Mokken scaling analyses of the full scales; specifically the items conforming to invariantly ordered Mokken scales. In Chapter 5 17 items of the ACE-III met this criterion whereas only 11 ACE-R items were included in the IIO hierarchy forming the basis for item selection in Chapter 7.

While IIO is a strong psychometric feature the limited range of items for selection for the Short ACE-R raises the question of whether this criterion for item selection was too restrictive to permit the most appropriate scale in terms of item content. Only 11 items

CHAPTER 12: DISCUSSION

conformed to the IIO hierarchy of ACE-R items in the analysis of SDRIR participants in Chapter 6 which formed the basis for item selection for the Short ACE-R. Of these 11 items only one item from the memory domain—‘recognition’—was included. A brief cognitive tool in the assessment of dementia would be expected to assess memory. However the inclusion of ‘recognition’ in the scale is redundant without ‘name and address learning’. ‘Name and address learning’ was not included in the IIO hierarchy but was included in the shortened scale nonetheless due to the significance of its content and relevance to ‘recognition’ which was included in the formal hierarchy.

Perhaps, given the extent of exclusions made due to violations of IIO from the analyses in Chapter 6 which resulted in a smaller item pool for selection for the Short ACE-R, drawing from the larger item pool of items meeting Mokken’s less restrictive MHM would have permitted the development of a superior scale. Several key items in the assessment of dementia were excluded for violating IIO which otherwise demonstrated good item properties. Only two items; ‘repeat single multi-syllabic words’ and ‘repeat: no ifs, ands or buts’ failed to meet the assumptions of the MHM. The remaining 24 items were sufficiently *discriminatory* and assessed a wider range of *difficulty* than the 11 item hierarchical scale. These items could have been incorporated into the new scale to provide a wider range of assessment. The items of the IIO hierarchy were also restricted in the range of item *difficulty*. The most *difficult* item in the hierarchy was ‘recognition’ with a mean score of 0.51. The more inclusive MHM item pool would extend the level of measurement to the assessment of more severe dementia with ‘name and address recall’ (mean=0.07) and ‘three item recall’ (0.24) and offer a wider selection of items from each cognitive domain. Additionally, despite the items not being selected from an invariantly ordered hierarchy it is possible that the brief scale would still form such a hierarchy were it re-analysed. A future study could investigate

CHAPTER 12: DISCUSSION

this possibility and compare the scale performance of this scale in comparison to the Short ACE-R developed from the IIO hierarchy in Chapter 5.

Beyond the desire to include a diverse range of items to enable the domains of the full version to be assessed the item properties as determined by Mokken scaling analysis guided item selection. Item *difficulty* was considered to ensure a reasonable breadth of measurement and the items' *discriminatory* value was used to ensure that the most sensitive item assessing each of the domains was selected. The analysis and consideration of these item parameters in the creation of these new scales was in accordance with Mungas and Reed's (2000) criteria for an ideal measure of global functioning in dementia; a scale that *discriminates* at high levels of ability as well as very low levels of ability.

Relative to the 17-item ACE-III IIO hierarchy in Chapter 5 the items of the 11 ACE-R item IIO hierarchy in Chapter 7 did not include high *difficulty* items. The most *difficult* item in this hierarchy was 'recognition' with a mean value of 0.51 whereas the level of impairment assessed by the hierarchy in Chapter 5 extended to a mean score of 0.33 for 'name and address recall'.

While the Mini-ACE appears to be the stronger of the two brief versions of the ACE the Short ACE-R was derived using a sample that was more representative of the general dementia population. The sample used to develop the Mini-ACE comprised a preponderance of patients with progressive primary aphasia. The use of such a sample for scale development raises concerns regarding the generalizability of the scale.

Both short versions of the ACE developed in using Mokken scaling analysis were based on analyses of two different versions of the ACE; the Short ACE-R was derived from analysis of data from the ACE-R and the Mini-ACE was developed using ACE-III data. ACE-R items identified as poorly performing were changed or removed in the development

CHAPTER 12: DISCUSSION

of the ACE-III ('follow written command-close eyes' was removed, the phrases for repetition were changed, 'naming 1' and 'naming 2' were amalgamated to form one item, 'draw intersecting pentagons' was replaced by 'draw intersecting infinity loops' and 'write a sentence' became 'write two sentences'). These changes resulted in a scale with two fewer items in the ACE-III. The reduction of item numbers reduces the number of item ordering comparisons in the assessment of IIO, which could result in fewer violations. The modifications to the other ACE-R items could also have increased the scalability of the ACE-III items.

The development of the Mini-ACE in Chapter 5 was based on the analysis of ACE-III data. All other analyses of current cognitive impairment were based on the ACE-R, the predecessor of the ACE-III. The ACE-III performed better than the ACE-R in terms of the number of items retained in the final hierarchical Mokken scale. The analysis of ACE-III data in Chapter 5 from which the Mini-ACE was derived found 17 items meeting Mokken scaling criteria inclusion in a hierarchical scale. Across all analyses of the ACE-R fewer items were retained with hierarchies ranging from 10 to 14 items.

Further analyses of the ACE-III are required to determine if the item removals and alteration of ACE-R items in the development of this version of the scale are the cause of the superior performance of this scale. To determine the influence of these changes on the results of the ACE-III analysis the analysis of ACE-R data in Chapters 4 and 6 could be replicated without the superfluous ACE-R items. Furthermore, this analysis would also test the performance of the scale with two fewer items to determine the effect of both the change in item content, range of *difficulty* and number of items analysed on the results.

As discussed in Chapter 5 the Mini-ACE developed from analysis of the ACE-III was subsequently validated using data from the ACE-R. The validation of the Mini-ACE carried

CHAPTER 12: DISCUSSION

out by Hsieh et al. (2015) also used ACE-R data. The items of the Mini-ACE are included in both full scale versions of the ACE. Therefore as both validation analyses isolated the scores of the pertinent items from the ACE-R the differences between the ACE-R and ACE-III would not influence these analyses.

However as the ACE-R and ACE-III differ slightly in length, administration order and item content it is possible that changes to other items (not included in the Mini-ACE) could affect performance on the Mini-ACE items indirectly. In the ACE-III the fluency items are assessed earlier in the test following ‘three item recall’, ‘follow written command-close eyes’ was removed, the *difficulty* of ‘syntactical comprehension’ was increased by raising the syntactical complexity of the commands, the assessment of sentence writing increased in *difficulty* from writing one to two sentences, the phrases for repetition were changed, naming 1 and 2 were combined to assess the ability to name all objects within the one item, ‘draw overlapping pentagons’ was replaced by ‘draw intersecting infinity loops’. While most of these changes were small and may not have any influence on the performance on the items of the Mini-ACE one alteration in the item ordering could potentially have an effect on a patient’s performance on one of the Mini-ACE items. In the ACE-R the ‘name and address learning’ occurs earlier in the test, prior to the assessment of verbal fluency. In the ACE-III the verbal fluency assessment precedes the ‘name and address learning’. The fluency items add at least two minutes (one minute per item). This means that there is a longer time delay between the name and address learning and subsequent name and address recall in the ACE-R than in the ACE-III. While the timespan between learning and recall in the ACE-III will be sufficient for many patients with dementia to suffer memory loss it is possible that the shorter difference in time in the ACE-III could increase the ability of some patients to retain the name and address details. To determine whether performance on the Mini-ACE items differs depending on the full length version administered further analyses could be performed to

CHAPTER 12: DISCUSSION

compare the item scores. Whether patients perform better on these same items when assessed using the Mini-ACE itself should also be assessed particularly as the score on the Mini-ACE can be derived from the ACE-R and ACE-III. If the increased testing burden of the longer scales affects patients' performance the comparison of Mini-ACE scores derived from the full scale with that of the brief version is not valid.

The Mini-ACE was conceived as a short alternative to the MMSE which due to changes in copyright is no longer freely available for clinical or research use. An additional benefit of the creation of a brief scale is the reduction in time required to complete and score the test.

Assessing the item content of both the Mini-ACE and the Short ACE-R in terms of the length of time each would take reveals some differences. While both scales include 'name and address learning' and one of the two orientation items which would take the same length of time, the inclusion of 'verbal fluency-animal' and 'draw a clock' in the Mini-ACE would take longer than the counterparts of these items in the Short ACE-R ('semantic comprehension' and 'identify fragmented letters'). Furthermore the inclusion of 'draw a clock' in the Mini-ACE requires the use of a pencil and paper which limits the use of this item in certain clinical groups such as stroke patients or those with arthritis, or severe pain. Performance on this item could be compounded by impaired manual dexterity. This is less relevant in the full version of ACE-R or ACE-III where performance on the other perceptual items; 'count dot arrays' and 'identify fragmented letters' can be applied to differentiate between visuospatial impairment and motor symptoms. However in the short versions where only one item forms the basis for domain performance the potential limitations to 'draw a clock' must be considered. Therefore in terms of practical considerations the items of the Short ACE-R are quicker to administer and applicable to a wider range of patients.

CHAPTER 12: DISCUSSION

However, in terms of item content the Mini-ACE appears to have the advantage particularly in the visuospatial, language and memory domains. 'Draw a clock' is a valuable item in the Mini-ACE as it assesses visuospatial construction, abstract conceptualisation along with verbal and numeric memory (Koretz & Moore, 2001). The assessment of visuospatial functioning in the Short ACE-R with 'identify fragmented letters', as a measure of perceptual ability is unaffected by apraxia but is less comprehensive in terms of assessment than 'draw a clock'.

With regards to the assessment of language the items of both scales differ; the Mini-ACE includes 'verbal fluency-animal' and the Short ACE-R includes 'semantic comprehension'. 'Verbal fluency-animal' involves the timed associative exploration, retrieval and generation of words. The search for words is constrained by the semantic category of animals. The process of finding semantic extensions of the target of animals largely depends on the integrity of semantic associations (Rohrer, Salmon, Wixted & Paulsen, 1999). Therefore poor performance in the generation of animal names may be the result of semantic memory deficits and not executive dysfunction. This makes the language assessment in both scales more comparable. However the executive component of the verbal fluency item, the necessary organisation of verbal retrieval and recall along with self-initiation and inhibition of inappropriate responses, could contribute to greater variation in responses and differentiation between patients with this item.

Assessment of memory in the Mini-ACE using 'name and address learning' and assessing the subsequent retention of this information with 'name and address recall' is likely to detect earlier episodic memory impairments than use of 'name and address learning' and subsequent recognition of the information.

CHAPTER 12: DISCUSSION

Recall and recognition memory tests place different demands on prefrontal versus medial temporal lobe functioning which could contribute to different profiles of memory in each of the two scales. Practically speaking, the less *difficult* task of ‘recognition’ may be a less daunting memory assessment for patients who may be anxious about their memory performance.

As a brief assessment tool the Short ACE-R would not be used to make a differential diagnosis in isolation. However both the MMSE and Mini-ACE have been found to produce distinct cognitive profiles with the Mini-ACE demonstrating better differentiation within the memory and language domains (Hsieh et al., 2015).

Delayed episodic memory tests have been identified as good predictors of bv-FTD and AD diagnosis (Hornberger, Piguet, Graham, Nestor & Hodges, 2010) whereas performance on memory recognition has found to be the least predictive indicator of episodic memory differences between clinical groups (Hornberger & Piguet, 2012). Therefore due to the assessment of memory recognition in the Short ACE-R the scores on the scale may not provide distinctive profiles of cognitive impairment.

The Mini-ACE appears to be the more valuable scale in terms of the range of assessment, the ability of the items to *discriminate* between different degrees of cognitive impairment, the item content and potential for providing distinctive cognitive profiles which could be used to support a differential diagnosis.

12.9.2 Mini-NART

The Mini-NART derived from Mokken scaling analyses in Chapter 8 conformed to a formal hierarchy. The significance and implications of the development of this scale and the findings of this analysis goes beyond the creation of a new hierarchical scale however. This 23-item scale offers predictive accuracy effectively equal to that of the full scale. Importantly, the

CHAPTER 12: DISCUSSION

Mokken scaling analysis determined that of the 50 items within the NART 22 items did not contribute to the estimation of premorbid cognitive ability in the sample analyses.

The Wechsler Test of Adult Reading (WTAR, Wechsler, 2001) is a more recently developed measure co-developed and co-normed with the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997) for the estimation of premorbid intelligence. Like the NART, the WTAR assesses the pronunciation of irregular words.

The WTAR could be refined using Mokken scaling analysis as in the case with the NART by identifying and deleting or replacing poorly functioning items. Additionally a combined analysis of the 50 items of the NART and the 50 items of the WTAR could determine the hierarchical pattern of the items of each scale and could lead to an IIO hierarchical scale which could be compared to the short NART scale developed in Chapter 8. A hierarchical scale comprised of the most *discriminatory* items from both scales could potentially increase the predictive power of either scale in isolation.

12.10 General limitations

Over the course of my PhD studies I have identified some methodological issues and shortcomings. While some of these have been addressed earlier the more general limitations to these analyses and finding will be discussed here.

12.10.1 Sample size

While there are no strict guidelines for minimum sample size requirements for Mokken scaling Straat (2012) demonstrated how item scalability coefficients are related and inversely proportional to sample size in simulation study. Investigating the minimum required sample size for Mokken scaling analysis involves the assessment of the ability to correctly partition items into Mokken scales. In the absence of any analytical method for determining the

CHAPTER 12: DISCUSSION

minimum sample size Straat (2012) used a simulation study to assess the required sample size to correctly allocate items into Mokken scales using the automated item selection procedure by examining the effect of 16 different sample sizes (ranging for $N = 50$ to $N = 3500$) on the allocation of items into Mokken scales.

Using scalability coefficients from the ACE-R and ACE-III analyses as a barometer of adequate sample size as advocated by Straat (2012) and Watson et al. (2014) in the context of the current results does not demonstrate a wholly consistent pattern of low scalability in small samples. Table 12.3 presents the number of items with low scalability coefficients across all analyses of the ACE-R and ACE-III. While nine items were excluded from two samples with relatively small samples ($N = 137$, $N = 100$) analyses of some of the other small samples resulted in only one item with poor *discrimination* (which is indexed by the item scalability coefficient) ($N = 147$; $N = 122$; $N = 113$). While most items were excluded from the smaller samples and only two items were excluded from the largest samples ($N = 808$; $N = 539$) there were some inconsistencies to this pattern. There were other small samples ($N = 113$, $N = 122$, $N = 147$) where the items demonstrated good *discrimination* (with only one item excluded). In terms of *discrimination* the items performed best in the analyses in Chapter 5, particularly the results of the sensitivity analyses. These samples are the only ones in this thesis to include data from controls in addition to diagnostic groups. The only sample in this chapter not to include control data was the Mini-ACE development sample. Five items were excluded due to poor item *discrimination* in this sample. The pattern of low *discrimination* in the results of Chapter 5 suggests the addition of control data increases the item scalability. Based on the number of low *discrimination* items and the sample size of the ‘other frontotemporal lobe degenerative disorders’ ($N = 100$) I believe this sample size is too small.

CHAPTER 12: DISCUSSION

The examination of item scalability coefficients produced in the functional scale analyses raises some questions about the certainty with which the strength of scalability coefficients can be applied to reflect sample size adequacy. Despite the small sizes of some of the functional assessment samples (for example, the analyses of mixed AD VaD sample ($N=138$) and non-AD sample ($N=142$) in Chapter 10) all item scalability coefficients were greater than the lower bound threshold. One possible reason for the superior scalability of functional scales in comparison to the ACE is the presence of fewer items in the functional scales than in the cognitive scales which decreases the number of possible violations. Therefore the superior item performance of the functional items could be an artefact of the number of items under examination. However, it has also been identified that the better the data meet the assumptions of IRT the smaller the sample size required (Thissen, 2003). Based on this reasoning and scalability values it appears that the item response data from the Lawton IADL and PSMS scales easily meet IRT assumptions.

Straat's (2012) simulation study determined that $N > 1500$ is a safe sample size, particularly for the investigation of the dimensionality of the data for various lowerbounds c (e.g. $c=0.00$, $c=0.05$, $c=0.10$, $c=0.50$). This sample size requirement is beyond the scope of any of the analyses in this thesis. However, in the case of high quality items the procedure of partitioning of items in small samples performs well (Straat et al., 2014). For example, in the Mokken scaling analysis of a Chinese version of the Myocardial Infarction Dimensional Assessment Scale (MIDAS) in a relatively small sample of 180 (Watson, Wang, Ski & Thompson, 2012) the range of item scalability coefficients (0.39-0.60) indicate that in this case the sample is adequate based on the findings of the Straat et al. (2014) simulation study. An overview of the range of scalability coefficients of scale items analysed found the lowest range of H_i values in the ACE-R analyses in Chapter 6. The lowest and most restricted range of H_i within this chapter was found in the analysis of the ACE-R in the AD VaD sample

CHAPTER 12: DISCUSSION

(0.31-0.44). The results of the simulation study suggest that where the item scalability coefficient is close to the lowerbound threshold a larger sample size is required (Straat, van der Ark & Sijtsma, 2014). The other analyses in Chapter 6 also tended toward the lower end of item scalability; the full SDRIR sample (0.34-0.47), late-onset AD (0.33-0.57), early and late-onset AD (0.36-0.49). The range of item scalability coefficients for the functional scale analyses were considerably higher with $H_i=0.43$ the lowest item scalability coefficient found across the three chapters. The range of H_i were superior to that of the Watson et al. study (2012) which was found to have an adequate sample size on this basis (Straat et al., 2014). Therefore while the samples in some of the functional scale analyses were relatively small (e.g. $N=138$, $N=142$) the scalability coefficients appear to indicate adequate sample size.

From the results across these analyses and previous research on sample size requirements it appears that item scalability coefficients can act as a useful guide in the assessment of the adequacy of the sample size. For example, where many items fail to meet the minimum lowerbound threshold a larger sample is most likely required. Alternatively, where the analysis of a relatively small sample produces a good range of item scalability coefficients it is likely that the scale contains high quality items with the data easily meeting IRT assumptions.

It would be interesting to determine whether for each of the analyses of ACE-R data conducted here all items could be retained in a Mokken scale in larger samples and if so how large a sample would be required. This possibility casts some doubt over the reliability of some of the results of Mokken scaling of the ACE-R reported here. Perhaps several items excluded from some analyses would have been sufficiently *discriminatory* in a larger sample. Several items such as ‘recognition’, ‘name and address recall’, ‘naming 1’, ‘naming 2’, and ‘semantic comprehension’ were only excluded from one analysis for poor *discrimination*. While some items were consistently poor in terms of their scalability (‘count dot arrays’ and

‘repetition 1’ for example) some items excluded that were retained in the majority of the analyses were potentially eliminated from further analyses due to inadequate sample size. Further analyses are required to determine if these items display more consistent *discrimination* in larger samples.

12.10.2 Manipulation of lowerbound threshold

In exploring the data for multiple dimensions I did not adjust the value of the lowerbound threshold c . The process of incrementally increasing the value of c in 0.05 increments is advocated by Hemperker et al. (1995) and Meijer and Baneke (2001). Taking the reliability of the scales produced from the incremental adjustments of the lowerbound threshold into account a suitable balance between the number of reliable scales ($MS > 0.8$) formed and those with an inconsequentially small number of items (i.e. fewer than three items) can be found (Watson et al., 2014).

The process of manipulating the lowerbound threshold enables the identification of a well-considered lowerbound threshold value, which permits the effective identification of unidimensional scales (Hemker, Sijtsma & Molenaar, 1995). MSP5 for Windows is recommended for this analysis as it contains very convenient functions for assessing the influence of varying values of lowerbound c . It would be interesting to examine the reliability of the Mokken scales formed in these analyses at a lowerbound threshold value of 0.15 for example.

12.10.3 Significance of standard errors and confidence intervals

Standard errors must be considered to allow for valid assessment of scalability coefficients, as they are fundamental in the interpretation of the size of the effect of the estimated value (Kuijpers et al., 2013). Kuijpers et al. (2013) demonstrated how failure to take standard errors

CHAPTER 12: DISCUSSION

into account can lead to inaccurate inferences regarding scalability. For example, a scale with H of 0.53 would be considered a strong Mokken scale before taking a high standard error of 0.09 into account. This scale should no longer be considered strong as the standard error implies that the population value of H may well be lower than 0.53.

While standard errors for scalability coefficients were assessed in all analyses of the ACE-R and in the analysis of the Lawton IADL scale in Chapter 9 standard errors were not computed in the analyses of the Mini-ACE in Chapter 5, the NART in Chapter 8, the Lawton IADL scale in Chapter 10 or both PSMS and Lawton IADL scales in Chapter 11.

Standard errors for scalability coefficients for dichotomous items have been derived using marginal models (van der Ark, Croon & Sijtsma, 2008). All possible patterns of response must be assessed using this method, which presents a computational difficulty. In the context of the NART which 50 dichotomously scored items the number of possible response patterns is $2^{50}=1,125,899,906,842,624$. In the case of polytomously scored items such as the ACE-R where the number of possible response patterns = $6^5 \times 4^3 \times 8^4 \times 5^4 \times 2^6 \times 3^3 \times 11 = 2.421657^{e+16}$ the number is even greater. To solve the problem of the scope of these computations Kuijpers, van der Ark and Croon (2011) generalised the approach to scalability coefficients for polytomous items. Kuijpers et al., (2011) based their calculations of standard errors by assuming that the observed frequencies of the item response patterns would conform to a multinomial distribution. Prior to these developments standard errors could only be calculated for dichotomous items providing the number of items was small.

However it was only recently that these methods of standard error computation were implemented in software packages, which limited the calculation and reporting of standard errors considerably. Kuijpers et al. (2013) extended the application of marginal modelling approach for calculating standard errors of scalability coefficients to large numbers of items

CHAPTER 12: DISCUSSION

and to polytomous items and crucially this method was incorporated into the Mokken scaling package in R (van der Ark, 2012). For a detailed description of the derivation of standard errors for large item sets and polytomous items see Kuijpers et al. (2013).

I did not calculate confidence intervals (CIs) for items and item pairs. While in some of the analyses standard errors were examined these were not used to calculate 95% CIs. This is of particular concern given the small sample sizes and high standard errors of items in some cases. The lowerbound 95% CI for items should not cross the lowerbound criteria for items and scales (0.3) and the lowerbound 95% CI for item pairs should not fall below 0 (Kuijpers, van der Ark & Croon, 2013). The lack of this information has considerable implications for the Mokken scales formed in these analyses. It is possible that some items seemingly meeting Mokken scaling criteria have lowerbound 95% CIs including the value 0.3.

IRF pairs were not plotted in every analysis. Plotting IRFs for item pairs is recommended to visually ascertain whether any of the scale items are close in proximity or even intersecting which would violate IIO. These plots can also determine whether any of the items are located at some distance from the other items. It is possible that the presence of such outlying items can lead to the interpretation of apparent IIO despite some potential intersection of other items (Watson et al., 2014).

IRF pairs were not plotted in any of the functional scale analyses. Examining the mean scores across these analyses does not immediately identify any potential sources of the high H^T values. ‘Telephone use’ is the least *difficult* Lawton IADL item in Chapter 9 with a mean value in this sample of 0.92. The second least *difficult* item; ‘housework’ has a mean score of 0.77, which creates a gap in the measurement of functional impairment of 0.15. While this does not immediately provoke any cause for concern in the absence of IRF pair

plots this item could be removed to determine whether this outlying item is having an effect on IIO.

12.10.4 Local stochastic independence

In Chapter 4 some ACE-R items were identified as possible sources of violations of local stochastic independence (LSI). LSI as one of the assumptions of Mokken scaling analysis is a crucial property to consider. Related to the concept LSI is local item dependence (LID). LID describes the enduring dependence in the data beyond what the model accounts for (Balazs & deBoeck, 2006). LID can arise from several testing conditions and formats such as practice, fatigue, external help, hastiness, item or response format and items where an explanation of the previous item is required for example (Yen, 1993). There are two main classifications of sources of item dependencies; *item chains*, where the response to an item can depend on the response to a previous item (Thissen et al., 1992) and *item overlap*, where the items include very similar concepts. Item chains can occur where the answer to one item must be applied to respond to another item or where there is a logical connection between two or more items, for example in an assessment of physical ability where cumulative running distances are assessed; the ability to run a five miles is predicated on the ability to run one mile. Therefore anyone endorsing the ‘run five miles’ item will logically have also endorsed the ‘run one mile’ item.

‘Name and address learning’, ‘name and address recall’ and ‘recognition’ form an *item chain* as theoretically or otherwise it is not possible to respond to either ‘name and address recall’ or ‘recognition’ without having been exposed to ‘name and address learning’. ‘Three item registration’ and ‘3-item recall’ form another such *item chain*. Therefore, while the results of the analysis excluding these items do not confirm this, it is highly likely that

CHAPTER 12: DISCUSSION

these items represent *item chains*, which are potential sources of LSI violations (Balazs & deBoeck, 2006).

In Chapter 6 an additional analysis of the ACE-R data was carried out excluding items identified in Chapter 4 as potential sources of violations of local stochastic independence ('3-item registration', '3-item recall', 'name and address learning', 'name and address recall' and 'name and address recognition'). The results from this further analysis did not provide evidence for the high values of H^T being driven by items violating LSI. While the H^T values are very high they do not appear to be driven by violations of LSI.

It is possible that some items of the ACE-R also display item overlap due to potentially overlapping processes underlying item performance; the three repetition items for example and 'draw a cube' and 'draw intersecting pentagons'. Within the framework of regression these types of items could be considered to demonstrate multicollinearity. These items raise concern for LSI and item redundancy. Whether LSI arises from the design of the items and tests, as in the case of the item chains in the ACE-R, or through the same underlying cognitive processes such as the potential overlapping items in the ACE-R the analysis of the scale should be replicated once the methods to estimate LSI become available (Straat, 2012).

Inspecting the items of the NART it is not likely that violations of LSI would have occurred. The items of the NART are irregular words, which the respondent attempts to pronounce correctly. The response to each item is assumed to be independent from all other responses. However, within the NART some items have religious origins (e.g. prelate, psalm, beatify) or are medical terms (e.g. puerperal, syncope) or are derived from French (e.g. façade, naïve, bouquet), which could provide some linking between the responses to these items, but are very unlikely to demonstrate sufficient overlap as to violate LSI.

CHAPTER 12: DISCUSSION

The items of the functional assessment scales analysed in this thesis (Lawton IADL and PSMS scales) assessing physical abilities and behaviours are not likely to violate LSI as the items are quite distinct and independent; it is not logical that shopping performance is predicated on the ability to prepare food or vice versa. Analyses of some alternative measures of functional ability such as the MOS 36-item short-form health survey (Ware, Kosinski & Gandek, 2003) are however likely to involve violations of LSI due to the item dependencies implied within the scale. For example, the response to each of three items assessing cumulative walking distances ('walking more than one mile', 'walking several blocks' and 'walking one block') is dependent on or implied from the response on the other items. These items form an item chain based on their assessment of the ability to walk increasing distances; a respondent who endorses 'walking more than one mile' will therefore also endorse the other lesser walking distance items and a respondent who does not endorse 'walking one block' will not endorse any of the greater walking distances.

12.10.5 Items excluded from Mokken scales

Mokken scaling analysis involves an iterative process whereby ill-fitting items are removed one by one with the re-assessment of remaining items each time an item is excluded. Throughout the analyses in this thesis items were excluded due to their failure to meet the standards of the heuristic rules involved in the interpretation of values of the scalability coefficients of for violations of IIO.

Throughout the analyses items were excluded for one of two reasons: low scalability (poor *discrimination*) or IIO violations. The items removed due to poor *discrimination* may have led to slightly "stronger" scales according to Mokken scaling criteria however these exclusions may not always be in the best interest with regards to measurement. Excluding scale items can lead to the loss of information relevant to the measurement of a domain that

CHAPTER 12: DISCUSSION

may be relevant or valuable. It is important to consider the advantages and disadvantages of discarding any items from a scale. For example, does the increase in strength warrant the potential loss of information? Rather than automatically removing an item the information provided by the item should be carefully considered. Would the loss of this information diminish the insight into a specific or general characteristic of a measured domain? Or would the removal of the item compromise the validity? Would a different item designed to assess the same domain perform better? For example, the assessment of language skills may be better assessed without the dependence on hearing as in the case of the verbal repetition items.

Items identified for removal due to low *discrimination* such as ‘follow written command-close eyes’ demonstrate poor association to the measurement of the latent trait. In cases such as this the loss of item content may be more justified than the removal of an otherwise good item identified for exclusion due to violations of IIO (such as ‘name and address recall’ in the analysis of the full SDRIR sample in Chapter 6).

While items are identified for exclusion due to either poor scalability or violations of IIO it is important to consider the effect of the scale as a whole as well as the remaining items. This is particularly relevant to items violating IIO. This feature, although psychometrically desirable, can result in the exclusion of items that can affect the construct being assessed (Meijer & Egberink, 2012). It is important to consider the purpose and goal of the scale when making exclusions due to IIO violations.

In an analysis by of the Mental Health Inventory, a 38 item scale of psychological distress and well-being item pair plots revealed that the most *difficult* item ‘during the past month, did you think about taking your own life?’ was located at some distance from the other items at that range of assessment (Watson et al., 2014). The extreme nature of this item

CHAPTER 12: DISCUSSION

was found to be influencing the invariant item ordering. However, the authors could not remove this item assessing suicidal ideation as it contributes significant and valuable clinical information. Had this important marker of extreme distress been removed from the scale due to its responsibility for the apparent IIO in the scale important clinical information would be overlooked, which is not in accordance with the aim of the scale. This exemplifies the importance of keeping the objective of the scale in mind when exploring the data. While such items should not be excluded in practice, items violating IIO can be removed to assess the quality of the remaining items. This allows psychometric improvements, if necessary, to be made to the scale by the removal or modification of the remaining items.

From a practical point of view it is imperative to assess the value of each item removed from a scale as it may be of sufficient importance or relevance to warrant the development of new alternative items to assess this domain. These newly created items could then be assessed to determine if any improvement in measurement has been gained. This would be preferred than refining a scale to meet stringent IIO properties only to find that it can no longer assess the latent trait effectively due to the removal of key items.

In the analysis of the full SDRIR sample ($N=808$) in Chapter 6 12 items were excluded due to IIO violations. These items included ‘name and address recall’ and both fluency items. The inclusion of ‘name and address recall’ in the scale is important, as it appears to be particularly sensitive to AD (Mathuranath et al., 2000) and assesses the early stages of cognitive decline. For these reasons ‘name and address recall’ and other items assessing delayed recall play an important role within dementia assessment scales. Additionally, the assessment of verbal fluency is also valuable as it is included in the VLOM ($[(\text{verbal fluency} + \text{language})/(\text{orientation} + \text{memory})]$) ratio which is used to differentiate between Alzheimer’s disease and frontotemporal dementia based on the ratio of scores on verbal fluency plus language to orientation plus name and address recall (Mioshi et al., 2006).

CHAPTER 12: DISCUSSION

The absence of both fluency items would limit the degree of differentiation possible using the ACE measures. Therefore the removal of these items from a scale due to IIO violations is not necessarily advocated. While IIO is a crucial property of Mokken scales the exploration of IIO can be considered more theoretically informative in some cases rather than being of practical importance.

The decision to exclude items across this thesis should have been based on a better consideration of theoretical, practical and qualitative reasoning. For example the IIO violation of ‘name and address recall’ in the analysis of SDRIR participants in Chapter 6 due to IIO violations should perhaps have been disregarded due to the contribution of this item to early cognitive decline in Alzheimer’s disease. This item was in fact subsequently re-incorporated into Chapter 7’s development of the ACE-R due to its clinical and practical significance.

Just as it is important not to completely disregard the item content and significance within the scale, it is also important not to completely disregard the output of Mokken scaling analysis: a balance between quantitative and qualitative reasoning is required. While a wider consideration should be made the removal of all items can also be of value psychometrically if not always practically. For example, items can be removed to assess the psychometric properties and quality of the remaining scale items.

In the initial analyses in this thesis I accepted the arbitrary numerical values, cut-offs and thresholds outlined to denote how well items meet Mokken scaling criteria rather than a full evaluation and consideration of the each item identified for having fallen short of one of the analytic thresholds. Scales are designed to assess a latent construct and the items within them have been developed to fulfil the measurement of particular levels, domains or features

CHAPTER 12: DISCUSSION

of the latent trait. Therefore before any item is discarded the item content, validity, wording and variance should be assessed or at least considered.

Some items could have been removed from further analyses due to the low variance of response. This could have been the case with any of the very high or low *difficulty* items where performance was generally at floor or ceiling. An item demonstrating low variance of response distribution will have low covariance and correlation with the other scale items. This will affect the scalability of the item and increase the likelihood of the item demonstrating poor *discrimination*. Examining items of extreme levels of *difficulty* across these analyses and chapters provides some insight into this potential unjustified item removal. Had some of these items been analysed in samples representing greater disease severity perhaps they would have been retained. This could be explored in further analyses of data from samples with wider and differing range of disease severity.

Another oversight in the automatic exclusion of items based on quantitative values is the consideration of item wording and comprehensibility. An item with unusual item response patterns could be due to difficulties in the comprehension of the item. Should some respondents incorrectly interpret the wording or the item content the response are likely to be less closely associated with the latent trait. This would be reflected by poor item scalability. These examples of analytical shortcomings here also emphasise the potential for Mokken scaling analysis to identify and correct these measurement issues.

In the NART the item content of poor *discriminatory* items is less important. The 50 items of the NART were selected due to their violation of the typical grapheme-phoneme rules. Each of the 50 words was chosen to assess respondents' familiarity with the words as opposed to the ability to apply typical phonetic rules to decode the pronunciation of the word.

CHAPTER 12: DISCUSSION

Therefore the meaning of the item is irrelevant and all items are considered equal in terms of content.

Items identified as poorly *discriminatory* should be removed from the analysis to enable the assessment of the remaining scale items. Given that the items of the NART are designed to provide a score from which premorbid intelligence can be estimated the formation of a separate item cluster demonstrates that the NART includes items that are not measuring the latent trait. This means the full NART score is contaminated by ‘noise’ from unidentified traits. The three items—‘prelate’, ‘drachm’ and ‘topiary’—within this cluster need to be assessed to determine what these items are assessing if they are not contributing to the estimation of premorbid intelligence.

While generally the Lawton IADL and PSMS scales performed very well with most or all items retained in the majority of analyses considering the items excluded under the light of more substantive reasoning can offer some insights. In the analysis of the Lawton IADL scale in the full mixed sample in Chapter 10 one item (‘food preparation’) was removed from the ultimate hierarchical Mokken scale. ‘Food preparation’ was the second most *difficult* item in the scale for this sample and was therefore not ‘outlying’ in terms of *difficulty*. However the differential item functioning by gender is likely to have been responsible for some of the violations of item ordering in this mixed sample. Therefore while this analysis does not necessarily advocate the removal of this item it is certainly important to consider the effect of gender. Reformulations of some functional items are necessary to eliminate any gender bias.

12.11 Future directions and recommendations

The application of item response theory analyses to clinical assessments is becoming more frequent. As IRT can focus on patterns of item response and performance rather than having to rely on total summed scores this level of analyses has the potential to expand the level of understanding of cognitive and functional decline in dementia. For example, examining item responses over time may enable the identification of patterns of scores that are predictive of conversion from mild cognitive impairment to different types of dementia. In this way applying the results and findings of IRT analyses could potentially lead to the reduction of the size of patient sample required in clinical trials (Balsis et al., 2012). This example demonstrates the clinical significance of these analyses. A limitation to this predictive insight being gained in this thesis was the limited sample size and cross-sectional analyses. Having demonstrated the suitability of IRT to dementia assessment analyses the extension of the methods used across these chapters to the analyses of large numbers of longitudinal data is an important direction for future research.

12.11.1 Access to and analyses of large databases

IRT analyses require large data samples for reliable analysis. Within the UK researchers have access to a wealth of large-scale population cohort studies. These studies include data on a variety of phenotypic, biological and lifestyle variables throughout the life course for the assessment of health and wellbeing. Within these studies there is potential for comprehensive, thorough and extensive analysis of the determinants of cognitive decline and dementia.

The UK Biobank, primarily funded by the Medical Research Council and the Wellcome Trust, contains data from 500,000 people aged between 40-69 years of age. These study participants have undergone extensive assessment providing different types of data including blood pressure and lung function, blood, lifestyle and medical history information

CHAPTER 12: DISCUSSION

and importantly have consented to have their health followed. Web-based measurement scales are under development for addition to the biobank assessment including measures of cognitive function. These measures could be assessed using IRT methods.

This database is an important resource within the Medical Research Council Dementias Platform UK (DPUK). The DPUK will see the creation of the largest dementia research group in the world. The scale of the DPUK will allow for the acceleration of progress in research aimed to prevent or delay the onset and progression of dementia. The DPUK has the capability to provide a greater level of understanding of who is at risk of developing dementia and to determine why the rate and pattern of decline differs between people. Researchers will have access to data provided by over two million individuals over the age of 50 in addition to data from laboratory measures. The Platform aims to examine the causes of a range of types of dementia and neurodegenerative disorders including Parkinson's and Motor Neurone Disease. The scope of this resource would enable sufficient stratified samples to be analysed using IRT methods.

The array of different measurements and variables available mean Mokken scaling analysis could be performed on several different samples within the Platform. For example, stratified analyses by age, gender, diagnosis, co-morbidities, weight, lifestyle, early life factors, and genetic variation could be performed. This level of analysis would enable the identification of any differences in patterns of cognitive and functional decline between these samples.

The effects of existing drugs could be examined and the study of the efficacy of new medicines and therapies could be supported by IRT analyses. These analyses could be applied to develop a scale designed to detect small changes in cognition at specific levels of *difficulty*

and also to assess the degree of change in cognition across time with longitudinal IRT analyses.

Analyses of lifestyle choices and behaviours would allow for the detection of different item ordering between various samples. For example, does the pattern of decline differ between smokers and non-smokers or those who participate in sport or exercise and those who lead sedentary lives? IRT analyses would permit the investigation of these possibilities.

The study includes hearing discrimination data. Due to potential for variation in responses to several items of the ACE, the repetition items for example, due to poor hearing this information could be used to control for hearing problems. Participants with difficulty hearing could be removed from the analysis to provide a more reliable analysis of how the ability to repeat verbal information is related to cognitive decline.

12.11.2 Longitudinal analyses

Longitudinal IRT analyses would also allow the analysis of trajectories over time and disease progression for different patient groups. The results of the cross-sectional analyses in this thesis pertain to an ordering of cognitive impairments at one period in time. There are limitations to this level of analysis. The so called patterns of cognitive decline determined in these analyses are not patterns at all nor decline as there is no analysis of change in these analyses. The hierarchies of these studies provide an insight into the difficulty patients have in responding to various scale items. While there is value in these hierarchies, determining the sequence of decline over time would be a significant addition to the understanding of cognitive impairment in dementia and the different trajectories of various patient groups. These cross-sectional studies could form the first in a series of longitudinal analyses. The future analyses could determine whether the relationships between the items change and to monitor disease progression across different patient groups and diagnoses.

It is possible that repeat assessments over a couple of days could produce different results. An average across repeat testing could lead to more reliable assessment of a patient's current level of cognition. Concentration, frustration and attentional capacities will vary across assessments, which may reduce the reliability of scores examined here.

12.11.3 Person fit analysis

Another direction for further research is the additional assessment of person fit. It would be interesting to determine whether for specific persons the ordering is different. This can be examined by applying person-fit statistics. Most person-fit statistics are sensitive to the number of Guttman violations (Meijer, Niessen & Tendeiro, 2014). The more frequent these errors, where a more *difficult* item is answered correctly and a less *difficult* item is answered incorrectly, the more atypical or abnormal the item response pattern is. A suitable program for person-fit analysis is Perfit (Tendeiro & Tendeiro, 2014). Perfit examines the consistency of item response patterns and as such can detect invalid test scores by computing the normed number of Guttman violations. This package within R can be applied to calculate person-fit statistics and to interpret abnormal item response patterns. These atypical responses can then be removed from the dataset following inspection. Perfit also provides plots of nonparametric person response functions (PRFs; Emons, Sijtsma & Meijer, 2004; Sijtsma & Meijer, 2001) for item response patterns for dichotomous data. These plots provide graphical representation of the fit of an individual item response pattern by plotting the probability of responding to an item correctly as a function of the *difficulty* of the item.

Nonparametric person-fit statistics are valuable in the detection of unexpected patterns of response arising from a wide range of abnormal response behaviours. It should be noted that extreme person-fit scores imply that the item response pattern is very unlikely. Yet it is interesting to determine the underlying mechanism driving the unusual response pattern.

CHAPTER 12: DISCUSSION

Such response behaviours can include cheating, guessing or prior knowledge of the test items. These factors are unlikely to occur in the assessment of dementia due to the nature of the scales. For example, it would be very difficult for a participant to cheat in the ACE-R. However factors that could lead to aberrant test scores within these analyses could be the misinterpretation of test questions or an atypical form of cognitive decline. Atypical patterns of endorsing items of dementia assessment scales such as the ACE-R may provide interesting cognitive diagnostics.

12.12 Conclusion

This thesis drew attention to the use of Mokken scaling techniques in order to evaluate, enhance and develop cognitive and functional status scales used as outcome measures in dementia assessment. The analyses of measures of cognitive and functional ability serve to emphasise the value of hierarchical scales, which can be used to offer clinicians, researchers, families and caregivers predictive insights into the expected pattern of decline. This information has potential implications in the development of interventions, particularly with regards to functional decline where the adequate level of assistance can be offered to patients. Assessing diagnostic or gender differential item functioning permits the examination of whether the scores and item orderings have the same meaning across these relevant groups.

Two new scales were developed through Mokken scaling analyses in this thesis. These new hierarchical scales demonstrate the utility of Mokken scaling in identifying and selecting items for a shortened scale. The estimation and interpretation of item parameters across the analyses in this thesis demonstrate how IRT methods can be applied in test development, analysis and administration.

REFERENCES

References

- Adlam, A. L., Patterson, K., & Hodges, J. R. (2009). "I remember it as if it were yesterday": Memory for recent events in patients with semantic dementia. *Neuropsychologia*, 47(5), 1344-1351.
- Aggen, S.H., Neale, MC, Kendler, K.S. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, 35, 475-487.
- Alladi, S., Xuereb, J., Bak, T., Nestor, P., Knibb, J., Patterson, K., & Hodges, J. R. (2007). Focal cortical presentations of Alzheimer's disease. *Brain*, 130(10), 2636-2645.
- American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders, (DSM-5®). American Psychiatric Pub.
- Andersen, C. K., Wittrup-Jensen, K. U., Lolk, A., Andersen, K., & Kragh-Sørensen, P. (2004). Ability to perform activities of daily living is the main factor affecting quality of life in patients with dementia. *Health and quality of life outcomes*, 2(1), 52.
- Arbuckle JL (2009). *Amos (Version 18)*. Chicago, IL: SPSS.
- Ashford, J.W., Kolm, P., Colliver, J.A., Bekian, C., Hsu, L.N. (1989). Alzheimer patient evaluation and the Mini-Mental State: Item characteristic curve analysis. *Journal of Gerontology*, 44(5), 139-146.
- Balazs, K., & De Boeck, P. (2006). Detecting local item dependence stemming from minor dimensions: Interuniversity Attraction Pole statistics network [technical report].
- Balsis, S., Lowe, D.A., Bengtson, J.F. (2012). Shifts in measuring dementia. *Neurodegenerative Disease Management*, 2(5), 443-445.

REFERENCES

- Balsis, S., Unger, A. A., Bengtson, J. F., Geraci, L., & Doody, R. S. (2012). Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimer's & Dementia*, 8(4), 288-294.
- Barberger-Gateau, P., Fabrigoule, C., Helmer, C., Rouch, I., & Dartigues, J. F. (1999). Functional impairment in instrumental activities of daily living: an early clinical sign of dementia?. *Journal of the American Geriatrics Society*, 47(4), 456-462.
- Beardsall, L., & Brayne, C. (1990). Estimation of verbal intelligence in an elderly community: a prediction analysis using a shortened NART. *British Journal of Clinical Psychology*, 29(1), 83-90.
- Bengtson, J.F., Balsis, S., Garaci, L., Massman, P.J., Doody, R.S. (2009). How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease? *Dementia and Geriatric Cognitive Disorders*, 28(1), 63-69.
- Bengtson, S., Miller, T.M., Bengtson, J.F., Doody, R.S. (2011). Dementia staging across three different methods. *Dementia and Geriatric Cognitive Disorders*, 31(5), 328-333.
- Benson, A.D., Slavin, M.J., Tran, T.T., Petrella, J.R., Doraiswamy, P.M. (2005). Screening for early Alzheimer's disease: is there still a role for the Mini-Mental State Examination? *Primary Care Companion to the Journal of Clinical Psychiatry*, 7, 62-69.
- Bentler, P.M. (1980). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Birks, J (2006) Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database of Systematic Reviews* 2006, Issue 1. Art. No. : CD005593
- Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: a revision of the National Adult Reading Test. *The Clinical Neuropsychologist*, 3(2), 129-136.

REFERENCES

- Blessed, G.T., Roth, B.E., Tomlinson, M. (1968). The association between quantitative measures of dementia and of senile changes in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry*, 114, 797-811
- Bond, T., Ughrin, T. & Fox, C. (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates Publishers.
- Bonner, M. F., Ash, S., & Grossman, M. (2010). The new classification of primary progressive aphasia into semantic, logopenic, or nonfluent/agrammatic variants. *Current Neurology and Neuroscience Reports*, 10(6), 484-490.
- Borson, S., Scanlan, J. M., Chen, P., & Ganguli, M. (2003). The Mini-Cog as a screen for dementia: validation in a population-based sample. *Journal of the American Geriatrics Society*, 51(10), 1451-1454.
- Bourne J. *Improving Services and Support for People with Dementia*. National Audit Office, 2007.
- Boyle, P. A., Cohen, R. A., Paul, R., Moser, D., & Gordon, N. (2002). Cognitive and motor impairments predict functional declines in patients with vascular dementia. *International journal of geriatric psychiatry*, 17(2), 164-169.
- Boyle, P. A., Malloy, P. F., Salloway, S., Cahn-Weiner, D. A., Cohen, R., & Cummings, J. L. (2003). Executive dysfunction and apathy predict functional impairment in Alzheimer disease. *The American journal of geriatric psychiatry*, 11(2), 214-221.
- Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16(3), 201-213.
- Bright, P., Jaldow, E. & Kopelman, M. (2002). The National Adult Reading Test as a measure of premorbid intelligence: A comparison with estimates derived from demographic variables. *Journal of the International Neuropsychological Society*, 8, 847-854.

REFERENCES

- Brooks, B. R., Miller, R. G., Swash, M., & Munsat, T. L. (2000). World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and other Motor Neuron Disorders*, *1*(5), 293-299.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in multivariate analysis* (pp. 72–141). Cambridge, England: Cambridge University Press.
- Browne, M.W., & Cudeck, R. (1993), Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Brugnolo, A., Nobili, F., Barbieri, M. P., Dessi, B., Ferro, A., Girtler, N., ... & Rodriguez, G. (2009). The factorial structure of the mini mental state examination (MMSE) in Alzheimer's disease. *Archives of Gerontology and Geriatrics*, *49*(1), 180-185.
- Bucks, R. S., Ashworth, D. L., Wilcock, G. K., & Siegfried, K. (1996). Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. *Age and Ageing*, *25*(2), 113-120.
- Byrne, B. M., & van De Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107-132.
- Cahn-Weiner, D. A., Malloy, P. F., Boyle, P. A., Marran, M., & Salloway, S. (2000). Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. *The Clinical Neuropsychologist*, *14*(2), 187-195.
- Cahn-Weiner, D. A., Ready, R. E., & Malloy, P. F. (2003). Neuropsychological predictors of everyday memory and everyday functioning in patients with mild Alzheimer's disease. *Journal of geriatric psychiatry and neurology*, *16*(2), 84-89.

REFERENCES

- Carthery-Goulart, M. T., Knibb, J. A., Patterson, K., & Hodges, J. R. (2012). Semantic dementia versus nonfluent progressive aphasia: neuropsychological characterization and differentiation. *Alzheimer Disease & Associated Disorders*, *26*(1), 36-43.
- Chan, K. S., Kasper, J. D., Brandt, J., & Pezzin, L. E. (2012). Measurement equivalence in ADL and IADL difficulty across international surveys of aging: findings from the HRS, SHARE, and ELSA. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *67*(1), 121-132.
- Chandler, M.J., Lacritz, L.H., Cicerello, A.R., Chapman, S. B., Honig, L. S., Weiner, M. F., & Cullum, C. M. (2004). Three-word recall in normal aging. *Journal of Clinical and Experimental Neuropsychology*, *26*, 1128–1133.
- Chapman, L.J., Chapman, J.P. (1973). Problems in the measurement of cognitive deficit. *Psychological Bulletin*, *79*(6), 380–385.
- Chen, S. T., Sultzer, D. L., Hinkin, C. H., Mahler, M. E. & Cummings, J. L. (1998). Executive dysfunction in Alzheimer's disease: Association with neuropsychiatric symptoms and functional impairment. *Journal of Neuropsychiatry and Clinical Neurosciences*, *10*, 426-43.
- Chiu, Y., Fritz, S.L., Light, K.E. & Velozo, C.A. (2006). Use of item response analysis to investigate measurement properties and clinical validity of data for the dynamic gait index. *Journal of Physical Therapy*, *86*, 778-787.
- Clarke, D. E., van Reekum, R., Simard, M., Streiner, D. L., Conn, D., Cohen, T., & Freedman, M. (2008). Apathy in dementia: Clinical and sociodemographic correlates. *The Journal of Neuropsychiatry and Clinical Neurosciences*. *20*(3), 337-347.
- Coltheart, M., Patterson, K., & Marshall, J.C. (Eds.). (1987). *Deep dyslexia* (2nd ed.). New York: Routledge & Kegan Paul.

REFERENCES

- Cotrell, V., & Schulz, R. (1993). The perspective of the patient with Alzheimer's disease: a neglected dimension of dementia research. *The Gerontologist, 33*(2), 205-211.
- Crane, P.K., Gibbons, L.E., Jolley, L., van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44*(11), S115-S123.
- Crane, P.K., Narasimhalu, K., Gibbons, L.E., Pedrazz, O., Mehta, K.M., Tang, Y., Manly, J.J., Reed, B.R., Mungas, D.M. (2008). Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society, 14*(5), 746-759.
- Crawford, J. R. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford, W. McKinlay, & D. M. Parker (Eds.), *A Handbook of Neuropsychological Assessment* (pp. 21-49). London: Erlbaum.
- Crawford, J. R., Deary, I. J., Starr, J., & Whalley, L. J. (2001). The NART as an index of prior intellectual functioning: a retrospective validity study covering a 66-year interval. *Psychological Medicine, 31*(03), 451-458.
- Crawford, J. R., Moore, J. W., & Cameron, I. M. (1992). Verbal fluency: a NART-based equation for the estimation of premorbid performance. *British Journal of Clinical Psychology, 31*(3), 327-329.
- Crawford, J. R., Parker, D. M., & Besson, J. A. (1988). Estimation of premorbid intelligence in organic conditions. *The British Journal of Psychiatry, 153*(2), 178-181.
- Crawford, J. R., Parker, D. M., Allan, K. M., Jack, A. M., & Morrison, F. M. (1991). The Short NART: Cross-validation, relationship to IQ and some practical considerations. *British Journal of Clinical Psychology, 30*(3), 223-229.

REFERENCES

- Crawford, J. R., Parker, D. M., Stewart, L. E., Besson, J. A. O., & Lacey, G. (1989). Prediction of WAIS IQ with the National Adult Reading Test: Cross-validation and extension. *British Journal of Clinical Psychology, 28*(3), 267-273.
- Crawford, J.R., Stewart, L.E., Cochrane, R., Foulds, J., Besson., J.A.O, & Parker, D.M. (1989). Estimating premorbid IQ from demographic variables: A regression equation derived from a UK sample. *British Journal of Clinical Psychology, 28*(3), 275-278.
- Croon, M.A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315-331.
- Daltroy, L.H., Logigian, M., Iversen, M.D., & Lian, M.H. (1992). Does musculoskeletal function deteriorate in a predictable sequence in the elderly? *Arthritis & Rheumatism, 5*(3), 146-150.
- De Jong, A. & Molenaar, I.W. (1987) An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research, 21*, 137-149.
- de Morton, N.A., Keating, J.L., Davidson, M.(2008). Rasch analysis of the Barthel index in the assessment of hospitalized older patients after admission for an acute medical condition. *Archives of Physical Medicine and Rehabilitation, 89*, 641-647.
- Deary, I. J., Gow, A. J., Pattie, A., & Starr, J. M. (2012). Cohort profile: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology, 41*, 1576-1584.
- Deary, I. J., Watson, R., Booth, T., & Gale, C. R. (2013). Does cognitive ability influence responses to the Warwick-Edinburgh Mental Well-Being Scale? *Psychological Assessment, 25*(2), 313.

REFERENCES

- Deary, I.J., Gow, A.J., Taylor, M.D., Corley, J., Brett, C., Wilson, V.... Starr, J.M. (2007). The Lothian Birth cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatrics*, 7(1), 28.
- Deary, I.J., Whalley, L.J., Starr, J.M. (2009). *A Lifetime of Intelligence: Follow-Up Studies of the Scottish Mental Surveys of 1932 and 1947*. Washington, DC: American Psychological Association.
- DeBettignies, B.H., Mahurin, R.K., Pirozzolo, F.J.(1990). Insight for impairment in independent living skills in Alzheimer's disease and multi-infarct dementia. *Journal of Clinical and Experimental Neuropsychology*, 12(2), 355-363.
- DeJong, A., & Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research*, 21(2), 137-149.
- del Toro, C.M., Bislick, L.P., Comer, M., Velozo, C., Romero, S., Gonzalez Rothi, L.J., Kendall, D.L.(2011). Development of a short form of the Boston naming test for individuals with aphasia. *Journal of Speech, Language, and Hearing Research*, 54(4), 1089-1100.
- Delva, F., Edjolo, A., Pérès, K., Berr, C., Barberger-Gateau, P., & Dartigues, J. F. (2014). Hierarchical structure of the activities of daily living scale in dementia. *The Journal of Nutrition, Health & Aging*, 18(7), 698-704.
- Desai, A. K., Grossberg, G. T., & Sheth, D. N. (2004). Activities of Daily Living in patients with Dementia. *CNS Drugs*, 18(13), 853-875.
- Diesfeldt, H.F. (2004). Executive functioning in psychogeriatric patients: scalability and construct validity of the Behavioral Dyscontrol Scale (BDS) *International Journal of Geriatric Psychiatry*, 19(11), 1065-1073.

REFERENCES

- Dodge, H.H., Meguro, K., Ishii, H., Yamaguchi, S., Saxton, J.A., Ganguli, M. (2009). Cross-cultural comparisons of the Mini-mental State Examination between Japanese and U.S cohorts. *International Psychogeriatrics*, 21(1), 113-122.
- Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., ... & Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, 6(8), 734-746.
- Dykiert, D., & Deary, I. J. (2013). Retrospective validation of WTAR and NART scores as estimators of prior cognitive ability using the Lothian Birth Cohort 1936. *Psychological Assessment*, 25(4), 1361.
- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modelling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1), 5-18.
- Edelen, M.O, Thissen, D, Teresi, J.A., Kleinman, M., Ocepek-Welikson, K. (2006) Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. *Medical Care*, 44(11), S134-S142.
- Egberink, I. J., & Meijer, R. R. (2011). An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment*, 18, 201-212.
- Elfgrén, C. I., Ryding, E., & Passant, U. (1996). Performance on neuropsychological tests related to single photon emission computerised tomography findings in frontotemporal dementia. *The British Journal of Psychiatry*, 169(4), 416-422.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Psychology Press.

REFERENCES

- Emons, W.H.M., Meijer, R.R., Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: evaluating type-D personality and its assessment using item response theory. *Journal of Psychosomatic Research*, 63, 27-39.
- Emons, W.M., Sijtsma, K., & Meijer, R.R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioural Research*, 39, 1-35.
- Escobar, J.L., Burnam, A., Karno, M., Forsythe, A., Landsverk, J., Golding, M. (1986). Use of the Mini-Mental State Examination (MMSE) in a community population of mixed ethnicity. *The Journal of Nervous and Mental Disease*, 174, 607-614.
- Fang, R., Wang, G., Huang, Y., Zhuang, J. P., Tang, H. D., Wang, Y., ... & Ren, R. J. (2013). Validation of the Chinese version of Addenbrooke's Cognitive Examination-Revised for screening mild Alzheimer's disease and mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, 37, 223-231.
- Fernandez-Blazquez, M.A., Ruiz-Sanchez de Leon, J.M., Lopez-Pina, J.A., Llanero-Lurqu, M., Montenegro-Pena, M., Montejo-Carrasco, P. (2012). A new shortened version of the Boston Naming Test for those aged over 65: An approach from item response theory. *Revista de Neurologia*, 55(7), 399-407.
- Fieo, R. A., Austin, E. J., Starr, J. M., & Deary, I. J. (2011). Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC Geriatrics*, 11(1), 42.
- Fieo, R., Watson, R., Deary, I. J., & Starr, J. M. (2010). A revised activities of daily living/instrumental activities of daily living instrument increases interpretive power: Theoretical application for functional tasks exercise. *Gerontology*, 56(5), 483-490.
- Fillenbaum, G. G. (2013). Multidimensional functional assessment of older adults: The Duke Older Americans Resources and Services procedures. Psychology Press.

REFERENCES

- Fillenbaum, G.G., Wilkinson, W.E., Welsh, K.A., Mohs, R.C. (1994). Discrimination between stages of Alzheimer's disease with subsets of Mini-Mental State Examination items. An analysis of Consortium to Establish a Registry for Alzheimer's Disease data. *Archives of Neurology*, 51, 916–921
- Finch, M., Kane, R.L., & Philip, I.(1994). Developing a new metric for ADLs. *Journal of American Geriatric Society*, 43, 877-884.
- Finlayson, M., Mallinson, T., & Barbosa, V. M. (2005). Activities of daily living (ADL) and instrumental activities of daily living (IADL) items were stable over time in a longitudinal study on aging. *Journal of clinical epidemiology*, 58(4), 338-349.
- Fisher, A. G., & Jones, K. B. (1999). Assessment of motor and process skills (p. 2006). Fort Collins, CO: Three Star Press.
- Fisher, W.P. Measurement-related problems in functional assessment. (1993). *American Journal of Occupational Therapy*, 47, 331-338.
- Fisher, W.P., & Fisher, A.G. (1993). Applications of Rasch analysis to studies in occupational therapy. *Physical Medicine and Rehabilitation Clinics of North America*, 4, 551-569.
- Flanagan, E. C., Tu, S., Ahmed, S., Hodges, J. R., & Hornberger, M. (2014). Memory and orientation in the logopenic and nonfluent subtypes of primary progressive aphasia. *Journal of Alzheimer's Disease*, 40(1), 33-36.
- Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 57(5), S275-S284.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3), 189-198.

REFERENCES

- Force, A P.T. (1998). Guidelines for the evaluation of dementia and age-related cognitive decline. Washington, DC: American Psychological Association.
- Fraley, R.C., Waller, N.G., Brennan, K.A. (2000) An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78, 350-365.
- Galasko, D. (1998). An integrated approach to the management of Alzheimer's disease: assessing cognition, function and behaviour. *European Journal of Neurology*, 5(S4), S9-S17.
- Galasko, D., Bennett, D., Sano, M., Ernesto, C., Thomas, R., Grundman, M., & Ferris, S. (1997). An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. *Alzheimer Disease & Associated Disorders*, 11, 33-39.
- Galasko, D., Edland, S.D., Morris, J.C., Clark, C., Mohs, R., Koss, E. (1995). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Part XI. Clinical milestones in patients with Alzheimer's disease followed over 3 years. *Neurology*, 45, 1451-1455.
- Galton, C. J., Patterson, K., Xuereb, J. H., & Hodges, J. R. (2000). Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain*, 123(3), 484-498.
- Gaugler, J. E., Ascher-Svanum, H., Roth, D. L., Fafowora, T., Siderowf, A., & Beach, T. G. (2013). Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database. *BMC geriatrics*, 13(1), 137.
- Gearing, M., Mirra, S. S., Hedreen, J. C., Sumi, S. M., Hansen, L. A., & Heyman, A. (1995). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part X. Neuropathology confirmation of the clinical diagnosis of Alzheimer's disease. *Neurology*, 45(3), 461-466.

REFERENCES

- General Register Office (1951) "Classification of Occupations, 1950", H.M.S.O., London.
- Gibbons, L. E., van Belle, G., Yang, M., Gill, C., Brayne, C., Huppert, F. A., ... & Larson, E. (2002). Cross-cultural comparison of the Mini-Mental State Examination in United Kingdom and United States participants with Alzheimer's disease. *International Journal of Geriatric Psychiatry, 17*(8), 723-728.
- Gibbons, R. D., Clark, D. C., VonAmmon Cavanaugh, S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of psychiatric research, 19*(1), 43-55.
- Gorno-Tempini, M. L., Brambati, S. M., Ginex, V., Ogar, J., Dronkers, N. F., Marcone, A., ... & Miller, B. L. (2008). The logopenic/phonological variant of primary progressive aphasia. *Neurology, 71*(16), 1227-1234.
- Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., ... & Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology, 76*(11), 1006-1014.
- Graf, C. (2009). The Lawton instrumental activities of daily living (IADL) scale. *The Gerontologist, 9*(3), 179-186.
- Graves, R.E., Bezeau, S.C., Fogarty, J., Blair, R. (2004). Boston naming test short forms: a comparison of previous forms with new item response theory based forms. *Journal of Clinical Experimental Neuropsychology, 26*(7), 891-902.
- Gregory, C.A., Orrell, M., Sahakian, B., Hodges, J.R. (1997). Can frontotemporal dementia and Alzheimer's disease be differentiated using a brief battery of tests? *International Journal of Geriatric Psychiatry, 12*, 375-383.
- Grossman M, Mega M, Cummings JL, Joynt RJ, Griggs RC. In: Clinical Neurology. Baker AB, Joynt RJ, editors. Lippincott Williams and Wilkins; Philadelphia: 2004.

REFERENCES

- Grossman, M., Mickanin, J., Onishi, K., Hughes, E., D'Esposito, M., Ding, X. S., ... & Reivich, M. (1996). Progressive nonfluent aphasia: language, cognitive, and PET measures contrasted with probable Alzheimer's disease. *Journal of Cognitive Neuroscience*, 8(2), 135-154.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 139-150.
- Guttman, L. The basis for scalogram analysis. In *Studies in social psychology in World War II: Measurement and Predication. Volume 4*. Edited by Stouffer SA, Guttman LA, Suchman FA, Lazarfeld PF, Star SA, Clausen JA. Princeton: Princeton University Press; 1950, 60-90
- Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing, Boston.
- Hambleton, R.K., Swaminathan, H., Rogers, H.J. *Fundamentals of Item Response Theory (Measurement Methods for the Social Science)*. Newbury Park, CA: Sage; 1991
- Handels, R. L., Wolfs, C. A., Aalten, P., Verhey, F. R., & Severens, J. L. (2013). Determinants of care costs of patients with dementia or cognitive impairment. *Alzheimer Disease & Associated Disorders*, 27(1), 30-36
- Harper, D. G., Stopa, E. G., Kuo-Leblanc, V., McKee, A. C., Asayama, K., Volicer, L., ... & Satlin, A. (2008). Dorsomedial SCN neuronal subpopulations subserve different functions in human dementia. *Brain*, 131(6), 1609-1617.
- Harrison, J.E.(2007) Measuring cognitive change in Alzheimer's disease clinical drug trials. *Journal of Nutrition Health and Aging*, 11(4), 327.
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38(9 Suppl), I128.

REFERENCES

- Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous mokken IRT model. *Applied Psychological Measurement, 19*(4), 337-352.
- Hill, C.D., Edwards, M.C., Thissen, D., Langer, M.M., Wirth, R.J., Burwinkle, T.M., Varni, J.W. (2007). Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Medical Care, 45*(5 Suppl. 1), S39-47.
- Hinton, L., Franz, C., Friend, J. (2004). Pathways to dementia diagnosis: Evidence for cross-ethnic differences. *Alzheimer Disease and Associated Disorders, 18*, 134–144.
- Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain: a journal of neurology, 115*, 1783-1806
- Hohl, U., Grundman, M., Salmon, D.P., Thomas, R.G., Thal, L.J. (1999). Mini-Mental State Examination and Mattis Dementia Rating Scale performance differs in Hispanic and non-Hispanic Alzheimer's disease patients. *Journal of the International Neuropsychological Society, 5*, 301-307.
- Hokoishi, K., Ikeda, M., Maki, N., Nomura, M., Torikawa, S., Fujimoto, N., Fukuhara, R., Komori, K. & Tanabe, H. (2001). Interrater reliability of the Physical Self-Maintenance Scale and the Instrumental Activities of Daily Living Scale in a variety of health professional representatives. *Aging & Mental Health, 5*(1), 38-40.
- Holland, P.W., Wainer, H. *Differential item functioning*. Hillsdale, NJ: Erlbaum; 1993.
- Hornberger, M., & Piguet, O. (2012). Episodic memory in frontotemporal dementia: a critical review. *Brain, 135*(3), 678-692.

REFERENCES

- Hornberger, M., Piguet, O., Graham, A. J., Nestor, P. J., & Hodges, J. R. (2010). How preserved is episodic memory in behavioral variant frontotemporal dementia?. *Neurology, 74*(6), 472-479.
- Hsieh, H., McGrory, S., Leslie, F., Dawson, K., Ahmed, S., Butler, C. R., ... & Hodges, J. R. (2015). The Mini-Addenbrooke's Cognitive Examination: A New Assessment Tool for Dementia. *Dementia and Geriatric Cognitive Disorders, 39*(1-2), 1-11.
- Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., & Hodges, J. R. (2013). Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 36*(3-4), 242-250.
- Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In *Essays on item response theory* (pp. 221-244). Springer New York.
- Ideno, Y., Takayama, M., Hayashi, K., Takagi, H., Sugai, Y. (2012) Evaluation of a Japanese version of the Mini-Mental State Examination in elderly persons. *Geriatrics & Gerontology International, 12*, 310–316.
- Iliffe, S., Manthorpe, J., & Eden, A. (2003). Sooner or later? Issues in the early diagnosis of dementia in general practice: a qualitative study. *Family Practice, 20*(4), 376-381.
- Jacobs, D., Sano, M., Marder, K., Bell, K., Bylsma, F., Lafleche, G., ... & Stern, Y. (1994). Age at onset of Alzheimer's disease Relation to pattern of cognitive dysfunction and rate of decline. *Neurology, 44*(7), 1215-1215.
- Jefferies, K., & Agrawal, N. (2009). Early-onset dementia. *Advances in Psychiatric Treatment, 15*(5), 380-388.
- Jefferson, A. L., Cahn-Weiner, D., Boyle, P., Paul, R. H., Moser, D. J., Gordon, N., & Cohen, R. A. (2006). Cognitive predictors of functional decline in vascular dementia. *International Journal of Geriatric Psychiatry, 21*(8), 752-754.

REFERENCES

- Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V., & Stewart-Brown, S. (2001). Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *Journal of Neurology, Neurosurgery & Psychiatry, 71*(2), 220-224.
- Jette, A.M., Haley, S.M., Coster, W.J., Kooyoomjian, J.T., Levenson, S., Heeren, T., Ashba, J. (2002). Late life function and disability instrument: I. Development and evaluation of the disability component. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 57*, 209-216.
- Jonsson, M., Edman, Å., Lind, K., Rolstad, S., Sjögren, M., & Wallin, A. (2010). Apathy is a prominent neuropsychiatric feature of radiological white-matter changes in patients with dementia. *International Journal of Geriatric Psychiatry, 25*(6), 588-595.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.
- Kaplan, E., Goodglass, H., & Weintraub, S. (1983). Boston Naming Test. Philadelphia: Lea & Febiger
- Karantzoulis, S., & Galvin, J. E. (2011). Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Review of Neurotherapeutics, 11*(11), 1579-1591.
- Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. *Jama, 185*(12), 914-919.
- Kaufers, D. I., Cummings, J. L., Christine, D., Bray, T., Castellon, S., Masterman, D., ... & DeKosky, S. T. (1998). Assessing the impact of neuropsychiatric symptoms in Alzheimer's disease: the Neuropsychiatric Inventory Caregiver Distress Scale. *Journal of the American Geriatrics Society, 46*(2), 210-215.

REFERENCES

- Kempen, G. I. J. M., & Suurmeijer, T. P. (1990). The development of a hierarchical polychotomous ADL-IADL scale for noninstitutionalized elders. *The Gerontologist*, 30(4), 497-502.
- Kempen, G.I.J.M., Myers, A.M., & Powell, L.E. (1995). Hierarchical structure in ADL and IADL: analytical assumptions and applications for clinicians and researchers. *Journal of Clinical Epidemiology*, 48(11), 1299-1305.
- Kertesz, A., McMonagle, P., Blair, M., Davidson, W., & Munoz, D. G. (2005). The evolution and pathology of frontotemporal dementia. *Brain*, 128(9), 1996-2005.
- Klatka, L. A., Schiffer, R. B., Powers, J. M., & Kazee, A. M. (1996). Incorrect diagnosis of Alzheimer's disease: a clinicopathologic study. *Archives of Neurology*, 53(1), 35-42.
- Kline, R.B. (2005). Principles and practice of structural equation modelling. New York, NY: Guilford Press.
- Knibb, J. A., Xuereb, J. H., Patterson, K., & Hodges, J. R. (2006). Clinical and pathological characterization of progressive aphasia. *Annals of Neurology*, 59(1), 156-165.
- Koedam, E. L., Lauffer, V., van der Vlies, A. E., van der Flier, W. M., Scheltens, P., & Pijnenburg, Y. A. (2010). Early-versus late-onset Alzheimer's disease: more than age alone. *Journal of Alzheimer's Disease*, 19(4), 1401-1408.
- Koretz, B. K., & Moore, A. A. (2001). Assessment of the geriatric patient: a practical approach. *Journal of Clinical Outcomes Management*, 8(7), 35-40.
- Korner, A., Brogaard, A., Wissum, I., Petersen, U. (2012). The Danish version of the Baylor Profound Mental State Examination. *Nordic Journal of Psychiatry*, 66(3), 198-202.
- Kucukdeveci, A.A., Kutlay, S., Elhan, A.H., Tennant, A. (2005). Preliminary study to evaluate the validity of the mini-mental state examination in a normal population in Turkey. *International Journal of Rehabilitation Research*, 28(1), 77-79.

REFERENCES

- Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*(1), 42-69.
- Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*(1), 42-69.
- Landes, A. M., Sperry, S. D., & Strauss, M. E. (2005). Prevalence of apathy, dysphoria, and depression in relation to dementia severity in Alzheimer's disease. *The Journal of neuropsychiatry and clinical neurosciences, 17*(3), 342-349.
- Langenbucher, J. W., Labouvie, E., Martin, C. S., Sanjuan, P. M., Bavly, L., Kirisci, L., & Chung, T. (2004). An Application of Item Response Theory Analysis to Alcohol, Cannabis, and Cocaine Criteria in DSM-IV. *Journal of Abnormal Psychology, 113*(1), 72-80.
- Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist, 9*(3), 179-186.
- Lázaro Alquézar, A., Rubio Aranda, E., Sánchez Sánchez, A., & García Herrero, J. C. (2007). Functional capacity for daily living activities among senior citizens attending Community Centers in the city of Zaragoza, Spain, 2005. *Revista Española de Salud Pública, 81*(6), 625-636.
- Lechowski, L., de Stampa, M., Denis, B., Tortrat, D., Chassagne, P., Robert, P., ... & Vellas, B. (2007). Patterns of loss of abilities in instrumental activities of daily living in Alzheimer's disease: the REAL cohort study. *Dementia and Geriatric Cognitive Disorders, 25*(1), 46-53.
- Lechowski, L., Van Pradelles, S., Le Crane, M., d'Arailh, L., Tortrat, D., Teillet, L., & Vellas, B. (2010). Patterns of loss of basic activities of daily living in Alzheimer

REFERENCES

- patients: A cross-sectional study of the French REAL cohort. *Dementia and Geriatric Cognitive Disorders*, 29(1), 46-54.
- Letz, R., DiIorio, C. K., Shafer, P. O., Yeager, K. A., Henry, T. R., & Schomer, D. L. (2003). A computer-based reading test for use as an index of premorbid general intellectual level in North American English-speaking adults. *Neurotoxicology*, 24(4), 503-512.
- Libon, D. J., Massimo, L., Moore, P., Coslett, H. B., Chatterjee, A., Aguirre, G. K., ... & Grossman, M. (2007). Screening for frontotemporal dementias and Alzheimer's disease with the Philadelphia Brief Assessment of Cognition: a preliminary analysis. *Dementia and Geriatric Cognitive Disorders*, 24(6), 441-447.
- Ligtvoet, R. Essays on invariant item ordering. GildeprintDrukkerijen, Enschede (2010)
- Ligtvoet, R., van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, 76(2), 200-216.
- Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, 70, 578-595.
- Lillo, P., & Hodges, J. R. (2009). Frontotemporal dementia and motor neurone disease: overlapping clinic-pathological disorders. *Journal of Clinical Neuroscience*, 16(9), 1131-1135.
- Lillo, P., Garcin, B., Hornberger, M., Bak, T. H., & Hodges, J. R. (2010). Neurobehavioral features in frontotemporal dementia with amyotrophic lateral sclerosis. *Archives of Neurology*, 67(7), 826-830.
- Lillo, P., Mioshi, E., Zoing, M. C., Kiernan, M. C., & Hodges, J. R. (2011). How common are behavioural changes in amyotrophic lateral sclerosis?. *Amyotrophic Lateral Sclerosis*, 12(1), 45-51.

REFERENCES

- Lillo, P., Savage, S., Mioshi, E., Kiernan, M. C., & Hodges, J. R. (2012). Amyotrophic lateral sclerosis and frontotemporal dementia: a behavioural and cognitive continuum. *Amyotrophic Lateral Sclerosis*, *13*(1), 102-109.
- Lim, A., Tsuang, D., Kukull, W., Nochlin, D., Leverenz, J., McCormick, W., ... & Larson, E. B. (1999). Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. *Journal of the American Geriatrics Society*, *47*(5), 564-569.
- Lima-Silva, T. B., Bahia, V. S., Carvalho, V. A., Guimarães, H. C., Caramelli, P., Balthazar, M. L. F., ... & Yassuda, M. S. (2014). Direct and Indirect Assessments of Activities of Daily Living in Behavioral Variant Frontotemporal Dementia and Alzheimer Disease. *Journal of Geriatric Psychiatry and Neurology*, *28*(1), 19-26
- Lindeboom, R., Schmand, B., Holman, R., de Haan, R.J., Vermeulen, M. (2004). Improved brief assessment of cognition in aging and dementia. *Neurology*, *63*, 543-546.
- Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4), i-49.
- Lomen-Hoerth, C., Anderson, T., & Miller, B. (2002). The overlap of amyotrophic lateral sclerosis and frontotemporal dementia. *Neurology*, *59*(7), 1077-1079.
- Looi, J. C., & Sachdev, P. S. (1999). Differentiation of vascular dementia from AD on neuropsychological tests. *Neurology*, *53*(4), 670-670.
- Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc., 365 Broadway, Hillside, NJ 07642; 1980
- Lowenthal, M. F. *Lives in distress*. New York: Basic Books, 1964.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490.

REFERENCES

- Mackinnon, A., Ritchie, K., & Mulligan, R. (1999). The measurement properties of a French language adaptation of the National Adult Reading Test. *International Journal of Methods in Psychiatric Research*, 8(1), 27-38.
- Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English-and Spanish-speaking older adults. *Psychology and Aging*, 12(4), 718-725.
- Martin, A., Wiggs, C. L., Lalonde, F., & Mack, C. (1994). Word retrieval to letter and semantic cues: A double dissociation in normal subjects using interference tasks. *Neuropsychologia*, 32(12), 1487-1494.
- Massoud, F., Devi, G., Stern, Y., Lawton, A., Goldman, J. E., Liu, Y., ... & Mayeux, R. (1999). A clinicopathological comparison of community-based and clinic-based cohorts of patients with dementia. *Archives of Neurology*, 56(11), 1368-1373.
- Mathew R., Bak T.H., Hodges, J.R. (2012) Diagnostic criteria for corticobasal syndrome: A comparative study. *Journal of Neurology Neurosurgery & Psychiatry*, 83, 405-410.
- Mathias, J. L., Burke, J. (2009). Cognitive functioning in Alzheimer's and vascular dementia: A meta-analysis. *Neuropsychology*, 23(4), 411.
- Mathuranath, P. S., Nestor, P. J., Berrios, G. E., Rakowicz, W., & Hodges, J. R. (2000). A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. *Neurology*, 55(11), 1613-1620.
- Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R Package for Kernel Smoothing in Item Response Theory. *Journal of Statistical Software*, 58(6), 1-34.
- McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34,100-117.
- McGough, E. L., Kelly, V. E., Logsdon, R. G., McCurry, S. M., Cochrane, B. B., Engel, J. M., & Teri, L. (2011). Associations between physical performance and executive

REFERENCES

- function in older adults with mild cognitive impairment: gait speed and the timed “up & go” test. *Physical Therapy*, 91(8), 1198-1207.
- McGurn, B., Starr, J.M., Topfer, J.A., Pattie, A., Whiteman, M.C., Lemmon, H.A., ...Deary, I.J. (2004). Pronunciation of irregular words is preserved in dementia, validating premorbid IQ estimation. *Neurology*, 62(7), 1184-1186.
- McHorney, C.A., Cohen, A.S. (2000). Equating health status measures with item response theory: illustrations with functional status items. *Medical Care*, 38, 43-59.
- McHorney, C.A., Hayley, S.M., & Ware, J.E. (1997). Evaluation of the MOS SF-36 physical functioning scale (PF-10) II: comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology*, 50, 451-461.
- McKeith, I. G., Dickson, D. W., Lowe, J., Emre, M., O'brien, J. T., Feldman, H., ... & Yamada, M. D. L. B. (2005). Diagnosis and management of dementia with Lewy bodies Third report of the DLB consortium. *Neurology*, 65(12), 1863-1872.
- McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., ... & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 263-269.
- Meijer, R. R. (2010). A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale?. *Personality and Individual Differences*, 48(4), 502-503.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological methods*, 9(3), 354.

REFERENCES

- Meijer, R. R., & Egberink, I. J. (2012). Investigating Invariant Item Ordering in Personality and Clinical Scales Some Empirical Findings and a Discussion. *Educational and Psychological Measurement, 72*(4), 589-607.
- Meijer, R.R., & Tendeiro, J. N. (2014). The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us. Law School Admission Council, Research report, 14-03.
- Meijer, R.R., Niessen, A.S.M., & Tendeiro, J.N. (2014). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*.
- Meijer, R.R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*(3), 283-298.
- Mendez, M. F., Lauterbach, E. C. and Sampson, S. M. (2008). ANPA committee on research. An evidence-based review of the psychopathology of frontotemporal dementia: a report of the ANPA committee on research. *The Journal of Neuropsychiatry and Clinical Neuroscience, 20*, 130–149
- Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*(4), 308-312.
- Merritt, B. K., & Fisher, A. G. (2003). Gender differences in the performance of activities of daily living. *Archives of Physical Medicine and Rehabilitation, 84*(12), 1872-1877.
- Mesulam, M. (2001). Primary progressive aphasia. *Annals of Neurology, 49*(4), 425-432.
- Mesulam, M., Wicklund, A., Johnson, N., Rogalski, E., Léger, G. C., Rademaker, A., ... & Bigio, E. H. (2008). Alzheimer and frontotemporal pathology in subsets of primary progressive aphasia. *Annals of Neurology, 63*(6), 709-719.

REFERENCES

- Mion, M., Patterson, K., Acosta-Cabronero, J., Pengas, G., Izquierdo-Garcia, D., Hong, Y. T., ... & Nestor, P. J. (2010). What the left and right anterior fusiform gyri tell us about semantic memory. *Brain, 133*(11), 3256-3268.
- Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., Hodges, J.R. (2006). The Addenbrooke's Cognitive Examination Revised (ACE-R): A brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry, 21*, 1078-1085.
- Mioshi, E., Hsieh, S., Savage, S., Hornberger, M., & Hodges, J. R. (2010). Clinical staging and disease progression in frontotemporal dementia. *Neurology, 74*(20), 1591-1597.
- Mioshi, E., Kipps, C. M., & Hodges, J. R. (2009). Activities of daily living in behavioral variant frontotemporal dementia: differences in caregiver and performance-based assessments. *Alzheimer Disease & Associated Disorders, 23*(1), 70-76.
- Mioshi, E., Kipps, C. M., Dawson, K., Mitchell, J., Graham, A., & Hodges, J. R. (2007). Activities of daily living in frontotemporal dementia and Alzheimer disease. *Neurology, 68*(24), 2077-2084.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research* (Vol. 1). Walter de Gruyter.
- Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In *Handbook of modern item response theory* (pp. 351-367). Springer New York.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417-430.
- Mokken, R.J., Lewis, C., Sijtsma, K.(1986). Rejoinder to “The Mokken Scale: A Critical Discussion”. *Applied Psychological Measurement, 10*(3), 279-285.
- Mokken, R.J.: Nonparametric models for dichotomous responses. In *Handbook of modern item response theory*, Springer New York 1997; pp 351-367.

REFERENCES

- Morales, L.S., Flowers, C., Gutierrez, P., Kleinman, M., Teresi, J.A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care*, 44 (11 Suppl 3), S143.
- Morris, J. C. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics*, 9(S1), 173-176.
- Morris, J. C. (1999). Clinical presentation and course of Alzheimer disease. Alzheimer disease. Philadelphia: Lippincott Williams & Wilkins, 11-24.
- Morris, J. N., Fries, B. E., & Morris, S. A. (1999). Scaling ADLs within the MDS. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 54(11), M546-M553.
- Mortimer, J. A., Ebbitt, B., Jun, S. P., & Finch, M. D. (1992). Predictors of cognitive and functional progression in patients with probable Alzheimer's disease. *Neurology*, 42(9), 1689-1689.
- Mungas, D., & Reed, B. R. (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine*, 19(11-12), 1631-1644.
- Mungas, D., Reed, B.R., Kramer, J.H. (2003). Psychometrically matched measures of global cognition, memory and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, 17(3), 380-392.
- Nader, I. W., Tran, U. S., Baranyai, P., & Voracek, M. (2012). Investigating dimensionality of Eskin's Attitudes Toward Suicide Scale with Mokken scaling and confirmatory factor analysis. *Archives of Suicide Research*, 16(3), 226-237.
- Nagi, S.Z. (1964) A study in the evaluation of disability and rehabilitation potential: concepts, methods, and procedures. *American Journal of Public Health*, 54, 1568–1579.

REFERENCES

- Nasreddine, Z.S., Phillips, N.A., Bedirian, V., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of American Geriatric Society*, 53, 695–699.
- National Institute for Health and Clinical Excellence/Social Care Institute for Excellence. *Dementia: supporting people with dementia and their carers in health and social care. NICE Clinical Guidance 42*. London: NICE, 2006.
- Neary, D., & Snowden, J. (1996). Fronto-temporal dementia: nosology, neuropsychology, and neuropathology. *Brain and Cognition*, 31(2), 176-187.
- Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S. A., ... & Benson, D. F. (1998). Frontotemporal lobar degeneration A consensus on clinical diagnostic criteria. *Neurology*, 51(6), 1546-1554.
- Nelson, H. E., & Willison, J. R. (1991). The revised national adult reading test—test manual. Windsor: NFER-Nelson.
- Nelson, H.E. (1982). *National Adult Reading Test (NART): Test Manual*. Windsor: NFER-Nelson.
- Nelson, H.E., & McKenna, P.A.T (1975). The use of reading ability in the assessment of dementia. *British Journal of Social and Clinical Psychology*, 14,259-267
- Nelson, H.E., & O’Connell, A. (1978). Dementia: the estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex*, 14(2), 234-244.
- Ng, T. P., Niti, M., Chiam, P. C., & Kua, E. H. (2006). Physical and cognitive domains of the instrumental activities of daily living: validation in a multiethnic population of Asian older adults. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(7), 726-735.
- Niemoller, K., & van Schuur, W. (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. *Data Analysis and the Social Sciences*, 120-170.

REFERENCES

- Njegovan, V., Man-Son-Hing, M., Mitchell, S. L., & Molnar, F. J. (2001). The hierarchy of functional loss associated with cognitive decline in older persons. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *56*(10), M638-M643.
- Noerholm, V., Groenvold, M., Watt, T., Bjorner, J. B., Rasmussen, N. A., & Bech, P. (2004). Quality of life in the Danish general population—normative data and validity of WHOQOL-BREF using Rasch and item response theory models. *Quality of Life Research*, *13*(2), 531-540.
- O'Carroll, R. (1992). Predicting premorbid intellectual ability in dementia. *The Clinical Neuropsychologist*, *6*(1), 113-115.
- O'Carroll, R. E. (1987). The inter-rater reliability of the National Adult Reading Test (NART): A pilot study. *British Journal of Clinical Psychology*, *26*, 229-230.
- Office of Population Censuses and Surveys, Classification of Occupations, HMSO, London (1980)
- Patterson, M. B., Mack, J. L., Neundorfer, M. M., Martin, R. J., Smyth, K. A., & Whitehouse, P. J. (1992). Assessment of functional ability in Alzheimer disease: a review and a preliminary report on the Cleveland Scale for Activities of Daily Living. *Alzheimer Disease & Associated Disorders*, *6*(3), 145-163.
- Pearl, G. S. (1997). Diagnosis of Alzheimer's disease in a community hospital-based brain bank program. *Southern Medical Journal*, *90*(7), 720-722.
- Pérès, K., Helmer, C., Amieva, H., Orgogozo, J. M., Rouch, I., Dartigues, J. F., & Barberger-Gateau, P. (2008). Natural History of Decline in Instrumental Activities of Daily Living Performance over the 10 Years Preceding the Clinical Diagnosis of Dementia: A Prospective Population-Based Study. *Journal of the American Geriatrics Society*, *56*(1), 37-44.

REFERENCES

- Petersen, R., Stevens, J., Ganguli, M. (2001). Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, *56*, 1133-1142.
- Petersen, R.C. (2004). Mild underlying cognitive dysfunction as a diagnostic entity. *Journal of Internal Medicine*, *256*, 183-194.
- Pigliautile, M., Ricci, M., Mioshi, E., et al. (2011). Validation study of the Italian Addenbrooke's Cognitive Examination Revised in a young-old and old-old population. *Dementia and Geriatric Cognitive Disorders*, *32*, 301-307.
- Piguet, O., Hornberger, M., Mioshi, E., & Hodges, J. R. (2011). Behavioural-variant frontotemporal dementia: diagnosis, clinical staging, and management. *The Lancet Neurology*, *10*(2), 162-172.
- Potkin, S.G. (2002). The ABC of Alzheimer's disease: ADL and improving day-to-day functioning of patients. *International Psychogeriatrics*, *14* (Suppl 1), 7-26.
- Prieto, G., Contador, I., Tapias-Merino, E., Mitchell, A.J., Bermejo-Pareja, F. (2012). The Mini-Mental-37 test for dementia screening in the Spanish population: an analysis using the Rasch Model, *The Clinical Neuropsychologist*, *26*(6), 1003-1018.
- Prieto, G., Delgado, A.R., Perea, M.V., Ladera, V. (2011). Differential functioning of minimal test items according to disease. *Neurologia*, *26*(8), 474-480.
- Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & Dementia*, *9*(1), 63-75.
- Quaranta, D., Marra, C., Rossi, C., Gainotti, G., & Masullo, C. (2012). Different Apathy Profile in Behavioral Variant of Frontotemporal Dementia and Alzheimer's Disease: A Preliminary Investigation. *Behavioral Assessment*, *29*, 34.

REFERENCES

R Development Core Team: R: A language and environment for statistical computing.

Vienna, Austria: R Foundation for Statistical Computing; 2011.

Radakovic, R., Harley, C., Abrahams, S., & Starr, J. M. (2014). A systematic review of the validity and reliability of apathy scales in neurodegenerative conditions. *International Psychogeriatrics*, 1-21.

Ramsay, J.O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data.

Ranginwala, N. A., Hynan, L. S., Weiner, M. F., & White III, C. L. (2008). Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years. *The American Journal of Geriatric Psychiatry*, 16(5), 384-388.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19, 49-57.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago; 1960

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., ... & Miller, B. L. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9), 2456-2477.

Rascovsky, K., Salmon, D. P., Lipton, A. M., Leverenz, J. B., DeCarli, C., Jagust, W. J., ... & Galasko, D. (2005). Rate of progression differs in frontotemporal dementia and Alzheimer disease. *Neurology*, 65(3), 397-403.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. *Assessing quality of life in clinical trials: methods of practice*, 2, 55-73.

Reid, M.C., Lachs, M.S., Feinstein, A.R. (1995). Use of methodological standards in diagnostic test research. *Journal of the American Medical Association*, 274, 645-651

REFERENCES

- Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228-238.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103.
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48.
- Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101.
- Reise, S., Haviland, M.G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228-238.
- Reise, S.P., Widaman, K.F., Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.
- Ringdal, K., Ringdal, G. I., Kaasa, S., Bjordal, K., Wisløff, F., Sundstrøm, S., & Hjermstad, M. J. (1999). Assessing the consistency of psychometric properties of the HRQoL scales within the EORTC QLQ-C30 across populations by means of the Mokken Scaling Model. *Quality of Life Research, 8*(1-2), 25-43.
- Rohrer, D., Salmon, D. P., Wixted, J. T., & Paulsen, J. S. (1999). The disparate effects of Alzheimer's disease and Huntington's disease on semantic memory. *Neuropsychology, 13*, 381-388.
- Román, G. C., & Royall, D. R. (1999). Executive control function: a rational basis for the diagnosis of vascular dementia. *Alzheimer Disease & Associated Disorders, 13*, S69-S80.

REFERENCES

- Roorda, L. D., Houwink, A., Smits, W., Molenaar, I. W., & Geurts, A. C. (2011). Measuring upper limb capacity in poststroke patients: development, fit of the monotone homogeneity model, unidimensionality, fit of the double monotonicity model, differential item functioning, internal consistency, and feasibility of the stroke upper limb capacity scale, SULCS. *Archives of Physical Medicine and Rehabilitation*, *92*(2), 214-227.
- Roorda, L. D., Scholtes, V. A., van der Lee, J. H., Becher, J., & Dallmeijer, A. J. (2010). Measuring mobility limitations in children with cerebral palsy: development, scalability, unidimensionality, and internal consistency of the mobility questionnaire, MobQues47. *Archives of Physical Medicine and Rehabilitation*, *91*(8), 1194-1209.
- Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *The American journal of psychiatry*, *141*(11), 1356-1364.
- Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement?. *Educational and psychological measurement*, *59*(2), 248-269.
- Sabourin, S., Valois, P., & Lussier, Y. (2005). Development and validation of a brief version of the dyadic adjustment scale with a nonparametric item analysis model. *Psychological Assessment*, *17*(1), 15.
- Sachs, J., Law, Y. K., & Chan, C. (2003). A nonparametric item analysis of a selected item subset of the Learning Process Questionnaire. *British Journal of Educational Psychology*, *73*(3), 395-423.
- Sackett, D.L. (1992). A primer on the precision and accuracy of the clinical examination. *Journal of the American Medical Association*, *267*, 2638-2644.

REFERENCES

- Salmon, D.P., Riekkinen, P.J., Katzman, R., Zhang, M.Y., Jin, H., Yu, E. (1989). Cross-cultural studies of dementia. A comparison of Mini-Mental State Examination in Finland and China. *Archives of Neurology*, 46, 769-72.
- Samejima F. The graded response model. In: van der Linden WJ, Hambleton R, editors. Handbook of modern item response theory. New York, NY: Springer; 1996. pp. 85–100.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4), 100.
- Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*, 37(1), 68.
- Sampson, E.L., Blanchard, M.R., Jones, L., Tookman, A., King M. (2009). Dementia in the acute hospital: prospective cohort study of prevalence and mortality. *British Journal of Psychiatry*, 195, 61-6.
- Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, 10(4), 345.
- Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, 6(3), 255.
- Santor, D.A., Ascher-Svanum, H., Lindenmayer, J.P., Obenchain, R.L. (2007). Item response analysis of the positive and negative syndrome scale. *BMC Psychiatry*, 7, 66-76.
- Scottish Council for Research in Education. (1933). *The intelligence of Scottish children: A national survey of an age-group*. London, England: University of London Press.

REFERENCES

Scottish Council for Research in Education. (1949). *The Trend of Scottish Intelligence*.

London: University of London Press

Shankle, W. R., Romney, A. K., Hara, J., Fortier, D., Dick, M. B., Chen, J. M., ... & Sun, X. (2005). Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(13), 4919-4924.

Sharpe, K., & O'Carroll, R. (1991). Estimating premorbid intellectual level in dementia using the National Adult Reading Test: A Canadian study. *British Journal of Clinical Psychology*, *30*, 381-384.

Sheehan, T.J., DeChello, L.M., Garcia, R., Fifield, J., Rothfield, N., Reisine, S. (2002). Measuring disability: application of the Rasch model to activities of daily living (ADL/IADL). *Journal of Outcome Measurement*, *5*, 839-863.

Sijtsma, K., & Meijer, R.R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191-207.

Sijtsma, K., & Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*(2), 183-200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*(1), 79-105.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*(1), 79-97.

Sijtsma, K., & Molenaar, I. W. (Eds.). (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage.

REFERENCES

- Sijtsma, K., Debets, P., & Molenaar, I.W. (1990) Mokken scaling analysis for polytomous items: theory, a computer program and an empirical application. *Quality & Quantity* 24(2), 173-188.
- Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the world health organization quality-of-life scale (WHOQOL-Bref). *Quality of Life Research*, 17(2), 275-290.
- Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, 50(1), 31-37.
- Sikkes, S. A. M., De Lange-de Klerk, E. S. M., Pijnenburg, Y. A. L., & Scheltens, P. (2009). A systematic review of Instrumental Activities of Daily Living scales in dementia: room for improvement. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(1), 7-12.
- Smits, I. A., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken Scale Analysis as a Dimensionality Assessment Tool Why Scalability Does Not Imply Unidimensionality. *Applied Psychological Measurement*, 36(6), 516-539.
- Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: A form of circumscribed cerebral atrophy. *Behavioural Neurology*, 2(3), 167-182.
- Spector, W. D. (1997). Measuring functioning in daily activities for persons with dementia. *Alzheimer Disease and Associated Disorders*, 11, 81-90.

REFERENCES

- Spector, W. D., & Fleishman, J. A. (1998). Combining activities of daily living with instrumental activities of daily living to measure functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 53(1), S46-S57.
- Spector, W.D, Katz, S., Murphy J.B., Fulton, J.P. (1978). The hierarchical relationship between activities of daily living and instrumental activities of daily living. *Journal of Chronic Diseases*, 40, 481-489
- Spector, W.D., Fleishman, J.A. (1998). Combining activities of daily living with instrumental activities of daily living to measure functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 53(Suppl 1), 46-57.
- Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., ... & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 280-292.
- Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology*, 12(1), 74.
- Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A., & Clausen, J.A. (1950) *Measurement and Prediction*, Vol 4. Princeton University Press, Princeton, NJ
- Straat, J. H. (2012). *Using scalability coefficients and conditional association to assess monotone homogeneity* (Doctoral dissertation, Tilburg University).
- Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum Sample Size Requirements for Mokken Scale Analysis. *Educational and Psychological Measurement*, 74(5), 809-822.

REFERENCES

- Straat, J.H. Using Scalability Coefficients and Conditional Association to Assess Monotone Homogeneity. Ridderkerk: Ridderprint BV; 2012.
- Suurmeijer, T. P., Doeglas, D. M., Moum, T., Briançon, S., Krol, B., Sanderman, R., ... & Van den Heuvel, W. J. (1994). The Groningen Activity Restriction Scale for measuring disability: its utility in international comparisons. *American Journal of Public Health, 84*(8), 1270-1273.
- Tang-Wai, D. F., Josephs, K. A., Boeve, B. F., Dickson, D. W., Parisi, J. E., & Petersen, R. C. (2003). Pathologically confirmed corticobasal degeneration presenting with visuospatial dysfunction. *Neurology, 61*(8), 1134-1135.
- Tellegen, P., & Laros, J. (1993). The construction and validation of a nonverbal test of intelligence: the revision of the Snijders-Oomen tests. *European Journal of Psychological Assessment, 9*(2), 147-157.
- Tendeiro, J. N., & Tendeiro, M. J. N. (2014). Package 'PerFit'.
- Teresi, J.A., Golden, R.R., Cross, P., Gurland, B., Kleinman, M., Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology, 48*(4), 473-483.
- Teresi, J.A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., Holmes, D. (2000). Applications of item response theory to the examination of the psychometric properties and differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and White non-Latino elderly. *Research on Aging, 22*(6), 738-773.
- Thies, W., & Bleiler, L. (2013). 2013 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association, 9*(2), 208-245.

REFERENCES

- Thissen, D. (2003). Estimation in multilog. In M. du Toit (Ed.), *IRT from SSI: Bilog-MG, multilog, parscale, testfact*. Lincolnwood, IL: Scientific Software International.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, 67-113.
- Townsend, P. (1962). *The last refuge: a survey of residential institutions and homes for the aged in England and Wales*. London, Routledge.
- Tschanz, J. T., Corcoran, C. D., Schwartz, S., Treiber, K., Green, R. C., Norton, M. C., ... & Lyketsos, C. G. (2011). Progression of cognitive, functional, and neuropsychiatric symptom domains in a population cohort with Alzheimer dementia: the Cache County Dementia Progression study. *The American Journal of Geriatric Psychiatry, 19*(6), 532-542.
- Valcour, V. G., Masaki, K. H., & Blanchette, P. L. (2002). The phrase: "no ifs, ands, or buts" and cognitive testing. Lessons from an Asian-American community. *Hawaii Medical Journal, 61*(4), 72-74.
- Valdmanis, P. N., Dupre, N., Bouchard, J. P., Camu, W., Salachas, F., Meininger, V., ... & Rouleau, G. A. (2007). Three families with amyotrophic lateral sclerosis and frontotemporal dementia with evidence of linkage to chromosome 9p. *Archives of Neurology, 64*(2), 240-245.
- van de Vijver, F., Hambleton, R.K.(1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-88.
- van der Ark LA: Mokken (Version 2.5.1): An R package for Mokken scale analysis [computer software]. 2011.
- van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

REFERENCES

- van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5), 1-27.
- van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, 73(2), 183-208.
- van der Lee, J.H., Roorda, L.D., Beckerman, H., Lankhorst, G.J., & Bouter, L.M. (2002). Improving the Action Research Arm test: a unidimensional hierarchical scale. *Clinical Rehabilitation*, 16(6), 646-653.
- Van Schuur, W.H. (2003). Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis*, 11(2), 139-163.
- Vetter, P. H., Krauss, S., Steiner, O., Kropp, P., Möller, W. D., Moises, H. W., & Köller, O. (1999). Vascular dementia versus dementia of Alzheimer's type: do they have differential effects on caregivers' burden?. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 54(2), S93-S98.
- Vittengl, J.R., White, C.N., McGovern, R.J., & Morton, B.J. (2006). Comparative validity of seven scoring systems for the instrumental activities of daily living scale in rural elders. *Aging and Mental Health*, 10, 40-47.
- Waehrens, E. E., Bliddal, H., Dannekiold-Samsøe, B., Lund, H., & Fisher, A. G. (2012). Differences between questionnaire-and interview-based measures of activities of daily living (ADL) ability and their association with observed ADL ability in women with rheumatoid arthritis, knee osteoarthritis, and fibromyalgia. *Scandinavian Journal of Rheumatology*, 41(2), 95-102.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185-201.

REFERENCES

- Waite, L. M., Broe, G. A., Grayson, D. A., & Creasey, H. (2000). Motor function and disability in the dementias. *International Journal of Geriatric Psychiatry, 15*(10), 897-903.
- Wang, B. W., Lu, E., Mackenzie, I. A., Assaly, M., Jacova, C., Lee, P. E., & ... Hsiung, G. R. (2012). Multiple pathologies are common in Alzheimer patients in clinical trials. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques, 39*(5), 592-599.
- Ware Jr., J.E., Kosinski, M., Bjorner, J.B., Bayliss, M.S., Batenhorst, A., Carl, G.H.,...Dowson, A. (2003). Applications of computerised adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935-952.
- Ware, J. E., Kosinski, M., Dewey, J. E., & Gandek, B. (2000). SF-36 health survey: manual and interpretation guide. Quality Metric Inc.
- Watson, R. (1996). The Mokken scaling procedure (MSP) applied to the measurement of feeding difficulty in elderly people with dementia. *International Journal of Nursing Studies, 33*(4), 385-393.
- Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological medicine, 38*(04), 575-579.
- Watson, R., Deary, I., & Austin E. (2007). Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI. *Personality and Individual Differences, 43*, 1460-1469.
- Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: how Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing, 21*(19pt20), 2736-2746.

REFERENCES

- Watson, R., Wang, W., & Thompson, D. R. (2014). Violations of local stochastic independence exaggerate scalability in Mokken scaling analysis of the Chinese Mandarin SF-36. *Health and Quality of Life Outcomes, 12*(149), 1-10.
- Watson, R., Wang, W., Ski, C. F., & Thompson, D. R. (2012). The Chinese version of the Myocardial Infarction Dimensional Assessment Scale (MIDAS): Mokken scaling. *Health and Quality of Life Outcomes, 10*(2), 1-4.
- Watson, R., Wang, W., Thompson, D. R., & Meijer, R. R. (2014). Investigating invariant item ordering in the Mental Health Inventory: An illustration of the use of different methods. *Personality and Individual Differences, 66*, 74-78.
- Wattmo, C., Wallin, Å. K., & Minthon, L. (2012). Functional response to cholinesterase inhibitor therapy in a naturalistic Alzheimer's disease cohort. *BMC Neurology, 12*(1), 134.
- Wattmo, C., Wallin, Å. K., Londos, E., & Minthon, L. (2011). Long-term outcome and prediction models of activities of daily living in Alzheimer disease with cholinesterase inhibitor treatment. *Alzheimer Disease & Associated Disorders, 25*(1), 63-72.
- Wechsler D. Manual for the Wechsler Adult Intelligence Scale (WAIS) The Psychological Corporation, New York (1955)
- Wechsler, D. (1997). Wechsler Adult Intelligence Scale—Third Edition (WAIS).
- Wechsler, D. (2001). Wechsler Test of Adult Reading: WTAR. Psychological Corporation.
- Weiss, C., Fried, L., Brandeen-Roche, K. (2007). Exploring the hierarchy of mobility performance in high-functioning older women. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 62*, 167–173.
- Welsh, K., Butters, N., Hughes, J. and Mohs, R.C. (1992) Detection and staging of dementia in Alzheimer's disease: Use of the neuropsychological measures developed for the

REFERENCES

- Consortium to Establish a Registry for Alzheimer's disease. *Archives of Neurology*, 49, 448-452.
- Welsh, K., Butters, N., Hughes, J., Mohs, R., and Heyman, A. (1991). Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Archives of Neurology*. 48, 278-281.
- Wicklund, A. H., Johnson, N., Rademaker, A., Weitner, B. B., & Weintraub, S. (2006). Word list versus story memory in Alzheimer disease and frontotemporal dementia. *Alzheimer Disease & Associated Disorders*, 20(2), 86-92.
- Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., Koepke, K. M., ... & ACTIVE Study Group. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *Jama*, 296(23), 2805-2814.
- Wilson, R. S., Rosenbaum, G., Brown, G., Rourke, D., Whitman, D. & Grisell, J. (1978). An index of premorbid intelligence. *Journal of Consulting and Clinical Psychology*, 46, 1554-1555.
- Wouters, H., van Gool, W. A., Schmand, B., Zwinderman, A. H., & Lindeboom, R. (2010). Three sides of the same coin: measuring global cognitive impairment with the MMSE, ADAS-cog and CAMCOG. *International Journal of Geriatric Psychiatry*, 25(8), 770-779.
- Wouters, H., van Gool, W.A., Schmand, B., Lindeboom, R. (2008). Revising the ADAS-cog for a more accurate assessment of cognitive impairment. *Alzheimer Disease & Associated Disorders*, 22(3), 236-244.
- Wouters, H., Zwinderman, A.H., van Gool, W.A., Schmand, B., Lindeboom, R. (2009). Adaptive cognitive testing in dementia. *International Journal of Methods in Psychiatric Research*, 18(2), 119-127.

REFERENCES

- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.
- Yew, B., Alladi, S., Shailaja, M., Hodges, J. R., & Hornberger, M. (2013). Lost and forgotten? Orientation versus memory in Alzheimer's disease and frontotemporal dementia. *Journal of Alzheimer's Disease, 33*(2), 473-481.
- Zheng, L., Mack, W.J., Chui, H.C., Heflin, L., Mungas, D., Reed, B., DeCarli, C., Weiner, M.W., Kramer, J.H. (2012). Coronary Artery Disease Is Associated with Cognitive Decline Independent of Changes on Magnetic Resonance Imaging in Cognitively Normal Elderly Adults. *Journal of the American Geriatric Society, 60*, 499–504.

Appendices


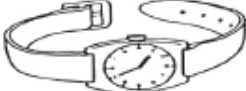










Appendix A: Addenbrooke's Cognitive Examination scales

ADDENBROOKE'S COGNITIVE EXAMINATION - ACE-R <i>Final Revised Version A (2005)</i>						
Name : Date of birth : Hospital no. :	Date of testing: / / Tester's name: Age at leaving full-time education: Occupation: Handedness:					
<i>Addressograph</i>						
ORIENTATION						
➤ Ask: What is the	Day	Date	Month	Year	Season	[Score 0-5] <input type="text"/> <input type="text"/>
➤ Ask: Which	Building	Floor	Town	County	Country	[Score 0-5] <input type="text"/> <input type="text"/>
REGISTRATION						
➤ Tell: 'I'm going to give you three words and I'd like you to repeat after me: lemon, key and ball'. After subject repeats, say 'Try to remember them because I'm going to ask you later'. Score only the first trial (repeat 3 times if necessary). Register number of trials						[Score 0-3] <input type="text"/> <input type="text"/>
ATTENTION & CONCENTRATION						
➤ Ask the subject: 'could you take 7 away from a 100? After the subject responds, ask him or her to take away another 7 to a total of 5 subtractions. If subject make a mistake, carry on and check the subsequent answer (i.e. 93, 84, 77, 70, 63 -score 4) Stop after five subtractions (93, 86, 79, 72, 65). ➤ Ask: 'could you please spell WORLD for me? Then ask him/her to spell it backwards:						[Score 0-5] <input type="text"/> <input type="text"/> (for the best performed task)
MEMORY - Recall						
➤ Ask: 'Which 3 words did I ask you to repeat and remember?'						[Score 0-3] <input type="text"/> <input type="text"/>
MEMORY - Anterograde Memory						
➤ Tell: 'I'm going to give you a name and address and I'd like you to repeat after me. We'll be doing that 3 times, so you have a chance to learn it. I'll be asking you later' Score only the third trial						[Score 0-7] <input type="text"/>
	1 st Trial	2 nd Trial	3 rd Trial			
Harry Barnes			
73 Orchard Close			
Kingsbridge			
Devon			
MEMORY - Retrograde Memory						
➤ Name of current Prime Minister ➤ Name of the woman who was Prime Minister ➤ Name of the USA president ➤ Name of the USA president who was assassinated in the 1960's						[Score 0 -4] <input type="text"/>

copyright 2000, John R. Hodges


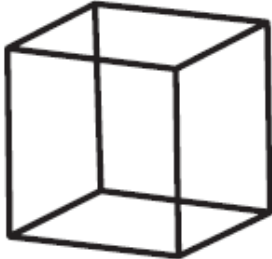
VERBAL FLUENCY - Letter 'P' and animals																				
<p>➤ Letters</p> <p>Say: 'I'm going to give you a letter of the alphabet and I'd like you to generate as many words as you can beginning with that letter, but not names of people or places. Are you ready? You've got a minute and the letter is P'</p>	<p>[Score 0 - 7]</p> <input style="width: 30px; height: 20px;" type="text"/>	Y																		
	<table style="font-size: small; border-collapse: collapse;"> <tr><td>>17</td><td>7</td></tr> <tr><td>14-17</td><td>6</td></tr> <tr><td>11-13</td><td>5</td></tr> <tr><td>8-10</td><td>4</td></tr> <tr><td>6-7</td><td>3</td></tr> <tr><td>4-5</td><td>2</td></tr> <tr><td>2-3</td><td>1</td></tr> <tr><td><2</td><td>0</td></tr> <tr><td>total</td><td>correct</td></tr> </table>	>17	7	14-17	6	11-13	5	8-10	4	6-7	3	4-5	2	2-3	1	<2	0	total	correct	C
>17	7																			
14-17	6																			
11-13	5																			
8-10	4																			
6-7	3																			
4-5	2																			
2-3	1																			
<2	0																			
total	correct																			
<p>➤ Animals</p> <p>Say: 'Now can you name as many animals as possible, beginning with any letter?'</p>	<p>[Score 0 - 7]</p> <input style="width: 30px; height: 20px;" type="text"/>	D																		
	<table style="font-size: small; border-collapse: collapse;"> <tr><td>>21</td><td>7</td></tr> <tr><td>17-21</td><td>6</td></tr> <tr><td>14-16</td><td>5</td></tr> <tr><td>11-13</td><td>4</td></tr> <tr><td>9-10</td><td>3</td></tr> <tr><td>7-8</td><td>2</td></tr> <tr><td>5-6</td><td>1</td></tr> <tr><td><5</td><td>0</td></tr> <tr><td>total</td><td>correct</td></tr> </table>	>21	7	17-21	6	14-16	5	11-13	4	9-10	3	7-8	2	5-6	1	<5	0	total	correct	L
>21	7																			
17-21	6																			
14-16	5																			
11-13	4																			
9-10	3																			
7-8	2																			
5-6	1																			
<5	0																			
total	correct																			
LANGUAGE - Comprehension																				
<p>➤ Show written instruction:</p>	<p>[Score 0-1]</p> <input style="width: 30px; height: 20px;" type="text"/> <input style="width: 20px; height: 20px; background-color: #cccccc;" type="checkbox"/>	E																		
<h1 style="margin: 0;">Close your eyes</h1>		G																		
<p>➤ 3 stage command: 'Take the paper in your right hand. Fold the paper in half. Put the paper on the floor'</p>	<p>[Score 0-3]</p> <input style="width: 30px; height: 20px;" type="text"/> <input style="width: 20px; height: 20px; background-color: #cccccc;" type="checkbox"/>	D																		
LANGUAGE - Writing																				
<p>➤ Ask the subject to make up a sentence and write it in the space below: Score 1 if sentence contains a subject and a verb (see guide for examples)</p>	<p>[Score 0-1]</p> <input style="width: 30px; height: 20px;" type="text"/> <input style="width: 20px; height: 20px; background-color: #cccccc;" type="checkbox"/>	N																		
		A																		
		L																		

APPENDIX A

LANGUAGE - Repetition	
<p>➤ Ask the subject to repeat: 'hippopotamus'; 'eccentricity'; 'unintelligible'; 'statistician' Score 2 if all correct; 1 if 3 correct; 0 if 2 or less.</p>	<p>[Score 0-2] <input type="text"/></p>
<p>➤ Ask the subject to repeat: 'Above, beyond and below'</p>	<p>[Score 0-1] <input type="text"/></p>
<p>➤ Ask the subject to repeat: 'No ifs, ands or buts'</p>	<p>[Score 0-1] <input type="text"/> <input type="checkbox"/></p>
LANGUAGE - Naming	
<p>➤ Ask the subject to name the following pictures:</p>	<p>[Score 0-2] pencil + watch <input type="text"/> <input type="checkbox"/></p>
<p>_____ <input type="text"/></p> 	<p>[Score 0-10] <input type="text"/></p>
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
<p>_____ <input type="text"/></p> 	
LANGUAGE - Comprehension	
<p>➤ Using the pictures above, ask the subject to:</p> <ul style="list-style-type: none"> • Point to the one which is associated with the monarchy _____ • Point to the one which is a marsupial _____ • Point to the one which is found in the Antarctic _____ • Point to the one which has a nautical connection _____ 	<p>[Score 0-4] <input type="text"/></p>

E
G
A
U
G
N
A
L

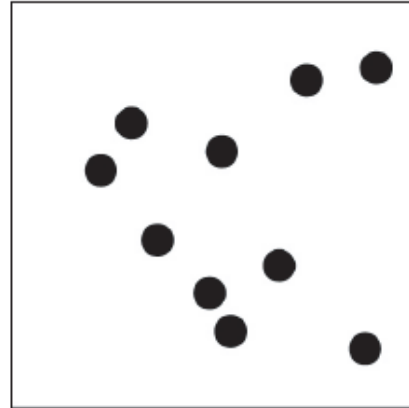
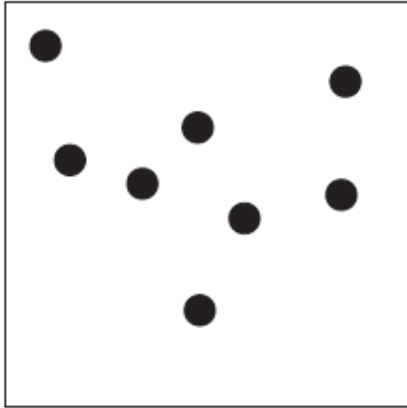
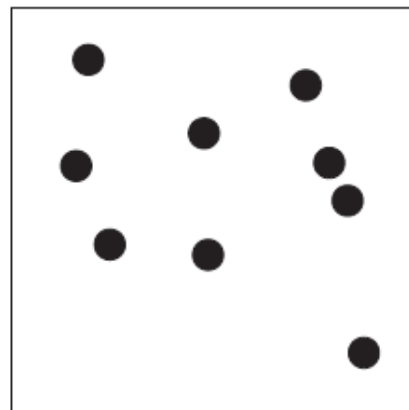
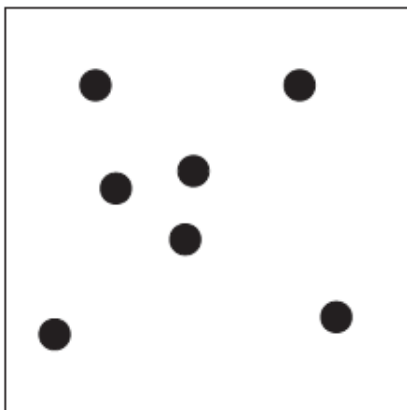
APPENDIX A

L A N G U A G E - Reading			
<p>➤ Ask the subject to read the following words: [Score 1 only if all correct]</p> <p style="text-align: center;"> sew pint soot dough height </p>	<p>[Score 0-1]</p> <input type="text"/>		L A N G U A G E
V I S U O S P A T I A L A B I L I T I E S			
<p>➤ Overlapping pentagons: Ask the subject to copy this diagram:</p>	<p>[Score 0-1]</p> <input type="text"/> <input type="text"/>		L A T I T I E S
			
<p>➤ Wire cube : Ask the subject to copy this drawing (for scoring, see instructions guide)</p>	<p>[Score 0-2]</p> <input type="text"/>		P S O U S I T Y
			
<p>➤ Clock: Ask the subject to draw a clock face with numbers and the hands at ten past five. (for scoring see instruction guide: circle = 1, numbers = 2, hands = 2 if all correct)</p>	<p>[Score 0-5]</p> <input type="text"/>		V

PERCEPTUAL ABILITIES





➤ Ask the subject to count the dots without pointing them

[Score 0-4]

L
A
I
T
A
P
S
O
U
S
I
V

APPENDIX A

ADDENBROOKE'S COGNITIVE EXAMINATION - ACE-R				Final Revised Version A (2005)	
PERCEPTUAL ABILITIES					
➤ Ask the subject to identify the letters				[Score 0-4] <input style="width: 40px;" type="text"/>	
<input style="width: 40px;" type="text"/>	<input style="width: 40px;" type="text"/>			V I S U O S P A T I A L	
<input style="width: 40px;" type="text"/>	<input style="width: 40px;" type="text"/>				
RECALL					
➤ Ask "Now tell me what you remember of that name and address we were repeating at the beginning"					[Score 0-7] <input style="width: 40px;" type="text"/>
Harry Bames 73 Orchard Close Kingsbridge Devon			Y O R O M E M O R Y	
RECOGNITION					
➤ This test should be done if subject failed to recall one or more items. If all items were recalled, skip the test and score 5. If only part is recalled start by ticking items recalled in the shadowed column on the right hand side. Then test not recalled items by telling "ok, I'll give you some hints: was the name X, Y or Z?" and so on. Each recognised item scores one point which is added to the point gained by recalling.					[Score 0-5] <input style="width: 40px;" type="text"/>
Jerry Bames 37	Harry Bames 73	Harry Bradford 76	recalled recalled	M E M O R Y	
Orchard Place	Oak Close	Orchard Close	recalled		
Oakhampton	Kingsbridge	Dartington	recalled		
Devon	Dorset	Somerset	recalled		
General Scores					
			MMSE	/30	
			ACE-R	/100	
Subscores					
			Attention and Orientation	/18	
			Memory	/26	
			Fluency	/14	
			Language	/26	
			Visuospatial	/16	

Normative values based on 63 controls aged 52-75 and 142 dementia patients aged 46-86

Cut-off <88 gives 94% sensitivity and 89% specificity for dementia
 Cut-off <82 gives 84% sensitivity and 100% specificity for dementia

copyright 2000, John R. Hodges













ADDENBROOKE'S COGNITIVE EXAMINATION – ACE-III English Version A (2012)																								
Name: _____ Date of Birth: _____ Hospital No. or Address: _____			Date of testing: ___/___/___ Tester's name: _____ Age at leaving full-time education: _____ Occupation: _____ Handedness: _____																					
ATTENTION																								
➤ Ask: What is the	Day _____	Date _____	Month _____	Year _____	Season _____	Attention [Score 0-5] <input style="width: 30px; height: 20px;" type="text"/>																		
➤ Ask: Which	No./Floor _____	Street/Hospital _____	Town _____	County _____	Country _____	Attention [Score 0-5] <input style="width: 30px; height: 20px;" type="text"/>																		
ATTENTION																								
➤ Tell: "I'm going to give you three words and I'd like you to repeat them after me: lemon, key and ball." After subject repeats, say "Try to remember them because I'm going to ask you later." ➤ Score only the first trial (repeat 3 times if necessary). ➤ Register number of trials: _____						Attention [Score 0-3] <input style="width: 30px; height: 20px;" type="text"/>																		
ATTENTION																								
➤ Ask the subject: "Could you take 7 away from 100? I'd like you to keep taking 7 away from each new number until I tell you to stop." ➤ If subject makes a mistake, do not stop them. Let the subject carry on and check subsequent answers (e.g., 93, 84, 77, 70, 63 – score 4). ➤ Stop after five subtractions (93, 86, 79, 72, 65): _____						Attention [Score 0-5] <input style="width: 30px; height: 20px;" type="text"/>																		
MEMORY																								
➤ Ask: "Which 3 words did I ask you to repeat and remember?" _____						Memory [Score 0-3] <input style="width: 30px; height: 20px;" type="text"/>																		
FLUENCY																								
➤ Letters Say: "I'm going to give you a letter of the alphabet and I'd like you to generate as many words as you can beginning with that letter, but not names of people or places. For example, if I give you the letter "C", you could give me words like "cat, cry, clock" and so on. But, you can't give me words like Catherine or Canada. Do you understand? Are you ready? You have one minute. The letter I want you to use is the letter "P"."						Fluency [Score 0 – 7] <input style="width: 30px; height: 20px;" type="text"/>																		
						<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td>≥ 18</td><td>7</td></tr> <tr><td>14-17</td><td>6</td></tr> <tr><td>11-13</td><td>5</td></tr> <tr><td>8-10</td><td>4</td></tr> <tr><td>6-7</td><td>3</td></tr> <tr><td>4-5</td><td>2</td></tr> <tr><td>2-3</td><td>1</td></tr> <tr><td>0-1</td><td>0</td></tr> <tr><td>total</td><td>correct</td></tr> </table>	≥ 18	7	14-17	6	11-13	5	8-10	4	6-7	3	4-5	2	2-3	1	0-1	0	total	correct
≥ 18	7																							
14-17	6																							
11-13	5																							
8-10	4																							
6-7	3																							
4-5	2																							
2-3	1																							
0-1	0																							
total	correct																							
➤ Animals Say: "Now can you name as many animals as possible. It can begin with any letter."						Fluency [Score 0 – 7] <input style="width: 30px; height: 20px;" type="text"/>																		
						<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td>≥ 22</td><td>7</td></tr> <tr><td>17-21</td><td>6</td></tr> <tr><td>14-16</td><td>5</td></tr> <tr><td>11-13</td><td>4</td></tr> <tr><td>9-10</td><td>3</td></tr> <tr><td>7-8</td><td>2</td></tr> <tr><td>5-6</td><td>1</td></tr> <tr><td><5</td><td>0</td></tr> <tr><td>total</td><td>correct</td></tr> </table>	≥ 22	7	17-21	6	14-16	5	11-13	4	9-10	3	7-8	2	5-6	1	<5	0	total	correct
≥ 22	7																							
17-21	6																							
14-16	5																							
11-13	4																							
9-10	3																							
7-8	2																							
5-6	1																							
<5	0																							
total	correct																							

APPENDIX A

MEMORY			
<p>> Tell: "I'm going to give you a name and address and I'd like you to repeat the name and address after me. So you have a chance to learn, we'll be doing that 3 times. I'll ask you the name and address later."</p> <p>Score only the third trial.</p>			<p>Memory [Score 0 – 7]</p> <input type="text"/>
	<i>1st Trial</i>	<i>2nd Trial</i>	<i>3rd Trial</i>
<p>Harry Barnes 73 Orchard Close Kingsbridge Devon</p>	<p>_____</p> <p>_____</p> <p>_____</p>	<p>_____</p> <p>_____</p> <p>_____</p>	<p>_____</p> <p>_____</p> <p>_____</p>
MEMORY			
<p>> Name of the current Prime Minister.....</p> <p>> Name of the woman who was Prime Minister</p> <p>> Name of the USA president.....</p> <p>> Name of the USA president who was assassinated in the 1960s.....</p>			<p>Memory [Score 0 – 4]</p> <input type="text"/>
LANGUAGE			
<p>> Place a pencil and a piece of paper in front of the subject. As a practice trial, ask the subject to "Pick up the pencil and then the paper." If incorrect, score 0 and do not continue further.</p> <p>> If the subject is correct on the practice trial, continue with the following three commands below.</p> <ul style="list-style-type: none"> • Ask the subject to "Place the paper on top of the pencil" • Ask the subject to "Pick up the pencil but not the paper" • Ask the subject to "Pass me the pencil after touching the paper" <p>Note: Place the pencil and paper in front of the subject before each command.</p>			<p>Language [Score 0-3]</p> <input type="text"/>
LANGUAGE			
<p>> Ask the subject to write two (or more) complete sentences about his/her last holiday/weekend/Christmas. Write in complete sentences and do not use abbreviations. Give 1 point if there are two (or more) complete sentences about the one topic; and give another 1 point if grammar and spelling are correct.</p>			<p>Language [Score 0-2]</p> <input type="text"/>
LANGUAGE			
<p>> Ask the subject to repeat: 'caterpillar'; 'eccentricity'; 'unintelligible'; 'statistician'</p> <p>Score 2 if all are correct; score 1 if 3 are correct; and score 0 if 2 or less are correct.</p>			<p>Language [Score 0-2]</p> <input type="text"/>


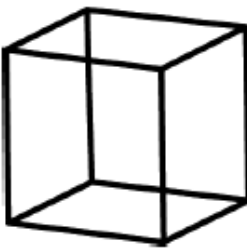
Updated 20/11/2012

APPENDIX A

LANGUAGE		
<p>➤ Ask the subject to repeat: 'All that glitters is not gold'</p>		<p>Language [Score 0-1] <input type="text"/></p>
<p>➤ Ask the subject to repeat: 'A stitch in time saves nine'</p>		<p>Language [Score 0-1] <input type="text"/></p>
LANGUAGE		
<p>➤ Ask the subject to name the following pictures:</p>		<p>Language [Score 0-12] <input type="text"/></p>
<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 
<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 
<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 
<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 	<p>_____ <input type="text"/></p> 
LANGUAGE		
<p>➤ Using the pictures above, ask the subject to:</p> <ul style="list-style-type: none"> • Point to the one which is associated with the monarchy • Point to the one which is a marsupial • Point to the one which is found in the Antarctic • Point to the one which has a nautical connection 		<p>Language [Score 0-4] <input type="text"/></p>

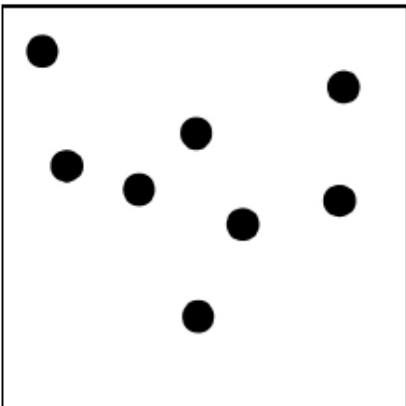
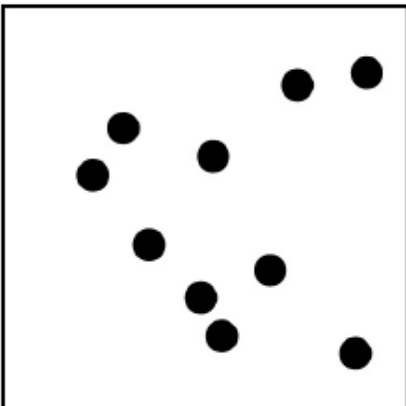
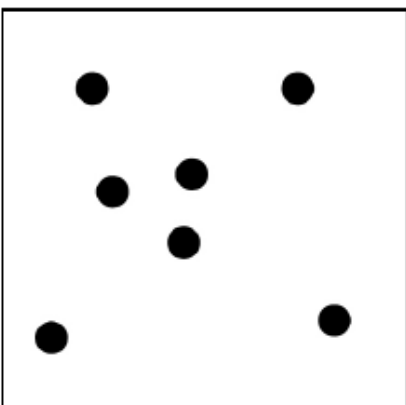
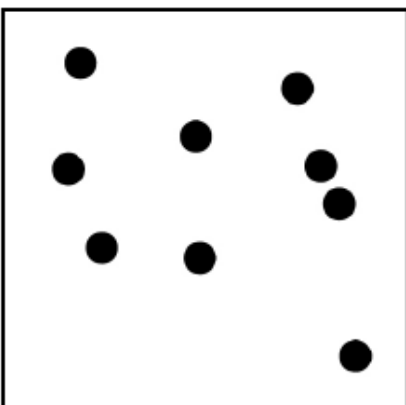
Updated 20/11/2012

APPENDIX A

LANGUAGE	
<p>> Ask the subject to read the following words: (Score 1 only if all correct)</p> <p style="text-align: center;"> sew pint soot dough height </p>	<p>Language [Score 0-1]</p> <input type="text"/>
VISUOSPATIAL ABILITIES	
<p>> Infinity Diagram: Ask the subject to copy this diagram</p>	<p>Visuospatial [Score 0-1]</p> <input type="text"/>
	
<p>> Wire cube: Ask the subject to copy this drawing (for scoring, see instructions guide).</p>	<p>Visuospatial [Score 0-2]</p> <input type="text"/>
	
<p>> Clock: Ask the subject to draw a clock face with numbers and the hands at ten past five. (For scoring see instruction guide: circle = 1, numbers = 2, hands = 2 if all correct).</p>	<p>Visuospatial [Score 0-5]</p> <input type="text"/>




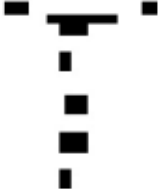
Updated 20/11/2012

APPENDIX A

VISUOSPATIAL ABILITIES	
<p>> Ask the subject to count the dots without pointing to them</p>	Visuospatial [Score 0-4] <input type="text"/>
<div style="border: 1px solid black; width: 150px; height: 150px; margin: 10px auto;"></div>	<div style="border: 1px solid black; width: 150px; height: 150px; margin: 10px auto;"></div>
<div style="border: 1px solid black; width: 150px; height: 150px; margin: 10px auto;"></div>	<div style="border: 1px solid black; width: 150px; height: 150px; margin: 10px auto;"></div>

Updated 20/11/2012

APPENDIX A

VISUOSPATIAL ABILITIES			
> Ask the subject to identify the letters			Visuospatial [Score 0-4] <input style="width: 30px; height: 15px;" type="text"/>
		<input style="width: 30px; height: 15px;" type="text"/>	<input style="width: 30px; height: 15px;" type="text"/>
		<input style="width: 30px; height: 15px;" type="text"/>	<input style="width: 30px; height: 15px;" type="text"/>
MEMORY			
> Ask "Now tell me what you remember about that name and address we were repeating at the beginning"			
Harry Barnes 73 Orchard Close Kingsbridge Devon	Memory [Score 0-7] <input style="width: 30px; height: 15px;" type="text"/>	
MEMORY			
> This test should be done if the subject failed to recall one or more items above. If all items were recalled, skip the test and score 5. If only part was recalled start by ticking items recalled in the shadowed column on the right hand side; and then test not recalled items by telling the subject "ok, I'll give you some hints: was the name X, Y or Z?" and so on. Each recognised item scores one point, which is added to the point gained by recalling.			Memory [Score 0-5] <input style="width: 30px; height: 15px;" type="text"/>
Jerry Barnes	Harry Barnes	Harry Bradford	recalled
37	73	76	recalled
Orchard Place	Oak Close	Orchard Close	recalled
Oakhampton	Kingsbridge	Dartington	recalled
Devon	Dorset	Somerset	recalled
SCORES			
TOTAL ACE-III SCORE			/100
Attention			/18
Memory			/26
Fluency			/14
Language			/26
Visuospatial			/16

Updated 20/11/2012

APPENDIX A

MINI – ADDENBROOKE'S COGNITIVE EXAMINATION UK Version A (2014)																								
Name: Date of Birth: Hospital No. or Address:			Date of testing: ___/___/___ Tester's name: _____ Age at leaving full-time education: _____ Occupation: _____ Handedness: _____																					
ATTENTION																								
> Ask: What is the	Day _____	Date _____	Month _____	Year _____	Attention [Score 0-4] <input style="width: 40px; height: 20px;" type="text"/>																			
MEMORY																								
> Tell: "I'm going to give you a name and address and I'd like you to repeat the name and address after me. So you have a chance to learn, we'll be doing that 3 times. I'll ask you the name and address later." Score only the third trial.					Memory [Score 0 – 7] <input style="width: 40px; height: 20px;" type="text"/>																			
	<i>1st Trial</i>	<i>2nd Trial</i>	<i>3rd Trial</i>																					
Harry Barnes 73 Orchard Close Kingsbridge Devon	_____ _____ _____ _____	_____ _____ _____ _____	_____ _____ _____ _____																					
FLUENCY – ANIMALS																								
> Animals Say: "Now can you name as many animals as possible. It can begin with any letter. You have one minute. Go ahead."					Fluency [Score 0 – 7] <input style="width: 40px; height: 20px;" type="text"/>																			
					<table border="1" style="font-size: small; border-collapse: collapse;"> <tr><td>≥ 22</td><td>7</td></tr> <tr><td>17-21</td><td>6</td></tr> <tr><td>14-16</td><td>5</td></tr> <tr><td>11-13</td><td>4</td></tr> <tr><td>9-10</td><td>3</td></tr> <tr><td>7-8</td><td>2</td></tr> <tr><td>5-6</td><td>1</td></tr> <tr><td><5</td><td>0</td></tr> <tr><td>total</td><td>correct</td></tr> </table>	≥ 22	7	17-21	6	14-16	5	11-13	4	9-10	3	7-8	2	5-6	1	<5	0	total	correct	
≥ 22	7																							
17-21	6																							
14-16	5																							
11-13	4																							
9-10	3																							
7-8	2																							
5-6	1																							
<5	0																							
total	correct																							

APPENDIX A

CLOCK DRAWING		
<p>➤ Clock: Ask the subject to draw a clock face with numbers and the hands at ten past five. (For scoring see instruction guide: circle = 1, numbers = 2, hands = 2 if all correct).</p>		<p>Visuospatial [Score 0-5]</p> <input type="text"/>
MEMORY RECALL		
<p>➤ Ask "Now tell me what you remember about that name and address we were repeating at the beginning"</p>		
<p>Harry Barnes</p> <p>73 Orchard Close</p> <p>Kingsbridge</p> <p>Devon</p>	<p>.....</p> <p>.....</p> <p>.....</p> <p>.....</p>	<p>Memory [Score 0-7]</p> <input type="text"/>
TOTAL SCORE		/ 30

Updated 25/05/2014

Appendix B: Systematic review search terms used for each database

PsychoInfo

The first search was performed via OvidSP using the following terms:

1. Item response theory/ or “difficulty level (test)”/ or “item analysis (statistical)”/
2. Mokken.tw.
- 3.1 OR 2
4. dementia/ or dementia with lewy bodies/ or vascular dementia/
or Alzheimer’s disease/
5. dementia.tw. or
6. Semantic dementia/
- 7.4 OR 5 OR 6
- 8.3 AND 7

Medline

The search was performed via OvipSP. The search terms used were:

1. “item response theory”.tw. or
2. IRT.tw. or
3. “item response analysis”.tw. or
4. “modern testing theory”.tw. or
5. (cumulative adj2 structure).tw. or
6. “scale construction”.tw. or
7. “guttman scaling”.tw. or
8. “guttman scale”.tw. or
9. Mokken.tw. or
10. rasch.tw or

APPENDIX B

11. uni?dimensional*.tw. or
12. “cumulative order”.tw. or
13. “item characteristic curve”.tw.
14. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13
15. dementia/ or Alzheimer disease/ or dementia, vascular/ or frontotemporal lobal degeneration/ or lewy body disease
16. dementia.tw.
17. 15 OR 16
18. 14 AND 17

Embase

The search was performed via OvipSP using the following search terms:

1. “item response theory”.mp. or
2. Mokken.mp. or
3. IRT.mp. or
4. “modern testing theory”.mp. or
5. (Cumulative adj2 structure).mp. or
6. “scale construction”.mp. or
7. “guttman scaling”.mp. or
8. “guttman scale”.mp. or
9. Rasch.mp. or
10. Uni?dimensional.mp. or
11. “cumulative order”.mp. or
12. “item characteristic curve”.mp. or
13. “item response analysis”.tw.
14. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13
15. Dementia/ or Alzheimer’s disease/ or frontotemporal dementia/ or multiinfarct dementia/
16. Dementia.tw. or
17. Diffuse Lewy body disease/
18. 15 Or 16 OR 17
19. 14 AND 18

APPENDIX B

CINAHL

Search was performed via EBSCO Host gateway. The following search terms were used:

1. TX("item response theory" or "item response analysis")
2. TX (Mokken or IRT)
3. TX ("modern testing theory" or rasch)
4. TX ("scale construction" or "item characteristic curve")
5. TX ("guttman scaling" or "guttman scale")
6. TX "cumulative order"
7. 1 OR 2 OR 3 OR 4 OR 5 OR 6
8. TX (Dementia or "Alzheimer's disease")
9. TX ("vascular disease" or "frontotemporal dementia")
10. TX "lewy body disease"
11. 8 OR 9 OR 10
12. 7 AND 11

APPENDIX C

Appendix C: Conversion of NART, Abbreviated NART and Mini-NART scores to predict premorbid IQ using regression equations derived from this study.

NART Score	Predicted premorbid IQ	Abbreviated NART score	Predicted premorbid IQ	Mini-NART score	Predicted premorbid IQ
50	121.27	38	116.26	23	118.88
49	119.96	37	114.73	22	116.53
48	118.66	36	113.20	21	114.19
47	117.35	35	111.67	20	111.84
46	116.05	34	110.14	19	109.50
45	114.74	33	108.61	18	107.15
44	113.43	32	107.07	17	104.81
43	112.13	31	105.54	16	102.46
42	110.82	30	104.01	15	100.12
41	109.52	29	102.48	14	97.77
40	108.21	28	100.95	13	95.43
39	106.90	27	99.42	12	93.08
38	105.60	26	97.89	11	90.74
37	104.29	25	96.36	10	88.39
36	102.90	24	94.83	9	86.05
35	101.68	23	93.30	8	83.70
34	100.37	22	91.76	7	81.36
33	99.07	21	90.23	6	79.01
32	97.76	20	88.70	5	76.67
31	96.46	19	87.17	4	74.32
30	95.15	18	85.64	3	71.98
29	93.84	17	84.11	2	69.63
28	92.54	16	82.58	1	67.29
27	91.23	15	81.15		
26	89.93	14	79.52		
25	88.62	13	77.98		
24	87.31	12	76.45		
23	86.01	11	74.92		
22	84.70	10	73.39		
21	83.40	9	71.86		
20	82.09	8	70.33		
19	80.78	7	68.80		
18	79.48	6	67.27		
17	78.17	5	65.74		
16	76.87	4	64.21		
15	75.56	3	62.68		
14	74.25	2	61.14		
13	72.95	1	59.61		
12	71.64				
11	70.34				
10	69.03				
9	67.72				
8	66.42				
7	65.11				
6	63.81				
5	62.50				
4	61.19				
3	59.89				
2	58.58				
1	57.28				

Note. NART=National Adult Reading Test.

Appendix D: Functional assessment scales

THE LAWTON INSTRUMENTAL ACTIVITIES OF DAILY LIVING SCALE

Ability to Use Telephone

- 1. Operates telephone on own initiative; looks up and dials numbers1
- 2. Dials a few well-known numbers1
- 3. Answers telephone, but does not dial1
- 4. Does not use telephone at all0

Shopping

- 1. Takes care of all shopping needs independently1
- 2. Shops independently for small purchases0
- 3. Needs to be accompanied on any shopping trip0
- 4. Completely unable to shop0

Food Preparation

- 1. Plans, prepares, and serves adequate meals independently1
- 2. Prepares adequate meals if supplied with ingredients0
- 3. Heats and serves prepared meals or prepares meals but does not maintain adequate diet0
- 4. Needs to have meals prepared and served0

Housekeeping

- 1. Maintains house alone with occasion assistance (heavy work)1
- 2. Performs light daily tasks such as dishwashing, bed making1
- 3. Performs light daily tasks, but cannot maintain acceptable level of cleanliness1
- 4. Needs help with all home maintenance tasks1
- 5. Does not participate in any housekeeping tasks0

Laundry

- 1. Does personal laundry completely1
- 2. Launders small items, rinses socks, stockings, etc1
- 3. All laundry must be done by others0

Mode of Transportation

- 1. Travels independently on public transportation or drives own car1
- 2. Arranges own travel via taxi, but does not otherwise use public transportation1
- 3. Travels on public transportation when assisted or accompanied by another1
- 4. Travel limited to taxi or automobile with assistance of another0
- 5. Does not travel at all0

Responsibility for Own Medications

- 1. Is responsible for taking medication in correct dosages at correct time1
- 2. Takes responsibility if medication is prepared in advance in separate dosages0
- 3. Is not capable of dispensing own medication0

Ability to Handle Finances

- 1. Manages financial matters independently (budgets, writes checks, pays rent and bills, goes to bank); collects and keeps track of income1
- 2. Manages day-to-day purchases, but needs help with banking, major purchases, etc1
- 3. Incapable of handling money0

Scoring: For each category, circle the item description that most closely resembles the client's highest functional level (either 0 or 1).

Lawton, M.P., & Brody, E.M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*, 9(3), 179-186.

Copyright © The Gerontological Society of America. Reproduced [Adapted] by permission of the publisher.

APPENDIX D

PHYSICAL SELF-MAINTENANCE SCALE (ACTIVITIES OF DAILY LIVING, OR ADLs)

In each category, circle the item that most closely describes the person's highest level of functioning and record the score assigned to that level (either 1 or 0) in the blank at the beginning of the category.

A. Toilet		_____
1.	Care for self at toilet completely; no incontinence	1
2.	Needs to be reminded, or needs help in cleaning self, or has rare (weekly at most) accidents	0
3.	Soiling or wetting while asleep more than once a week	0
4.	Soiling or wetting while awake more than once a week	0
5.	No control of bowels or bladder	0
B. Feeding		_____
1.	Eats without assistance	1
2.	Eats with minor assistance at meal times and/or with special preparation of food, or help in cleaning up after meals	0
3.	Feeds self with moderate assistance and is untidy	0
4.	Requires extensive assistance for all meals	0
5.	Does not feed self at all and resists efforts of others to feed him or her	0
C. Dressing		_____
1.	Dresses, undresses, and selects clothes from own wardrobe	1
2.	Dresses and undresses self, with minor assistance	0
3.	Needs moderate assistance in dressing and selection of clothes.	0
4.	Needs major assistance in dressing, but cooperates with efforts of others to help	0
5.	Completely unable to dress self and resists efforts of others to help	0
D. Grooming (neatness, hair, nails, hands, face, clothing)		_____
1.	Always neatly dressed, well-groomed, without assistance	1
2.	Grooms self adequately with occasional minor assistance, eg, with shaving	0
3.	Needs moderate and regular assistance or supervision with grooming	0
4.	Needs total grooming care, but can remain well-groomed after help from others	0
5.	Actively negates all efforts of others to maintain grooming	0
E. Physical Ambulation		_____
1.	Goes about grounds or city	1
2.	Ambulates within residence on or about one block distant	0
3.	Ambulates with assistance of (check one)	
	a () another person, b () railing, c () cane, d () walker, e () wheelchair	0
	1. ___Gets in and out without help. 2. ___Needs help getting in and out	
4.	Sits unsupported in chair or wheelchair, but cannot propel self without help	0
5.	Bedridden more than half the time	0
F. Bathing		_____
1.	Bathes self (tub, shower, sponge bath) without help.	1
2.	Bathes self with help getting in and out of tub.	0
3.	Washes face and hands only, but cannot bathe rest of body	0
4.	Does not wash self, but is cooperative with those who bathe him or her.	0
5.	Does not try to wash self and resists efforts to keep him or her clean.	0

Appendix E: Published papers

McGrory et al. *BMC Psychiatry* 2014, **14**:47
<http://www.biomedcentral.com/1471-244X/14/47>



RESEARCH ARTICLE

Open Access

Item response theory analysis of cognitive tests in people with dementia: a systematic review

Sarah McGrory^{1*}, Jason M Doherty², Elizabeth J Austin², John M Starr^{1,3,4} and Susan D Shenkin^{3,4}

Abstract

Background: Performance on psychometric tests is key to diagnosis and monitoring treatment of dementia. Results are often reported as a total score, but there is additional information in individual items of tests which vary in their *difficulty* and *discriminatory* value. Item *difficulty* refers to an ability level at which the probability of responding correctly is 50%. *Discrimination* is an index of how well an item can differentiate between patients of varying levels of severity. Item response theory (IRT) analysis can use this information to examine and refine measures of cognitive functioning. This systematic review aimed to identify all published literature which had applied IRT to instruments assessing global cognitive function in people with dementia.

Methods: A systematic review was carried out across Medline, Embase, PsychInfo and CINHAL articles. Search terms relating to IRT and dementia were combined to find all IRT analyses of global functioning scales of dementia.

Results: Of 384 articles identified four studies met inclusion criteria including a total of 2,920 people with dementia from six centers in two countries. These studies used three cognitive tests (MMSE, ADAS-Cog, BIMCT) and three IRT methods (Item Characteristic Curve analysis, Samejima's graded response model, the 2-Parameter Model). Memory items were most *difficult*. Naming the date in the MMSE and memory items, specifically word recall, of the ADAS-cog were most *discriminatory*.

Conclusions: Four published studies were identified which used IRT on global cognitive tests in people with dementia. This technique increased the interpretative power of the cognitive scales, and could be used to provide clinicians with key items from a larger test battery which would have high predictive value. There is need for further studies using IRT in a wider range of tests involving people with dementia of different etiology and severity.

Keywords: Item response theory, Dementia, Psychometrics, Cognition, Alzheimer disease, MMSE, Systematic review

Background

Global cognitive functioning measures are the mainstay diagnostic tool for dementia, in conjunction with determination of functional decline, and are also used to track and measure disease course. Measures of cognition in dementia should be able to both reliably detect the disease in its early stages and to evaluate the severity of the disease [1].

The most common method of scoring a cognitive test is to sum the raw score. The total score is used to aid diagnosis and to assess and monitor disease severity. This method is quick and simple to apply and is based on the premise of all test items reflecting a common unobservable trait or

ability range along which cognitive impairment can be measured [2].

However the simple summation of raw scores overlooks any differences between the items and information the pattern of response can provide. It may therefore lead to an inaccurate estimation of cognitive impairment [2].

Items within a measure will differ in several ways. Firstly some items may be more *difficult* than others, for example, for most people, repeating a noun would be less *difficult* than remembering a phrase or list of words. Secondly, some items may be more sensitive to the early stages of cognitive decline and others to the later stages of the disease. Thirdly, items may differ in how sensitive they are to clinical change. Finally, some items may be redundant and provide no meaningful variability to the

* Correspondence: S.McGrory@sms.ed.ac.uk

¹Alzheimer Scotland Dementia Research Centre, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK

Full list of author information is available at the end of the article



© 2014 McGrory et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

measure. These items could be removed to ease the burden on patients and clinicians.

The same total score can be achieved via many different patterns of response. For example, two individuals scoring 20 on the MMSE may have correctly and incorrectly answered completely different items. Likewise an individual obtaining the same total score before and after treatment would be considered as having experienced no change in cognitive impairment even if the pattern of response across the items had changed.

Therefore, there is a need to look beyond the total score and to investigate the pattern of response to the individual items. This can be done using the statistical method 'item response theory' (IRT) [3].

IRT is based on the probability of a person achieving a certain score on a test being a consequence of that person's ability on the latent construct [4]. As that ability, cognitive function in this case, changes, so too does the probability of the individual achieving a certain score, offering measurement precision that varies with ability level [5]. Unlike other statistical methods which use the aggregate raw score as an indication of ability, IRT is more concerned with individual test items.

IRT can provide two useful measures; *difficulty* and *discrimination*, both of which are technical properties of the Item Characteristic Curve (ICC). The ICC is a non-linear regression on ability of probability of a correct response to each item. *Difficulty* is the ability value that is associated with a 50% probability of scoring one (rather than zero) on an individual item [6].

Discrimination, reflecting the slope of the ICC in its middle section, is an index of how well an item can differentiate between patients of varying levels of severity. More *discriminating* items, with a steeper slope, are better able to differentiate among individuals in the range of the latent trait [7].

The performance of the overall scale can be measured using the Test Characteristic Curve (TCC). The TCC is a valuable tool for assessing the range of measurement and the degree of *discrimination* at various points along the ability continuum. Also the extent to which the TCC is linear illustrates the degree to which the scale provides interval scale or linear measurement.

Information is the equivalent of variance explained, showing how effectively a measure captures the latent trait. *Information* can be calculated for each ability level. The greater the amount of *information*, the more precision with which the ability can be estimated.

IRT could improve tests used for diagnosing and monitoring people with dementia. By determining the *difficulty* of items within a scale it is possible to develop a hierarchy of item *difficulty* i.e. a list of questions from those with lowest *difficulty* (where the expected probability of a correct answer of 50% is reached at a low overall score) to

those with highest *difficulty* (where the expected probability of a correct response of 50% is reached at a high score). This confirms the sequence of cognitive decline. Establishing a hierarchy of *difficulty* confirming the sequence of decline will allow clinicians and researchers to identify any deviations in the rate or sequence of cognitive decline from the usual trajectory of loss. Hierarchies of item *difficulty* may differ according to diagnosis or by country/region or by different translations of measures. Identifying unique sequences of cognitive decline for different forms of dementia could aid in diagnoses. Additionally being aware of the ordering of *difficulty* makes it possible for clinicians to tailor their assessments according to severity level, e.g., selecting less *difficult* items for patients with established dementia and the more *difficult* items for healthy elderly or those with mild or early stages of cognitive impairment [8].

IRT can also examine the sensitivities of the items within a measure. By examining the slope of the ICC the items *discrimination* can be assessed. The range of cognitive impairment at which the slope is the steepest is where that item will be maximally *discriminative*, differentiating well between various gradations of impairment and providing increased sensitivity to change. Determining the *discrimination* of items can reveal which items are most likely to expose changes in cognition and those with weaker *discriminatory* power that are unresponsive to such changes [9,10]. Looking at the item curves in relation to each other provides useful information on the breadth of measurement of an instrument. IRT can also identify key items which provide valuable information or whether any items within the scale are redundant, i.e. items with similar ICCs.

Applying IRT techniques to measures of cognitive functioning in dementia could have far reaching implications for clinicians and researchers leading to advancements in screening assessments and diagnosis, the charting of disease course and the measurement of change with disease progression and in response to treatment. In addition, IRT methodology will be useful to industry in the design of psychometric tests. IRT has been used to analyse clinical measures in several different fields: schizophrenia [11], depression [12], attachment [13], social inhibition [14] and quality of life [15]. IRT has also been used to examine ADL and Instrumental Activities of Daily Living (IADL) scales [16,17]. IRT methods have been successful in improving functional scales by establishing interval level measurement [18]; hierarchies of item *difficulty* [16,19,20]; *discrimination* of items [16,21]; as well as identifying ways of increasing measurement precision [18]. IRT analyses of measures of cognitive functioning in the general population have been described [22,23], including several papers with samples including some participants with dementia [24-28]. However, despite the strong theoretical basis outlined above for using IRT in people with dementia, there

is limited published data. Therefore we performed a systematic review of the published studies that use IRT to revise or develop instruments assessing cognitive ability in people with dementia.

Methods

Search strategy

Published studies were identified through searches of Medline (including work in progress from 1946 until 5th September 2013), Embase (1980 until 5th September 2013), PsychInfo (1806 until 5th September 2013) and CINAHL (1981 until 5th September 2013). Search filters included were keyword, title and abstract information. Search terms relating to IRT and dementia were combined. Articles with any combination of any of the IRT terms and any dementia term were reviewed. For full search strategy see Appendix 1. References of included studies were hand-searched and a forward citation search was performed on all included studies to establish all articles which cited them.

Data extraction

A total of 384 articles were identified from this search. After duplicates were removed the titles and abstracts of 203 articles were screened by two independent researchers. 160 articles were excluded on review of title and/or abstract (for example, non IRT methods, IRT analyses of functional or other non-cognitive assessments). 43 articles considered to be relevant were retrieved and assessed for agreement with the following inclusion and exclusion criteria. Data were extracted from original studies onto forms which were refined following piloting.

Figure 1 shows the flow chart for this review.

Inclusion criteria

This review aimed to include all published studies that applied item response theory methods to instruments with face validity for measuring global cognitive impairment in dementia. The initial search did not restrict results to those published in the English language.

Exclusion criteria

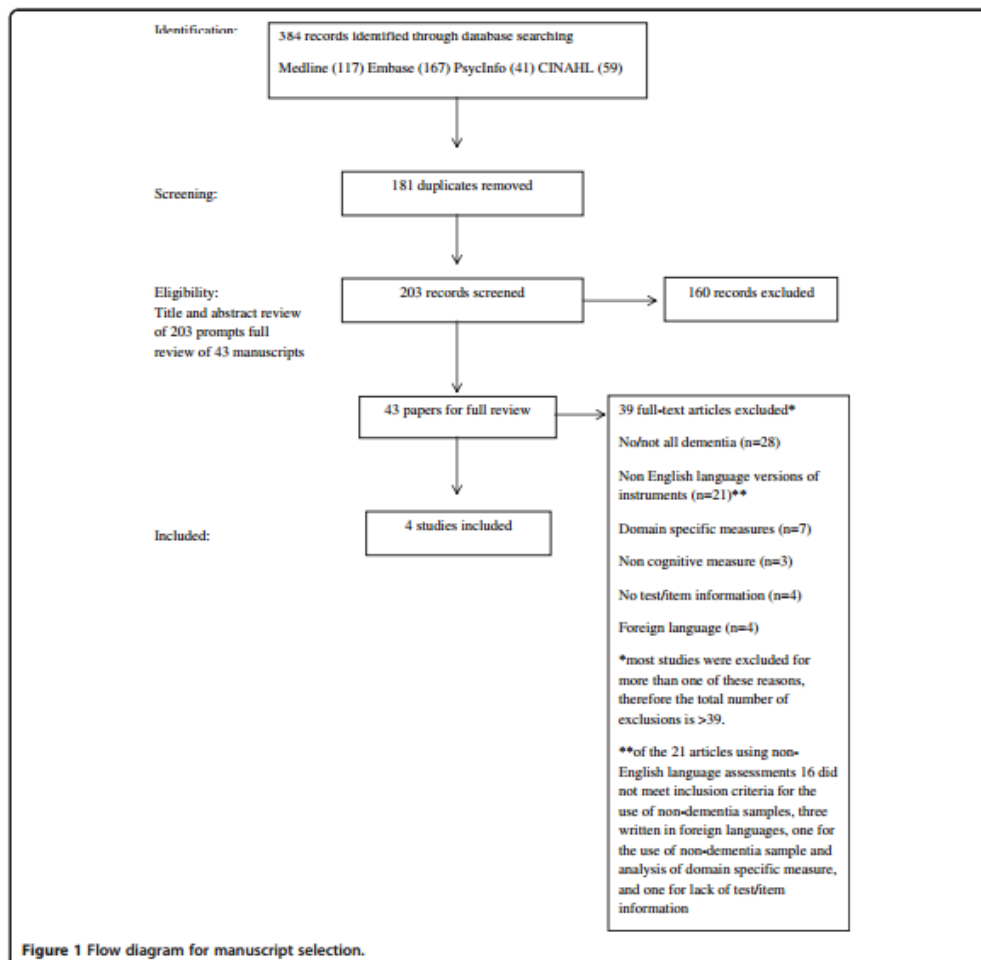
(i) Unpublished studies, dissertations, theses, journal conference abstracts and poster presentations; (ii) studies using proxy reports as there is evidence of discrepancy between self-report and informant measures of cognitive functioning [29]; (iii) studies with participants without diagnosed dementia; (iv) studies without details of dementia diagnosis criteria or percentages of participants with dementia; (v) studies reporting IRT applications to domain specific measures of cognition rather than global cognitive functioning, for example the Boston Naming Test [30] used to measure confrontational word retrieval; (vi) studies that did not provide information on item level performance or overall test

performance; (vii) studies examining non-cognitive scales, although studies which reviewed a range of outcomes had the results from the cognitive scales included; (viii) no language restrictions were made in the search, but non-English language articles were not included in the final review as they used non-English scales; (ix) use of Guttman scaling procedures [31].

While studies have found increased sensitivity of domain specific neuropsychological tests to early impairment than test of global cognition [32] this review chose to restrict its focus to IRT analyses of global cognitive instruments to increase clinical relevance as these are the most commonly used for testing in routine practice.

The decision to exclude Guttman scaling was based on the considerable evidence stating the inferiority of these methods in comparison to the more advanced item response methods [33]. The method was included in the search strategy; however, as some studies may have applied another method of analysis without indexing it and the exclusion of this term may have led to some relevant studies being overlooked.

Non-English language versions of cognitive measures were excluded. While several measures, most notably the MMSE [34], have been translated into many languages for use in different countries and cultures there are concerns over the cross-cultural validity. The language in which a test is administered can affect performance leading to a potential overestimation of cognitive impairment in individuals who do not speak English [35-37]. Differential Item Functioning (DIF) [38] can be applied to examine the effect of language bias of items and tests administered in different languages. For example, if patients of equal cognitive ability tested in English and Spanish have unequal probabilities of responding correctly to a particular item on a cognitive assessment, then the item functions differently with respect to language. The effect of different test languages of cognitive assessments has been examined in this way [27,39-43]. However these studies did not examine DIF in dementia populations and were therefore not included. Also the non-English language versions administered makes comparison with scales in English problematic because the semantic range of items cannot be assumed in translation [44], for example, repeating "No ifs, ands, or buts" corresponds to repeating "We put ones' efforts all together and pull the rope" in the Japanese version of the MMSE [26] and to a tongue-twisting phrase "en un trigal habia tres tigres" ("there were three tigers in a wheat field") in the Spanish version [45]. To avoid any potential confounding these articles were not included for full review [43]. The decision to exclude articles using non-English language assessments has no implications for the validity of cognitive testing in other languages.



Results

Four cross-sectional studies met inclusion criteria, including 2,920 patients from six centers in two countries: Table 1 describes the characteristics of the studies reviewed. In total dementia aetiologies comprise 74.1% (2165) probable Alzheimer's disease (AD), 9.3% (273) possible AD, 2% (60) Vascular dementia, 11.1% (325) mixed and other dementia. For individual studies see Table 1. Most patients fall within the moderate range of severity of dementia.

Three cognitive tests (MMSE, ADAS-cog, BIMCT) and three different IRT methods (Item Characteristic Curve analysis, Samejima's graded model, Two-parameter model) were used.

Ashford *et al.* [46] applied IRT techniques to identify the degree of AD severity at which individual items of the MMSE are lost and the rate at which they are lost at that level of severity. MMSE scores from 86 AD patients were analysed. Most people had moderately severe AD (mean MMSE score = 18).

A hierarchy of item *difficulty* was formed (see Table 2). Most *difficult* items were the three memory items and "Orientation to date" (which also tests recent memory), and "Serial sevens". These findings suggest that the mental functions assumed to underlie performance of these items—memory and attention and calculation— are lost earliest in the progression of AD. Least *difficult* items, i.e. late loss,

Table 1 Articles meeting inclusion criteria applying IRT methods to cognitive measures of dementia

Study	Ashford <i>et al.</i> [46]	Mungas & Reed [1]	Gibbons <i>et al.</i> [48]	Benge <i>et al.</i> [49]
Country	USA	USA	USA and UK	USA
Setting	Geriatric psychiatry Outpatient clinic	Two clinical sites of Alzheimer's disease centre	Two community based samples from USA and UK	Alzheimer's disease and memory disorders clinic
N	86	1207	540 (US: 401, UK: 139)	1087
Sex	73.2% female	64.7% female	(US) 64% female (UK) 75% female	66.6% female
Age			(US) (UK)	
Mean	74	76	82 84.7	75
SD	8	8.9	4.7 5.3	8.1
Range	53-91	39-100	> 75 >75	Not reported
Etiology; n (%)	Probable AD: 52 (60) Possible AD: 34 (40)	Probable AD: 592 (49.0) Possible AD: 176 (14.6) Vascular: 60 (5.0) Mixed and other dementia: 325 (26.9) No cognitive impairment: 27 (2.2) Diagnosis deferred: 27 (2.2)	UK: AD: 139 (100) US: Probable AD: 338 (84.2) Possible AD: 63 (15.7)	AD: 1044 (96) MCI: 43 (4)
Dementia severity	Mean MMSE=18 SD=7.1 Range=1-29	Mean MMSE= 17.7 SD=7.3 Range=0-30 Mean BIMCT= 16.9 SD=8.3 Range=0-33	US: Mean MMSE=19.6 SD=4.9 Range=1-29 UK: Mean MMSE=16.5 SD=5.5 Range=0-25	Mean ADAS cog=31.2 SD=16.5 Range=not reported
Cognitive measure	MMSE	MMSE, BIMCT	MMSE	ADAS-cog
IRT method	Item characteristic curve analysis	Two-parameter model	Item characteristic curve analysis	Samejima's graded model
Outcome	Hierarchy of item difficulty and discrimination	Hierarchy of item difficulty of Global function scale. Investigation of linearity of MMSE, BIMCT and global function.	Hierarchy of item difficulty from 2 samples	Discrimination and information statistics on ADAS-cog test as whole, plus domains and subscales

AD = Alzheimer's disease, MCI = mild cognitive impairment, MMSE = Mini Mental State Examination, ADAS-cog = Alzheimer's disease Assessment Scale-Cognitive subscale, BIMCT = Blessed Information Memory Concentration Test.

were "Verbal directions", "Name pencil" and "Repeat nouns". This pattern is consistent with the typical clinical course of AD starting with memory problems ultimately leading to problems with over-learned associations and early-learned verbal mimicking.

For one of the least *difficult* items "Name pencil" participants with a score of 6.6 had a 50% probability of getting this item correct. At a score of 10 participants had an almost 100% chance of correctly identifying the pencil. This is in sharp contrast to the most *difficult* items "Recall nouns". A participant with a score of 20 had approximately 25% chance of getting "Recall: Tree" correct. These recall items were answered incorrectly by approximately 83% of the participants.

Item *discrimination* was used as an index of the rate of loss. The most *discriminatory* items were: "Name pencil", "Write sentence", "Orientation to month", "Name watch", "Orientation to date", "Orientation to year", "Close eyes". For these items there is a sharp cut-off of ability level at which the item was passed or failed. The items with the lowest *discriminative* power are those items lost earliest; "Recall: Tree" and "Recall: Flag", and latest in disease course; "Verbal directions". Due to these items assessing abilities which are either lost almost immediately or not until very late stages the rate of loss is not meaningful but the items do serve a useful purpose as they measure ability at either extreme of the MMSE scale.

Table 2 Item difficulty comparison across studies

	Ashford et al. [46] (MMSE)	Gibbons et al. [48] UK (MMSE)	Gibbons et al. [48] US (MMSE)	Mungas and Reed [1] (BIMCT/MMSE)
Truncated above upper limit	Recall: tree Recall: flag	No ifs ands or buts Recall nouns		
1st quartile (most difficult)	Serial sevens: subtraction 5 Serial sevens: subtraction 3 Orientation to date Recall: Ball	Orientation to date Verbal directions Intersecting pentagons Serial sevens	Orientation to date No ifs ands or buts Intersecting pentagons Serial sevens	Recall '42' (BIMCT) Recall 'Market Street' (BIMCT) Recall 'John' (BIMCT) Recall 'Chicago' (BIMCT) Recall 'Brown' (BIMCT)
2nd quartile	Serial sevens: subtraction 4 Serial sevens: subtraction 2 Orientation to day Orientation to county Orientation to month Serial sevens: subtraction 1 Orientation to year Orientation to season Orientation to place Orientation to floor	Orientation to year Orientation to county/streets Orientation to day Orientation to month	Recall nouns Orientation to day Orientation to year Orientation to season Orientation to month Orientation to county/streets	Orientation to year (BIMCT/MMSE) Orientation to month (BIMCT/MMSE) Age (BIMCT)
3rd quartile	Orientation to city Intersecting pentagons Orientation to state Write sentence No ifs ands or buts Name watch Verbal directions: paper-on floor	Orientation to state/county Write sentence Orientation to Season Orientation to Address	Orientation to address Verbal directions Write sentence Orientation to place Orientation to city	Orientation to state (MMSE) Type of work (BIMCT) Count forward (BIMCT) Name watch (MMSE)
4th quartile (least difficult)	Close eyes Repeat: flag Name pencil Repeat: ball Repeat: tree Verbal directions: paper-take in right hand Verbal directions: paper-fold in half	Repeat nouns Orientation to city/town/village Orientation to room	Orientation to state Close eyes Name objects	Place of birth (BIMCT) Name pencil (MMSE) Name (BIMCT)
Truncated below 0		Close eyes Name objects	Repeat nouns	

MMSE = Mini Mental State Examination, BIMCT = Blessed Information Memory Concentration Test.
 Ashford et al. [46] and Gibbons et al. [48] test items divided into quartiles based on range of scores. Mungas and Reed [1] items divided into quartiles based on difficulty parameters.
 Most difficult items were truncated above upper limit as difficulty estimates were above the upper limit. Easiest items were truncated below 0 as even this low level of ability most participants were able to answer correctly.
 Some differences between MMSE versions between studies led to some discrepancies between items, e.g.: state/county.

Some limitations of this study include the fact that participants with possible AD were not excluded for sensitivity analysis. Also there was no explicit investigation of unidimensionality of the MMSE. However the item-by item analysis of the variability in AD

implies that there is a strong unidimensional component in the course of AD. There was no report of who administered the MMSE to the participants and whether they were blind to diagnoses. This introduces potential for bias.

Mungas and Reed [1] analysed MMSE and Blessed Information Memory Concentration Test; BIMCT [47] scores from 1207 participants. A very broad range of cognitive impairment across the full range of MMSE and BIMCT scores was represented. Here IRT methods were employed to evaluate existing measures and to develop a new global functioning measure by selecting items from the existing scales with *difficulty* ranges spanning the breadth of ability levels to increase *discrimination* at all ability levels.

Items were recoded as dichotomized variables for analysis. Ordinal scale items such as "World backwards" in the MMSE were converted to a number of dichotomous items equal to the maximum score on this item, leading to total scores of 30 for the MMSE, 33 for the BIMCT. Cognitive tests were administered by a neuropsychologist, neuropsychology trainee or a trained psychometrist. The authors did not mention if these individuals were blind to diagnoses.

Test characteristic curves (TCCs) for both scales were generated. TCCs of the MMSE and BIMCT were distinctly non-linear, showing decreased *discrimination* at both ends of the ability continuum with linear measurement for moderate levels of impairment. This indicates relative insensitivity to changes in ability at each end of the ability spectrum.

A more linear brief composite instrument; 'Global Function' was created. Items were selected from the MMSE, BIMCT and a functional measure; Blessed-Roth Dementia Rating Scale (BRDRS). Items fitting uniform distribution of *difficulty* across the spectrum of ability measured by the three measures were selected. The new scale showed improved *discrimination* at low ability levels but due to the relative absence of high *difficulty* items in the MMSE, BIMCT and BRDRS the scale showed decreased *discrimination* at high ability levels. This illustrates the need to develop and add more *difficult* items to existing and new measures to decrease ceiling effects. The hierarchy of item *difficulty* of the cognitive items from this measure is provided in Table 2. While this measure included functional items which is beyond the scope of this review the most *difficult* items were memory items which is in line with previous findings.

Again there was no assessment of whether the items in the tests are sufficiently unidimensional for the use of IRT. It was not reported whether those who tested the participants were involved in the analysis.

Gibbons *et al.* [48] used IRT to compare the relative *difficulties* of MMSE items between people with AD living in the US and UK. The 401 US participants were comparatively less impaired (mean MMSE 19.6) than the 139 UK participants (mean MMSE 16.5).

There were some differences between items used for the two samples. Orientation to state and county items in US sample were substituted for orientation to county

and 2 streets nearby for the UK cohort and the nouns to repeat and remember were also different for the two cohorts. Although these differences limit the direct comparison of difficulty between these items as the differences are limited to these items they are unlikely to explain the entire difference observed between the two samples. Reports indicate the interview structures did not differ between samples in any substantial way. For analysis all items which could have a score greater than one were dichotomized. All three nouns must be repeated and all stages of following the verbal directions must be carried out for these items to be scored as correct. "Recall nouns" was scored correctly if any one of the three nouns were recalled. Two points for "Serial sevens" were sufficient to be scored as correct. Therefore ability level was represented by the score of the 19 dichotomized items, excluding the score of the item under assessment resulting in score ranges from 0–18.

Gibbons *et al.* [48] established the relative *difficulties* of items for both cohorts, adjusted to an education level of high school or less.

UK results

The most *difficult* items were "No ifs, ands or buts" and "Recall nouns". At the uppermost score of 18 only an estimated 29% of participants could repeat the phrase "No ifs, ands or buts".

The easiest items were "Close eyes" and "Name objects". Here at an estimate of less than zero most participants could still answer correctly so again these estimates were truncated at 0. This reflects the relative simplicity of these items.

US results

The most *difficult* items were "Orientation to date" and "No ifs, ands or buts". At ability scores of 17.5 and 15.3 half of the participants could correctly identify the date and repeat "No ifs, ands or buts" respectively.

The easiest item was "Repeat nouns". The ability score was again truncated at 0 indicating that even at this low level of ability most participants were able to answer correctly. "Name objects" and "Close eyes" were also relatively easy items.

Hierarchies of item *difficulty* for both UK and US samples are presented in Table 2. Five items; "No ifs, ands or buts" "Recall nouns", "Orientation to state/county", "Repeat nouns" and "Verbal directions" were significantly more *difficult* for the UK sample. While some items were more *difficult* for the US cohort the differences were not significant. A score of 15.6 was necessary for a UK participant to have a 50% chance of correctly responding to "Verbal directions" in comparison to a US participant having the same probability at a score of seven.

Additional analyses excluding 'possible' AD, MMSE items which differed between samples, and accounting

for international differences in educational standards did not affect the results.

Attempting to control for the differing levels of severity between the samples, dementia severity (as assessed by the Dementia Rating Scale; DRS) along with age, education and gender were assessed as possible confounders of the relative *difficulty* of items. The relative *difficulty* of the items was not affected by the DRS. It is possible however that controlling for the DRS may not have been enough to compensate for the differences between the two groups.

The methodology applied here was rather robust given the additional analyses performed. However the researchers did not explicitly investigate unidimensionality of the instruments. The MMSE was administered at home by trained research interviewers for both cohorts. The scores used were taken from interviews preceding diagnosis which eliminated risk of bias. The diagnoses were not made by the researchers doing the analysis again limiting any potential bias.

Benge *et al.* [49] used IRT analyses to examine the measurement properties of the ADAS-cog across the spectrum of cognitive decline in AD. To determine the relationship between the level of impairment and the probability of achieving observed scores on the test as a whole and the test's subscales scores from 1087 AD participants were analysed. 43 patients with mild cognitive impairment (MCI), diagnosed using Petersen *et al.*, [50] criteria, were included. This is the only study to include MCI participants and although they account for only 4% of the sample it is worth keeping this difference in mind when interpreting the results. The mean ADAS-cog score was 31.2 indicative of moderate to severe dementia.

Benge *et al.* [49] assessed the unidimensionality of the ADAS-cog. Results from an exploratory factor analysis and confirmatory factor analysis confirmed the ADAS-cog as a one-factor scale.

The measure's subscales were grouped into three domains: memory, praxis and language for analysis. Curves permitting the comparison of the domain performance across the spectrum of cognitive decline were created. These curves indicate that memory has most *discriminative* power at the relatively milder stages of decline in comparison to language and praxis which were maximally *discriminative* at the same stages later in the disease course.

Analysis of the 11 subscales showed "Word recall" to be the most *discriminative* at mild stages of disease making it the best indicator of mild cognitive decline. "Recall of instructions" remained relatively unaffected until the later stages of disease. Praxis and language subscale curves indicate that as with the domains, these subscales maximally *discriminate* at moderate levels of decline. The curves for "Ideational praxis", "Construction" and "Word finding", "Speech comprehension", "Commands", "Speech content" and "Naming" overlap considerably implying that they yield

more or less the same information about patient's stage of cognitive decline. All items *discriminate* well at moderate levels of severity.

Information analysis found perhaps not surprisingly the highest level of *information* is found at moderate levels of cognitive dysfunction. At this level a unit change in cognitive dysfunction represents a greater change in performance than the same change at either ends of the range. This indicates that the ADAS-cog as a whole has relatively high levels of *discrimination* and can differentiate between various degrees of ability at this moderate stage.

This study was the only one to report an assessment of unidimensionality prior to IRT analyses. This is an important assumption underlying IRT theory and it is therefore important to have established that the ADAS-cog meets this assumption.

Analyses were carried out using the most recent of the patients' ADAS-cog scores. It was not reported whether the researchers who carried out the analysis also scored and diagnosed the patients. This introduces some possibility of bias.

Discussion

This is the first systematic review of studies applying IRT methods to the assessment of cognitive decline in dementia. This review employed a comprehensive search strategy and included a detailed narrative review of the studies meeting the inclusion criteria.

This review appraised four published studies of IRT analyses of the cognitive decline of 2,920 participants with dementia. The four studies reviewed provided demonstrations of the applicability of IRT to assessment of cognitive functioning in dementia.

Item difficulty

Three of the four studies established a hierarchy of item *difficulty* [1,46,48]. Two of these hierarchies were of the MMSE items [46,48] and the third was of the Mungas and Reed 'Global Function' scale [1]. The dichotomization of MMSE items in Gibbons [48] decreased the ease at which direct comparisons of item *difficulties* between different studies could be made. In an attempt to equate the different range of MMSE scores across the studies items were divided into quartiles based on score ranges and *difficulty* parameters.

Table 2 shows that "Orientation to date", "Recall nouns" and "Serial sevens" are consistently the most *difficult* items across studies. A clinician identifying problems with these tasks could expect the patient to develop further cognitive *difficulties* in the progression suggested by the hierarchies in Table 2. Generally the least *difficult* items were; "Name objects", "Repeat nouns" and "Close eyes". Problems with these items can help identify severe dementia. From a clinical perspective this information is very useful. It provides

a clearer insight into decline than the traditional scoring method. *Difficult* items are very informative as it is likely that a patient with no *difficulties* here will not have limitations with other less *difficult* items. The items most consistently found the least *difficult* could be used in a similar fashion. It is likely that a patient unable to correctly respond to these items would have problems with most of the other items in the scale. In this way IRT analyses can identify key items from a scale that can quickly inform clinicians of a patient's level of functioning, for example, a clinician could select from the most *difficult* items such as "Recall nouns" to identify potential early cognitive difficulties in the healthy elderly.

None of the studies attempted to determine whether the hierarchies of *difficulty* held at the individual level (ordering items in terms of *difficulty* does not necessarily mean the ordering is the same for every person; those with higher levels of ability may find one item more *difficult* than the other yet the ordering may be reversed for those with lower ability levels [46,51] by considering invariant item ordering (IIO). As invariantly ordered hierarchies are of great clinical value this should be included in future studies.

Discrimination

Two studies determined item *discrimination* [46,49]. Table 3 summarises the findings from these papers, showing the most *discriminatory* items at the various stages of disease. High *discrimination* for low *difficulty* items indicates that the abilities assessed by these items are lost at an advanced stage and that these losses are rapid once this stage has been reached. For more *difficult* items high *discrimination* means that these abilities are lost in the early stages and quickly at this stage.

Items with low *discrimination*; "Repeat nouns", "No ifs, ands or buts", "Orientation to day and season", "Orientation to country, floor and city", "Copy pentagons" also reveal valuable insights. For these items the range of scores

in which participants respond either correctly or incorrectly is wider than high *discriminating* items. Either the abilities being measured by these items are lost with more variability or more gradually or the functions measured here are assessed less concisely by these items.

Including more items like "Word recall" and "Orientation to date" may help to detect changes in milder stages of the disease as these abilities are lost quickly at an early stage. For severe dementia the inclusion of simple repetition tasks or non-cognitive functioning tasks could help to introduce greater *discrimination* in this stage. Items such as recalling or recognizing one's name, from the Severe Cognitive Impairment Rating Scale, measuring the ability of overlearned autobiographic memory, could be applied to broaden the range of assessment in cognitive instruments.

From a large battery of items those demonstrating the best *discrimination* across the disease course could be used to create an instrument to accurately measure patients in early and late stages. More precise assessment would lead to enhanced measurement of the rate of decline and improve predication of impending deterioration.

While these studies demonstrate the use of IRT to examine item *difficulty* and *discrimination* the investigation of item differences has also been addressed using classical test theory (CTT). Chapman and Chapman [52] identified the need to study these item parameters in their analyses of specific and differential deficits in psychopathology research, for example, specific deficits in schizophrenia or the analysis of domains or abilities which remain relatively intact in dementia. Chapman and Chapman's analyses of differential deficits is rooted in classical test theory (CCT) and IRT, as a newer statistical model, offers alternative means of exploring the differential deficit problem. When examining differential deficits between different groups IRT, unlike CCT, can offer estimates of measurement error for different levels of cognitive ability, without having to conduct separate studies, and can establish whether different items or measures are equally *difficult*.

Table 3 High discrimination items and disease stages

	Early disease/high difficulty	Moderate stages	Late disease/low difficulty
High discrimination	"Orientation to date" (MMSE)	ADAS-cog	"Name pencil" (MMSE)
	"Word recall" (ADAS-cog)	"Ideational praxis" (ADAS-cog)	"Close eyes" (MMSE)
		"Construction" (ADAS-cog)	"Name watch" (MMSE)
		"Word finding" (ADAS-cog)	
		"Speech comprehension" (ADAS-cog)	
		"Commands" (ADAS-cog)	
		"Speech content" (ADAS-cog)	
		"Naming" (ADAS-cog)	

Linearity and the assessment of change in severity

Two studies investigated whether the magnitude of cognitive dysfunction represented by each item on the cognitive scale was equal across the scale [1,49]. In a recent paper Balsis *et al.* [53] also drew attention to the limitations associated with the traditional method of measuring cognitive dysfunction with the ADAS-cog. This study was not included in the review as it did not provide information on the individual items or subscales however its analysis of IRT scoring of the ADAS-cog is worth noting. Balsis *et al.* [53] found that individuals with the same total score can have different degrees of cognitive impairment and conversely those with different total scores can have the same amount of cognitive impairment. These findings are supported by a similar study also failing to meet inclusion criteria due to some use of non-English language measures and a lack of information on test/item information [2]. Results indicate that participants with equal ADAS-cog scores had distinctly different levels of cognitive impairment. Equally, participants with the same estimated level of impairment had wide ranging ADAS-cog scores. The same differences in scores did not reflect the same differences in level of cognitive impairment along the continuum of test score range. Without equal intervals between adjacent test items change scores may reflect different amounts of change for subjects with differing levels of severity, or may fail to identify change at all [54]. Wouters *et al.* [2] revised the ADAS-cog scoring based on the results of this IRT analysis by weighting the items in accordance with their measurement precision and by collapsing their categories until each category was hierarchically ordered, ensuring the number of errors increase with a decline along the continuum of cognitive ability. Examining *difficulty* hierarchies of the error categories within the items revealed some disordered item categories. As the categories are only useful if they have a meaningful hierarchy of *difficulty* these disordered categories were collapsed until all categories were correctly ordered in hierarchies of *difficulty*. This revision resulted in a valid one to one correspondence between the summed ADAS-cog scores and estimated levels of impairment.

These studies demonstrate the potential to misinterpret test scores due to a lack of measurement precision. This is illustrated by Mungas and Reed's examination of linearity of the MMSE, BIMCT and the 'Global Function' scale [1]. The findings of non-linearity of the MMSE and BIMCT indicate that a change in total score is less for a given specified change in ability at the two ends of ability distribution than it is in the middle of the ability distribution. For example, a two standard deviation change in ability from 3.0 to 1.0 reflects an approximate five point MMSE score loss, whereas the same degree of change from 1.0 to -1.0 represents a 15 point MMSE score loss. A similar pattern was found for the BIMCT. IRT methods can be used to

create a scale with greater linearity by establishing item *difficulties*, as illustrated by the 'Global Function' scale [1]. The 'Global Function' scale shows promise of linear measurement throughout the majority of the continuum of ability. This new measure, along with any new IRT measure, would need to be cross-validated and directly compared to existing clinical instruments to ensure this test development technique is truly beneficial. It is worth noting that this measure also incorporates items assessing independent functioning. The inclusion of tasks such as these with meaningful variability even in the late stages of dementia could afford the test more *discriminatory* power increasing the information at this stage. While this review did not aim to include functional scales this study suggests that scales that combine cognitive and functional items, or concomitant use of both types, may provide added value. A limitation of this and many other cognitive functioning scales is the lack of items sensitive to very mild early stage of dementia. The inclusion of items capable of *discriminating* mild dementia could improve measurement properties in much the same way.

The measurement properties of a scale can impact the interpretation of clinical trials as change scores are used to determine the efficacy of interventions and treatments. A Cochrane review of AD pharmaceutical trials methods included ADAS-cog change scores to help ascertain the effectiveness of cholinesterase inhibitors [55]. Bengt *et al.* [49] confirmed that the degree of cognitive ability symbolized by each point on the ADAS-cog was not uniform across the scale. A three point change in raw scores can represent a change in cognitive abilities ranging from 0.85 standard deviations of cognitive functioning (representing a change from a score of 4 to 1) to 0.14 standard deviations of cognitive functioning (from a score of 37 to 34).

The observation of differences between and within people may be greatly aided using an IRT approach. In clinical trials it is possible that these analyses will lead to an increased ability to correctly identify group treatment differences and to recognize responders and nonresponders to treatment.

Information

Another advantage of IRT is the increased reliability it provides however, only Bengt *et al.* [49] estimated the *information* parameter. The ADAS-cog has the highest level of *information* at moderate levels of cognitive impairment. At milder levels of impairment the *information* function remains low which indicates that the test domains; language, memory and praxis, and the measure as a whole do a relatively poor job *discriminating* among the different levels of impairment in the mild severity range. The same can be said about the severe levels of impairment. That moderate levels have the highest *information* function is unsurprising as the ADAS-cog was originally designed to measure moderate AD. Decreased

information at mild and severe levels could affect the interpretation of the significance of the change scores at these levels of impairment.

This review excluded 28 studies using general populations, some of which included some dementia subgroups. In an effort to widen the scope of the review studies using general populations including some participants with dementia were looked at to determine if these dementia subgroups could be analysed separately. However it was determined that these papers failed to meet inclusion criteria for reasons beyond the sample characteristics, mostly for the use of non-English language measures, and therefore the authors of the papers were not contacted for further details. One such study analysed a Japanese version of the MMSE within a general population [26]. However the ordering of items was examined for the AD subgroup in isolation illustrating the sequence of cognitive decline. IRT analysis found the scale could be simplified with the removal of items showing similar ICCs and factor loadings, reflecting potential redundancy. "Naming" was deemed to be similar to "Three-step command" and was deleted along with "Read and follow instruction" showing similarity to "Repeat a sentence" and "Orientation to time" as its function was comparable to "Orientation to place". The ordering from least to most difficult was "Three-step command", "Registration", "Repeat a sentence", "Write a complete sentence", "Copy drawings of two polygons", "Delayed recall", "Orientation to place" and "Serial sevens".

21 studies were excluded for administering non-English measures. However, all except one were excluded for other reasons also (16 did not meet inclusion criteria for the use of non-dementia samples, three written in foreign languages, one for the use of non-dementia sample and analysis of domain specific measure, and one for lack of test/item information). The results of the single study [56] which was only excluded due to use of a Dutch version of the Baylor Profound Mental State Examination are discussed. Korner *et al.* [56] applied Mokken analysis and the one-parameter Rasch analysis in a validation study of the cognitive part of the Danish version Baylor Profound Mental State Examination. In doing so the relative difficulty of the test items were estimated. The difficulties of the 25 items were evenly distributed along the ability range with no redundant items. The least difficult items in this measure were; "What is your name?" and the repetition of the first word (one syllable). The most difficult item was the drawing of "Intersecting pentagons". While the other studies administering such measures would not have been included for various other reasons there are data that may be informative [24,26,28,57].

While global cognitive instruments such as the MMSE are probably the most commonly used measure of cognitive functioning domain specific neuropsychological tests have been demonstrated to show increased sensitivity to

early stages of cognitive impairment than measures of global cognition [32]. However of the seven studies applying IRT methods to domain specific measures identified [40,58-63] only one; Benge *et al.* [58] otherwise met inclusion criteria. This study's findings were briefly discussed here. Temporal ("Day of month", "Year", "Month", "Day of week" and "Season"), and spatial ("Name of hospital", "Floor", "Town", "Country" and "State") Orientation items of the MMSE, were analysed to determine their difficulty and discrimination parameters. The most difficult item was "Floor of hospital" and the least difficult item was "State". The full order of item difficulty was; "Floor", "Name of Hospital", "Date", "Day of Week", "Year", "Month", "Season", "Country", "Town" and "State". A relatively high level of ability (2.81SD) is required to have a 95% chance of correctly identifying the floor of the building which illustrates that knowing which floor of the hospital reflects a relatively high level of cognitive ability. Clinicians can use this sort of knowledge to help interpret the information they get from their assessments.

The spatial orientation items discriminate best at varying levels of cognitive ability with a wider range of difficulties assessed than the temporal items. Spatial items could be used to create a short scale sensitive to a relatively broad range of abilities. The temporal items assess a narrower breadth of abilities at a relatively modest degree of impairment and therefore would be best suited to identifying change within this range of cognition.

The value contributed by each item was examined to reveal key items and those whose function was largely redundant. "Year" and "Month" provide roughly the same information as they have similar levels of discrimination and difficulty, as do "State" and "Town". Both item pairs provide no meaningful variability to the set of items. One item from each pair would be sufficient to capture the same information as both. "Date", "Name of Hospital" and "State" together sample the range of cognitive abilities assessed by the orientation items and could together provide key information about a wide range of abilities.

Some limitations of this review should be acknowledged

While the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed insofar as they were applicable for methodological studies the studies identified in this review did not allow a statistical summary or to perform a meta-analysis due to the variety of subjects, sites, diagnostic criteria and the varied statistical item response theory methods applied. The four studies cross a 20 year span with the earliest data collection and diagnoses in 1984 [46] with the most recent in 2002 [49]. This will affect criteria for diagnosing dementia. With mostly moderate ranges of dementia the studies also represented a rather restricted range of

severity limiting the scope of the analysis as the findings cannot be extrapolated to mild or severe dementia.

IRT analyses assume unidimensionality which limits its application to measures assessing a single latent construct. However only one study reviewed here explicitly assessed unidimensionality prior to IRT analyses [49].

Three of the four studies failed to report who administered the test to participants and whether these individuals were blind to the diagnoses [1,46,49]. This introduces some potential bias in these studies.

This review was limited to analyses of only three global cognitive function; MMSE, BIMCT and ADAS-cog. This was a consequence of the articles meeting inclusion criteria. However, an analysis of the Baylor Profound Mental State Examination, while not reviewed due to use of a Dutch version, was mentioned in the discussion [56].

With the exception of Mungas and Reed [1] all studies solely included patients with Alzheimer's disease. This could have an impact on findings as there should be a different pattern of decline between different aetiologies. Of the excluded articles one included patients with amyotrophic lateral sclerosis and behavioural variant frontotemporal dementia which would have expanded the scope of this review [64]. However this study failed to provide data on the measure of cognition in isolation from the other outcomes studied and for this reason was excluded.

Conclusion

This systematic review of IRT use in cognitive tests in people with dementia found only four relevant published papers. These include heterogeneous populations, with widely varying sample sizes, different methods of dementia diagnosis (and inclusion of possible dementia or MCI), and samples are mostly derived from specialist clinical populations, with a risk of inclusion bias. Most participants had Alzheimer's dementia of moderate severity, and were resident in the United States, so the relevance of this method to other subtypes of dementia, and other countries, cannot be determined. Different cognitive tests, and IRT methods, were used, and different statistics were reported. However, the studies show that IRT can demonstrate which items within scales are most *difficult*, and *discriminatory*, at different severities of dementia. IRT analyses can also be used to reveal non-uniform distances between scale scores and facilitate the creation of scales with enhanced measurement properties allowing more accurate assessment of change across the ability spectrum.

There is a need for more IRT analyses of cognitive scales used to assess dementia. These should include standard methodologies, and report item *difficulty* and *discriminatory* statistics along with a measure of *information* and an assessment of linearity of measurement. They should include large numbers, from a variety of

countries (both English speaking and non-English-speaking), different dementia subtypes, the full range of severity of dementia, and a wider range of cognitive tests, focusing on those that are widely used in clinical practice. This will allow refinement of these tools to improve the information provided to clinicians on how performance on items within the scale is informative at different stages in dementia.

Appendix 1

Search strategy

PsychInfo

IRT terms:

1. Item response theory/or "difficulty level (test)"/or "item analysis (statistical)"/
2. Mokken.tw.

Dementia terms:

3. dementia/or dementia with lewy bodies/or vascular dementia/ or Alzheimer's disease/
4. dementia.tw. or
5. semantic dementia/

Medline

IRT terms:

1. "item response theory".tw. or
2. IRT.tw. or
3. "item response analysis".tw. or
4. "modern testing theory".tw. or
5. (cumulative adj2 structure).tw. or
6. "scale construction".tw. or
7. "guttman scaling".tw. or
8. "guttman scale".tw. or
9. Mokken.tw. or
10. rasch.tw or
11. uni?dimensional*.tw. or
12. "cumulative order".tw. or
13. "item characteristic curve".tw.

Dementia terms:

14. dementia/or Alzheimer disease/or dementia, vascular/or frontotemporal lobar degeneration/or lewy body disease
15. dementia.tw.

Embase

IRT terms:

1. "item response theory".mp. or
2. Mokken.mp. or

Lawton IADL scale in dementia: can item response theory make it more informative?

SARAH McGRORY¹, SUSAN D. SHENKIN^{2,3}, ELIZABETH J. AUSTIN⁴, JOHN M. STARR^{1,2,3}

¹Alzheimer Scotland Dementia Research Centre, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK

²Geriatric Medicine, University of Edinburgh, Edinburgh, UK

³Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

⁴Psychology, University of Edinburgh, Edinburgh, UK

Address correspondence to: Sarah McGrory. Tel: 01 31 650 4340; Fax: 01 31 650 3461. E-mail: S.McGrory@sms.ed.ac.uk

Abstract

Background: impairment of functional abilities represents a crucial component of dementia diagnosis. Current functional measures rely on the traditional aggregate method of summing raw scores. While this summary score provides a quick representation of a person's ability, it disregards useful information on the item level.

Objective: to use item response theory (IRT) methods to increase the interpretive power of the Lawton Instrumental Activities of Daily Living (IADL) scale by establishing a hierarchy of item 'difficulty' and 'discrimination'.

Methods: this cross-sectional study applied IRT methods to the analysis of IADL outcomes. Participants were 202 members of the Scottish Dementia Research Interest Register (mean age = 76.39, range = 56–93, SD = 7.89 years) with complete itemised data available.

Results: a Mokken scale with good reliability (Molenaar Sijtsma statistic 0.79) was obtained, satisfying the IRT assumption that the items comprise a single unidimensional scale. The eight items in the scale could be placed on a hierarchy of 'difficulty' (H coefficient = 0.55), with 'Shopping' being the most 'difficult' item and 'Telephone use' being the least 'difficult' item. 'Shopping' was the most discriminatory item differentiating well between patients of different levels of ability.

Conclusions: IRT methods are capable of providing more information about functional impairment than a summed score.

S. McGrory et al.

'Shopping' and 'Telephone use' were identified as items that reveal key information about a patient's level of ability, and could be useful screening questions for clinicians.

Keywords *item response theory, Mokken, dementia, functional measures, Lawton IADL, older people*

Introduction

Functional impairment is a core feature of dementia. Functional impairment is widely measured using 'Activities of Daily Living' (ADL) scales which encompass the assessment of both Basic Activities of Daily Living and Instrumental Activities of Daily Living (IADL). The former include activities such as dressing, bathing and feeding, while the latter concern more complex activities such as handling finances, taking medication and doing housework. IADL items requiring more complex neuropsychological organisation are highly dependent on adequate cognitive capacity and are therefore most susceptible to the early effects of cognitive decline [1]. Assessing IADL can consequently be useful in detecting and diagnosing early dementia [2].

Many IADL scales have been developed and they are conventionally scored by summing the responses to individual items on the scale to yield a total score. Despite the popularity of this method, there are issues that make it difficult to interpret [3]. For example, the total-score method weights each item equally, which assumes that all items represent equal levels of severity. This is very rarely the case [4]: e.g. doing laundry is more challenging than feeding [5]. Furthermore, the total-score method asserts that each item on the scale is equally related to the construct under examination which again is rarely the case [4]. These limitations can be overcome using item response theory (IRT) methods [6, 7].

According to IRT, the items on a scale are related to a latent construct; functional impairment in the case of the current study. IRT is based on the probability of a person achieving a particular score on a test given their standing on the latent construct [8]. This better reflects the underlying trait than traditional methods [9]. A more in-depth discussion of IRT is beyond the scope of this article, for further information see Hambleton and Swaminathan [6].

IRT provides two useful measures; item 'difficulty' and 'discrimination'. 'Difficulty' refers to the ability level necessary in terms of the latent construct for an individual to have a 0.5 probability of responding positively to a specific item. In the context of IRT, an item is considered 'difficult' if a high degree of ability is required in order to respond positively. Only those with a high level of ability will be able to endorse the 'difficult' items, whereas most will endorse or respond positively to the less 'difficult' items. From a clinical point of view 'difficulty' can be thought of as severity, for example the degree of functional impairment required to cause challenges with handling finances. With regards to IADL items the more 'difficult' a task is, the better the person's functional ability must be in order to be able to perform the task. A hierarchy of item 'difficulty' details the

expected order of functional impairment. Hierarchies of ADL/IADL scales have been confirmed using IRT methods [5, 10, 11]. These hierarchies found that items differed in terms of 'difficulty' with items such as 'Shopping' and 'Doing Laundry' being more 'difficult' than 'Dressing' and 'Transferring' [5] and 'Active recreation' and 'Volunteer job' being more 'difficult' than 'Taking care of health' and 'Personal care needs' [10]. However, these studies did not examine functional impairment within dementia populations. Hierarchies outlining the functional impairment in dementia would offer prognostic value to researchers and clinicians investigating functional impairment by identifying any deviations in the rate of decline from the typical trajectory of loss [11].

'Discrimination' is the extent to which the item distinguishes participants with relatively low functional ability from those with relatively high levels of ability. An item with poor 'discrimination' will distinguish poorly between mild and severe levels of functional impairment because the probability that the person will endorse the item is nearly the same across all levels of severity. An item with good 'discrimination' distinguishes well between varying levels of functional ability because as the level of severity increases so too does the probability that a respondent will be unable to perform the task. For example, Fieo et al. [11] determined 'Prepare a meal' had very weak discriminatory value and did not differentiate between people of different abilities, whereas 'Get on a bus' was the most discriminatory item differentiating between those with low functional ability and those with high functional ability. Determining item 'discrimination' can identify key items on a scale and highlight weaker items or those whose function is redundant [12].

IRT methods have been applied to ADL/IADL scales in general populations [5, 10, 11]. However, the Lawton IADL scale has not been analysed with IRT methods to investigate the pattern of functional impairment caused by dementia. This analysis in a sample comprises people with dementia could provide clinically useful information. Therefore, this study applied IRT methods to the Lawton IADL scale to establish a hierarchy of item 'difficulties' and to assess the 'discriminatory' power of the items in people with dementia.

Methods**Sample**

Data were obtained from the Scottish Dementia Research Interest Register, described in detail previously [13]. Participants

S. McGrory *et al.*

least 'difficult' item with problems performing this task indicating severe impairment. A patient reporting challenges with 'Telephone use' is very unlikely to be able to perform any other task in the scale. Likewise, it is likely that a patient reporting no problems with 'Shopping' or 'Food preparation' will have no limitations with other tasks.

These findings have useful clinical implications. People requiring assistance with the most 'difficult' item 'Shopping' should alert clinicians as, in the context of cognitive decline, it could herald the initial phase of functional impairment. Problems performing complex activities of daily living have been reported to precede dementia diagnosis by as much as 10 years [26]. As the items of the Lawton IADL scale conform to a formal hierarchy the most 'difficult' items such as 'Shopping' and 'Food preparation' can act as sensitive indicators of impending disability in the other activities [27].

Items with high 'discrimination' are better able to detect differences in effects of interventions or drug therapies [28]. Ideally, a measure should comprise items of differing degrees of 'difficulty' right across the spectrum of ability and demonstrate high levels of 'discrimination'. This ensures that changes at every point along the ability spectrum will be detected resulting in more reliable and accurate measurement.

The inclusion of items such as 'Shopping' and 'Food preparation' which showed high 'discrimination' may assist in the detection of small changes in milder stages of dementia as these abilities are lost rapidly at an early stage. 'Telephone use' discriminates well at the lower end of the hierarchy. The creation of more items such as this may help to introduce greater 'discrimination' in the more advanced stages. IRT analyses can be applied to IADL/ADL scales making them more sensitive to identifying and monitoring changes in both mildly and severely impaired patients. Better assessment of the rate of decline could enhance prediction of future deterioration.

The study was predominantly restricted to patients with mild-moderate dementia (mean MMSE score 22.1, SD = 5.05) with a range of aetiologies. Future research should investigate the loss-of-functional independence in more severe samples, and in specific dementia subtypes. For example, there is more rapid deterioration of functional abilities in patients with Frontotemporal dementia compared with Alzheimer disease [29]. The majority of this sample (80%) was taking Cholinesterase Inhibitors which are acknowledged to be effective in delaying or slowing the worsening of symptoms, although these effects are not large [30].

IRT has benefits not only in the monitoring of patients, but establishing the sequence of decline which can also help in characterising adaptations to disability and differences between subgroups. While additive summary scores can be helpful in summarising overall function, they can conceal as much information as they reveal, and IRT methods are a useful method to increase the information provided by simple functional scales.

Furthermore, simultaneous analyses of cognitive and functional scales could enable the discovery of more precise associations between cognitive and functional outcomes.

Key points

- IRT analyses provide more information than the summed score method.
 - The ability to shop and to use the telephone was identified as key items which could be valuable screening questions for clinic.
 - IRT analyses can make IADL scales more sensitive to identifying and observing changes in functional abilities in dementia.
-

Conflicts of interest

None declared.

Funding

This work was supported by a PhD studentship from Alzheimer Scotland.

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *Age and Ageing* online.

References

1. Njegovan V, Man-Son-Hing M, Mitchell SL, Molnar FJ. The hierarchy of functional loss associated with cognitive decline in older persons. *J Gerontol A Biol Sci Med Sci* 2001; 56: M638-43.
2. Desai AK, Grossberg GT, Sheth DN. Activities of daily living in patients with dementia: clinical relevance, methods of assessment and effects of treatment. *CNS Drugs* 2004; 18: 853-75.
3. Reise S, Henson J. A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *J Pers Assess* 2003; 81: 93-103.
4. Gibbons RD, Clark DC, Cavanaugh SV, Davis JM. Application of modern psychometric theory in psychiatric research. *J Psychiatr Res* 1985; 19: 43-55.
5. Spector WD, Fleishman JA. Combining activities of daily living with instrumental activities of daily living to measure functional disability. *J Gerontol B Psychol Sci Soc Sci* 1998; 53 (Suppl. 1): 46-57.
6. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston: Kluwer, 1985.
7. Fico RA, Austin EJ, Starr JM, Deary IJ. Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC Geriatr* 2011; 11: 42-57.
8. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. *J Pers Assess* 2005; 84: 228-38.
9. Chan KS, Kasper JD, Brandt J, Pezzin LE. Measurement equivalence in ADL and IADL difficulty across international

