Measuring Mindfulness and Health-Related Outcomes: Applying Rasch Analysis and Generalisability Theory to Improve Reliability and Validity of Scales

Oleg Medvedev

PhD

2017

Measuring Mindfulness and Health-Related Outcomes: Applying Rasch Analysis and Generalisability Theory to Improve Reliability and Validity of Scales

Oleg Medvedev

A thesis submitted to Auckland University of Technology
in fulfillment of the requirements for the degree of
Doctor of Philosophy (PhD)

2017

School of Public Health and Psychosocial Studies

Faculty of Health and Environmental Sciences

Abstract

There is growing evidence for mindfulness-based interventions in alleviating the symptoms and enhancing the coping abilities of people suffering from psychological and physical health conditions and improving overall well-being. In essence, mindfulness is our immediate, instant contact with our internal and external environments that is not contaminated by judgmental attitudes or habitual cognitions, and is associated subjectively with a greater clarity of consciousness. With increased application of mindfulness-based interventions, evaluation of their effectiveness requires more accurate measurement of both mindfulness and associated health-related outcomes. In particular, issues with measurement precision (e.g. ordinal rather than interval scaling), item functioning and the state-trait distinction have not been sufficiently addressed or resolved using appropriate modern statistical methods. Ordinal measures have limited precision, and using them with parametric statistical techniques violates the basic assumptions of these tests. The accurate distinction of state from trait and establishing measurement at an interval level are two essential steps for rigorously validating mindfulness and health outcome measures.

The initial part of this thesis focused on applying Rasch analysis to improve the scaling properties of ordinal mindfulness and outcome measures to interval-level scales suitable for parametric statistics. Four studies improved the psychometric properties of widely used and recently developed mindfulness measures including the Mindful Attention and Awareness Scale (MAAS), the Kentucky Inventory of Mindfulness Skills (KIMS), the Five Facet Mindfulness Questionnaire (FFMQ) and the Comprehensive Inventory of Mindfulness Experiences (CHIME). Three further studies improved the scaling properties of the Functional Assessment Measure UK FIM+FAM, the Oxford Happiness Questionnaire (OHQ), and the Perceived Stress Scale (PSS). These studies all employed Rasch analysis and developed conversion algorithms to transform ordinal responses into interval-level data. The second part of the thesis applied Generalisability Theory for the first time to distinguish quantitatively between state and trait components in a mindfulness measure. This study demonstrated that Generalisability Theory can be successfully applied to accurately distinguish between state and trait components in a psychometric measure, and it is recommended as the most applicable psychometric method to validate state and trait questionnaires in the future. Until now the distinction between state and trait has typically based upon a single correlation between total test scores on two different occasions. Consequently, poor items could ‘hide’ behind the other items undetected by test-retest correlation and may affect the overall performance of a scale. The proposed method estimates the extent to which a scale and every individual item are each measuring a state and a trait. Findings of this study have far-reaching implications to help improve the accuracy of distinction between state and trait in measurement of mindfulness and other areas of psychological assessment. Together, these studies analysed data representing 2,551 participants including community and clinical populations, as well as university students. Overall, this work contributed practical solutions and innovative methods to improve the reliability, validity and scaling properties of psychometric measures with a range of implications for mindfulness and health research practice.

Table of Contents

List of Figures

List of Tables

List of Commonly Used Abbreviations

ACT	**Acceptance and Commitment Therapy**
CFA	**Confirmatory Factor Analysis**
DBT	**Dialectical Behavioural Therapy**
DIF	**Differential Item Functioning**
EFA	**Exploratory Factor Analysis**
FFMQ	**Five Facet Mindfulness Questionnaire**
GT	**Generalisability Theory**
IRT	**Item Response Theory**
ICC	**Item Characteristic Curve**
KIMS	**Kentucky Inventory of Mindfulness Skills**
MAAS	**Mindful Attention and Awareness Scale**
MBCT	**Mindfulness-Based Cognitive Therapy**
MBSR	**Mindfulness-Based Stress Reduction**
MiCBT	**Mindfulness-integrated Cognitive Behaviour Therapy**
OHQ	**Oxford Happiness Questionnaire**
PCA	**Principle Component Analysis**
PSS	**Perceived Stress Scale**
TMS	**Toronto Mindfulness Scale**
UK FIMFAM	**UK Functional Assessment Measure**

Attestation of Authorship

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the co-authored works, acknowledgements or referenced in text), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning."

Signature: Oleg Medvedev Oleg Medvedev Date:

Co-Authored Works

Although, this thesis includes studies that were mostly my own ideas and work, they were only achievable with the advice and support of my supervisors and several collaborators.

All studies included within this thesis were conducted under the supervision of Professor Richard Siegert and Associate Professor Chris Krägeloh, and the study in Chapter Eleven included supervision from Professor Ajit Narayanan. All three supervisors provided advice for data analysis and interpretation of the results reported in this thesis.

For the studies described in Chapters Three to Seven and in Chapter Ten, my contribution covered all the main aspects of the projects such as developing a research proposal, study materials, obtaining ethics permissions, completing data collection and analysis, and writing and submitting the manuscripts to international psychology journals. Completion of this work was greatly enhanced through continuous support and advice of my supervisors to ensure achievement of the highest possible academic standard.

The study described in Chapter Three also involved other collaborators. PhD student Xuan Joanna Feng contributed general population data to the sample to make the findings more generalizable and helped with data analysis and reviewing the manuscript. Joanna's collaborator Jing Young Jang supported collection of general population data and her supervisor, Adjunct Professor Rex Billington, reviewed the final version of the manuscript.

Chapter Six describes a study conducted in collaboration with the University of Bern, Switzerland, represented by Dr Claudia Bergomi, and the Swiss Federal Institute of Sport represented by Dr Philipp Röthlin. The author met Drs Bergomi and Röthlin at the International Mindfulness Conference in Rome 2016 when presenting his earlier work that used Rasch analysis to enhance psychometric properties of the MAAS, the KIMS & the FFMQ. After considering the benefits of Rasch analysis, Dr Bergomi kindly provided her dataset (*n*=443) for Rasch analysis of the CHIME, a new multidimensional mindfulness measure developed by herself and colleagues (Bergomi, Tschacher, & Kupper, 2014) and both Claudia and Phillipp reviewed the manuscript at different stages of development.

Chapter Seven includes a study that involved both local (University of Auckland, NZ) and international (Westchester University, USA) collaboration. In particular, Dr Erin Hill (West Chester University) provided two datasets: one was collected in New Zealand for her PhD (*n*=300) and another collected in the US (*n*=300), both works were focused on investigating various stress effects on health. Dr Marcus Henning (University of Auckland) provided another dataset (*n*=300) collected in New Zealand for the study investigating effects of stress on motivational variables. My supervisor Chris Krägeloh helped with preparing randomised datasets and data analysis. My supervisor Richard Siegert together with Erin Hill, Rex Billington, Craig Webster, Roger Booth and Marcus Henning reviewed the final version of the manuscript.

The study described in Chapter Eight involved international collaboration with Dr Ahmed Mohamed (University of Nottingham, Malaysia), who provided additional data (*n*=101) collected in Malaysia to increase the generalisability of the findings. Dr Daniel Shepherd and Dr Erik Landhuis both helped with data collection in New Zealand and reviewed the final version of the manuscript.

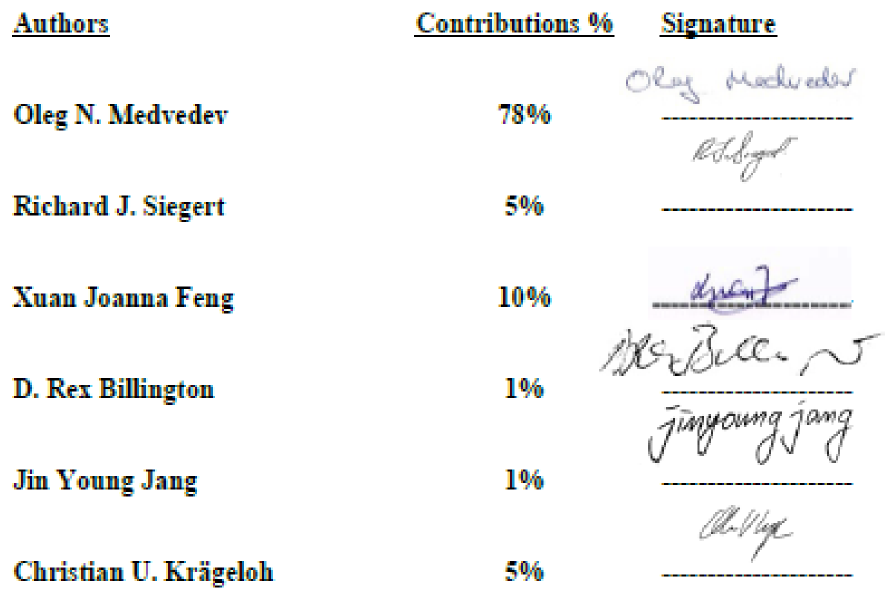
The study described in Chapter Nine involved international collaboration with Professor Lynne Turner-Stokes, Dr Roxana Vanderstay and Dr Stephen Ashford (King's College London, UK) on applying Rasch analysis to the Functional Assessment Measure (UKFIM+FAM) using a large stroke dataset from the UK rehabilitation outcomes consortium (UK ROC). In this project my contribution included applying modern strategies of Rasch analysis to investigate and improve psychometric properties of the UKFIM+FAM measure using dataset of *n* =320 stroke patients provided by Professor Turner-Stokes. Roxana and Stephan conducted preliminary statistical analysis on this dataset, which provided a base for my more extensive Rasch analysis. My contribution also included writing, preparing and submitting the resulting manuscript to the journal PLOS One.

Overall, my contribution to the studies included in this thesis was at least 80%, and I was the principal (first) author, writing the first complete draft, of all included publications.

Declarations confirming exact percentages of individual contributions for each work signed by co-authors and a qualitative supervisor statement confirming my contributions are included on the following pages.

Chapter Three

Medvedev, O. N., Siegert, R. J., Feng, X. J., Billington, D. R., Jang, J. Y., & Krägeloh, C. U. (2016). Measuring trait mindfulness: how to improve the precision of the Mindful Attention Awareness Scale using a Rasch model. *Mindfulness*, 7(2), 384-395.



Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	78%	Oleg Medvedev	22.12.2016
Richard J. Siegert	5%	R J Siegert	22.12.2016
Xuan Joanna Feng	10%		21.12.2016
D. Rex Billington	1%		31.01.2017
Jin Young Jang	1%	Jinyoung Jang	01.02.2017
Christian U. Krägeloh	5%		22.12.2016

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for all aspects of this study (Chapter Three). He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor. In this study Oleg used general population data provided by PhD student Xuan Joanna Feng to make the findings more generalizable and he has studiously acknowledged the use of these data set where relevant.

Professor Richard Siegert

R J Siegert

Chapter Four

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (2016b). Rasch Analysis of the Kentucky Inventory of Mindfulness Skills. *Mindfulness,* 7(2), 466-478.

Authors	**Contributions %**	**Signature**	**Date:**
Oleg N. Medvedev	**85%**	Oleg Medvedev	21.12.2016
Richard J. Siegert	**5%**	R.J. Siegert	21.12.2016
Paula Kersten	**5%**	PKersten	4/1/17
Christian U. Krägeloh	**5%**		21.12.2016

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for all aspects of this study (Chapter Four). He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor.

Professor Richard Siegert

R.J. Siegert

Chapter Five

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (20117). Improving the Precision of the Five Facet Mindfulness Questionnaire Using a Rasch Approach. *Mindfulness*, doi 10.1007/s12671-016-0676-8

Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	**85%**	Oleg Medvedev	21.12.2016
Richard J. Siegert	**5%**	R.J. Siegert	21.12.2016
Paula Kersten	**3%**		4/1/17
Christian U. Krägeloh	**7%**		21.12.2016

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for all aspects of this study (Chapter Five). He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor.

Professor Richard Siegert

R.J. Siegert

Chapter Six

Medvedev, O. N., Bergomi, C., Röthlin, P., & Krägeloh, C. U. (under review) Assessing the psychometric properties of the Comprehensive Inventory of Mindfulness Experiences (CHIME) Using Rasch Analysis (Submitted to *European Journal of Psychological Assessment*).

Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	**80%**	Oleg Medvedev	24.01.17
Claudia Bergomi	**10%**	Claudia Bergomi	15.02.17
Philipp Röthlin	**5%**		26.01.17
Christian U. Krägeloh	**5%**		8.02.17

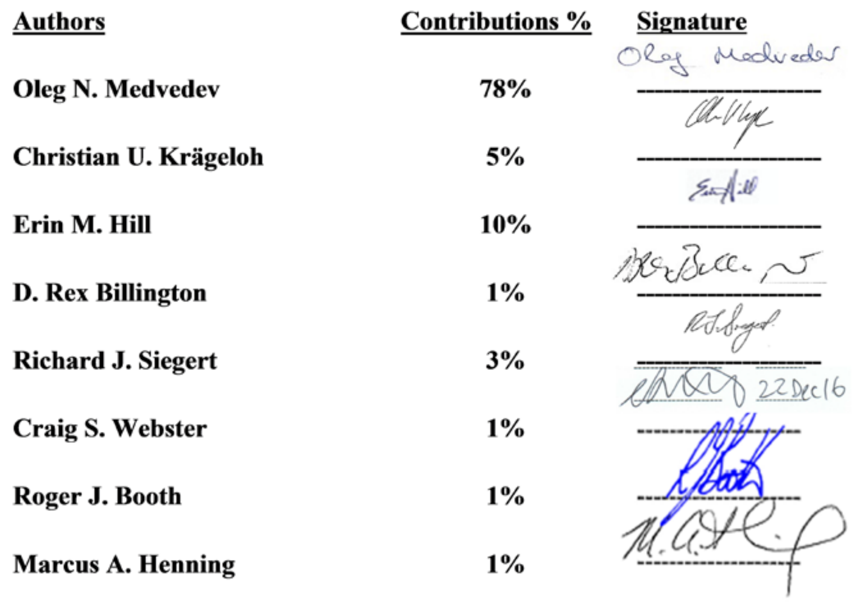
Qualitative Statement of Supervisor:

Oleg has been primarily responsible for the most aspects of this study (Chapter Six). He has been responsible for reviewing literature, developing hypotheses, designing the studies, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor. In this study Oleg used data from overseas kindly provided by Dr Bergomi (*n*=443) with whom he collaborated and he has studiously acknowledged the use of these data set where relevant.

Associaste Professor Chris Krägeloh

Chapter Seven

Medvedev, O. N., Krägeloh, C. U., Hill, E. M., Billington, R., Siegert, R. J., Webster, C. S., Booth, R. J., & Henning, M. A. (2017). Rasch analysis of the Perceived Stress Scale: Transformation from an ordinal to a linear measure. *Journal of Health Psychology*, doi:10.1177/1359105316689603



Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	78%	Oleg Medvedev	21.12.2016
Christian U. Krägeloh	5%		22.12.2016
Erin M. Hill	10%	Erin Hill	22.12.2016
D. Rex Billington	1%		31.01.2017
Richard J. Siegert	3%	R.J. Siegert	22.12.2016
Craig S. Webster	1%	22Dec16	22.12.2016
Roger J. Booth	1%		22.12.2016
Marcus A. Henning	1%		22.12.2016

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for the most aspects of this study (Chapter Seven). He has been responsible for reviewing literature, developing hypotheses, designing the studies, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor. In this study Oleg used data provided by Dr Erin Hill collected in New Zealand (*n*=300) and in the US (*n*=300), and by Dr Marcus Henning (*n*=300) collected in New Zealand with whom he collaborated and he has studiously acknowledged the use of these data sets where relevant.

Professor Richard Siegert

R.J. Siegert

Chapter Eight

Medvedev, O. N., Siegert, R. J., Mohamed, A. D., Shepherd, D., Landhuis, E., & Krägeloh, C. U. (2016) The Oxford Happiness Questionnaire: Transformation from an ordinal to an interval measure using Rasch analysis. *Journal of Happiness Studies,* doi:10.1007/s10902-016-9784-3

Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	**80%**	Oleg Medvedev	**09/01/2017**
Richard J. Siegert	**3%**	R.J.Siegert	**09/01/2017**
Ahmed D. Mohamed	**10%**	Ahmed	**09/01/2017**
Daniel Shepherd	**3%**	D Shepherd	**14/01/2017**
Erik Landhuis	**1%**		**16-01-2017**
Christian U. Krägeloh	**3%**		**09/01/2017**

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for all aspects of this study (Chapter Eight). He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor. In this study Oleg used data provided by Dr Ahmed Mohamed (University of Nottingham, Malaysia), (*n*=101) collected in Malaysia to increase the generalisability of the findings and he has acknowledged the use of these data set where relevant.

Professor Richard Siegert

R.J.Siegert

Chapter Nine

Medvedev, O. N., Vanderstay, R., Turner-Stokes, L., Stephen, A., & Siegert, R. J. (under review) Rasch analysis of the UK Functional Assessment Measure in patients with complex disability after stroke (submitted to *PLOS One*).

Authors	**Contributions %**	**Signature**	**Date:**
Oleg N. Medvedev	**80%**	Oleg Medvedev	**10.02.17**
Lynne Turner-Stokes	**8%**		**26.1.17**
Roxana Vanderstay	**5%**	R Vanderstay	**26.01.17**
Stephen Ashford	**2%**	S. Ashford	**09.02.17**
Richard J. Siegert	**5%**	R.J. Siegert	**09.02.17**

Qualitative Statement of Supervisor:

Oleg has been primarily responsible for the most aspects of this study (Chapter Nine). He has been responsible for reviewing literature, developing hypotheses, designing the studies, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor. In this study Oleg used dataset of *n* =320 stroke patients provided by Professor Turner-Stokes with whom he collaborated and he has studiously acknowledged the use of these data set where relevant.

Professor Richard Siegert

R.J. Siegert

Chapter Ten

Medvedev, O. N., Krägeloh, C. U., Narayanan, A., & Siegert, R. J. (2017). Measuring mindfulness: Applying Generalizability Theory to distinguish between state and trait. *Mindfulness,* doi: 10.1007/s12671-017-0679-0

Authors	Contributions %	Signature	Date:
Oleg N. Medvedev	**80%**	Oleg Medvedev	**7.01.17**
Christian U. Krägeloh	**6%**		**11.01.17**
Ajit Narayanan	**4%**		**17.01.17**
Richard J. Siegert	**10%**	R.J. Siegert	**11.01.17**

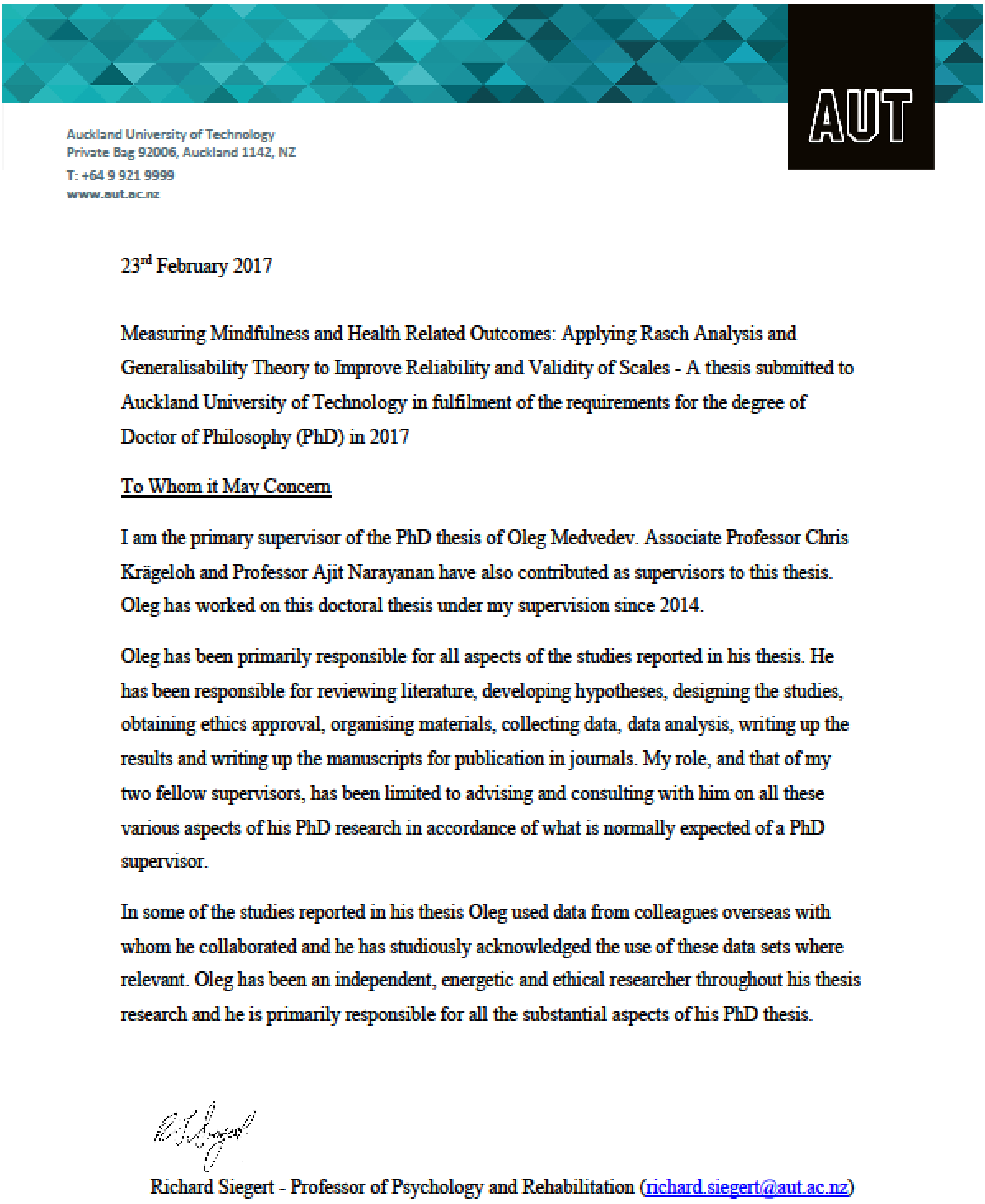
Qualitative Statement of Supervisor:

Oleg has been primarily responsible for all aspects of this study (Chapter Ten). He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in a journal. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor.

Professor Richard Siegert

R.J. Siegert

General Qualitative Statement of Supervisor



AUT

Auckland University of Technology
Private Bag 92006, Auckland 1142, NZ
T: +64 9 921 9999
www.aut.ac.nz

23rd February 2017

Measuring Mindfulness and Health Related Outcomes: Applying Rasch Analysis and Generalisability Theory to Improve Reliability and Validity of Scales - A thesis submitted to Auckland University of Technology in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) in 2017

To Whom it May Concern

I am the primary supervisor of the PhD thesis of Oleg Medvedev. Associate Professor Chris Krägeloh and Professor Ajit Narayanan have also contributed as supervisors to this thesis. Oleg has worked on this doctoral thesis under my supervision since 2014.

Oleg has been primarily responsible for all aspects of the studies reported in his thesis. He has been responsible for reviewing literature, developing hypotheses, designing the studies, obtaining ethics approval, organising materials, collecting data, data analysis, writing up the results and writing up the manuscripts for publication in journals. My role, and that of my two fellow supervisors, has been limited to advising and consulting with him on all these various aspects of his PhD research in accordance of what is normally expected of a PhD supervisor.

In some of the studies reported in his thesis Oleg used data from colleagues overseas with whom he collaborated and he has studiously acknowledged the use of these data sets where relevant. Oleg has been an independent, energetic and ethical researcher throughout his thesis research and he is primarily responsible for all the substantial aspects of his PhD thesis.

Richard Siegert - Professor of Psychology and Rehabilitation (richard.siegert@aut.ac.nz)

Acknowledgements

I would like to express my sincere gratitude to my supervisors Professor Richard Siegert, Associate Professor Chris Krägeloh, and Professor Ajit Narayanan for their great support and inspiration to do this research. In particular, I would like to thank Richard Siegert for his wise guidance to the Rasch and G (Generalisability Theory) worlds, Chris Krägeloh for travelling with me through these pathways of human intelligence and Ajit Narayanan for ensuring that we are on the right G-track.

I like to express my genuine thanks and admiration to my wife Svetlana who both motivated and supported this challenging project as well as showed immeasurable patience while listening to countless stories of statistical adventures. I also appreciate her help with proofreading of this thesis and included works. I thank my son Atisha for additional motivation to fast completion of this work and his tolerance for spending my time working on this thesis. My special thanks to my mother Liudmila, who always supported my learning and education and was very patient bearing consequences of my experiments conducted at home.

In addition, I am very grateful to all volunteers who took their time to participate in this research. Without their participation this work would not be possible.

Finally, I sincerely thank all my teachers for sharing their wisdom, knowledge and experience, which were the true foundations for completion of this work.

May this work contribute to health and happiness of all! Thank you.

Funding

I would like to acknowledge funding of my doctoral study by the Vice-Chancellor’s Scholarship of the Auckland University of Technology, which allowed me to conduct this research work on a full-time basis.

I would also like to acknowledge funding of the Rasch analysis software (RUMM2030) for my personal laptop and covering of costs for my attendance and presenting a paper from this thesis on the International Mindfulness Conference in Rome in 2016 by the AUT Strategic Research Investment Fund granted for mindfulness research, which was supported by my Supervisor Richard Siegert.

Intellectual Property Rights

There are no intellectual property rights related to this thesis.

Ethical Approval

For the studies described in Chapters Three, Four, Five and Nine, ethical approval was received from the Auckland University of Technology on the 24th of February 2014 and additional amendments for the study described in chapter ten were approved on the 18th of January 2016 (Ref:14/10 Measuring mindfulness: Determining the psychometric and neurophysiological correlates of mindfulness). Letters confirming the AUT ethics approval are included in Appendix A(1-6).

The study described in Chapter Three also included general population data provided by another PhD student Xuan Joanna Feng, which were collected to study aspects of mindfulness unrelated to this study, which required accuracy of measurement and thus benefited from the finding of this study. The letter confirming the AUT ethics approval is attached in Appendix A3.

The study described in Chapter Seven analysed three previously collected datasets. Two of them were contributed by Dr Erin Hill (the West Chester University): one was collected in New Zealand with ethical approval of the Auckland University of Technology for her PhD (n=300) and another collected in the US (n=300), with ethical approval of the West Chester University. Dr Marcus Henning (the University of Auckland) provided another dataset (n=300) collected in New Zealand for the study investigating effects of stress on motivational variables, which was approved by the University of Auckland Ethics Committee. Ethics approval letters for these studies are attached in Appendix A(6-8).

The study described in Chapter Six analysed general population data collected in Europe (n=443) contributed by Dr Claudia Bergomi from the University of Bern, Switzerland. Participants gave informed consent in compliance with Swiss ethics legislation, which allows scientific use of anonymized data. If the data are related to health and disease, or related to the function and structure of the human body, a study additionally needs the specific approval by the Cantonal Ethics Committee. This was not the case in the present study.

The study described in Chapter Eight analysed the data collected by the author earlier for the study on happiness, subjective well-being, quality of life and life satisfaction with life and the Auckland University of Technology ethical approval for this study is attached in Appendix A4 (Ref:11/209 Happiness, subjective well-being, quality of life and life satisfaction with life). This study also involved international collaboration with the

University of Nottingham, Malaysia, that provided ethical approval for the data collected by Dr Ahmed Mohamed in Malaysia, part of which was analysed in this study (Appendix A5).

The study described in Chapter Nine involved international collaboration with King's College London, UK, and included Professor Lynne Turner-Stokes, Dr Roxana Vanderstay and Dr Stephen Ashford as collaborators. This study used a large stroke data set from the UK rehabilitation outcomes consortium (UK ROC). The UKROC programme is registered as a multicentre service evaluation and as a Payment by Results Improvement Project. Collection and reporting of the UKROC dataset is a commissioning requirement according to the NHSE service specification for Level 1 and 2 Rehabilitation Services. According to the UK Health Research Authority, the publication of research findings from de-identified data gathered in the course of routine clinical practice does not require research ethics permission. Registration: The programme is registered with the NIHR Comprehensive Local Research Network: ID number 6352.

Introduction

Mindfulness practice is a safe, non-invasive method for the management of stress, emotional problems and for the improvement of psychological well-being (Baer, 2003; Brown & Ryan, 2003). Mindfulness refers to paying attention to and being aware of internal and external experiences of the present moment associated with a non-judgmental attitude (Segal, Williams, & Teasdale, 2013). There is a rapidly growing evidence base for the therapeutic application of mindfulness techniques for alleviating symptoms and enhancing the coping abilities of people suffering from anxiety, stress, depression, emotional instability, substance abuse, post-traumatic stress disorder, borderline personality disorder, psychophysiological disorders, and suicidal/self-harm behaviour (Chiesa & Serretti, 2010; Hofmann, Sawyer, Witt, & Oh, 2010; Zoogman, Goldberg, Hoyt, & Miller, 2015). In addition, mindfulness-based interventions (MBIs) were reported to enhance psychological well-being (Bennet & Dorje, 2015; Josefsson, Lindwall, & Broberg, 2014) and regulation of emotions (Chambers, Gullone, & Allen, 2009; Lyvers, Makin, Tomas, Thorberg, & Samios, 2014), and hence its baseline levels should be controlled when evaluating outcomes of MBIs (Visted, Vøllestad, Nielsen, & Nielsen, 2015). Consequently, there is a need for accurate measurement of mindfulness and related outcomes that accurately reflect psychological changes in people participating in MBIs.

Currently, measurement of both mindfulness and related outcomes is associated with issues such as scales' precision (e.g. ordinal rather than interval scaling), structural validity, item functioning and the state-trait distinction, which have not been sufficiently addressed using appropriate statistical methods (Park, Reilly-Spong, & Gross, 2013; Van Dam, Earleywine, & Borders, 2010). Establishing measurement at an interval level and the accurate distinction of state from trait are two essential steps for validating mindfulness and health outcome measures rigorously. Ordinal measures have limited precision, and using them with parametric statistical techniques violates the basic assumptions of these tests. In ordinal scales, responses are rank-ordered but distances between response options have no real meaning. Therefore, ordinal scales do not support mathematical operations of adding, subtracting, dividing and multiplying (Merbitz, Morris, & Grip, 1989). In contrast, interval level of measurement is characterised by the same distances between categories and refers to units of measurement. Commonly used example of an interval scale is temperature where the difference between 1 and 2 degrees Celsius is the same as between 2 and 3 and between 3 and 4. It should be noted that item summary scores may not be an accurate estimate of the latent trait as different items may explain a different amount of information relevant to the latent trait (e.g. mindfulness) (Stucki, Daltroy, Katz, Johannesson, & Liang, 1996; Allen & Yen, 1979). Usage of ordinal scales in research may compromise the validity of comparisons with interval level neurophysiological data (e.g. electroencephalogram, skin conductance, and heart rate), an especially important consideration in modern mindfulness research. Additionally, any specific element of mindfulness treatment can only be evaluated by comparing state and trait changes using techniques that allow such changes to be measured. If state and trait mindfulness cannot be reliably measured and distinguished in neurophysiological studies, then validity of their comparisons with neurophysiological data is confounded.

The initial part of this thesis aimed to improve the psychometric properties of mindfulness and related outcome measures up to an interval level scale suitable for parametric statistics and to address structural validity issues. Such investigation can be conducted using Rasch analysis, which employs a probabilistic logistic model and is particularly suited for this purpose (Tennant & Conaghan 2007; Rasch 1961). Rasch analysis has shown numerous advantages over other more traditional statistical methods, which has been extensively argued elsewhere (Rasch 1960; Wilson 2005; Wright and Stone 1979). The end product of Rasch analysis is algorithms to transform scores from an ordinal to an interval scale that increase precision of measurement, which has been demonstrated empirically (Norquist et al. 2004). Three studies described in Chapters Three to Six applied Rasch analysis to enhance the psychometric properties of the three widely used mindfulness measures and a newly developed multidimensional measure: the Mindful Attention and Awareness Scale (MAAS) (Brown & Rayan, 2003), the Kentucky Inventory of Mindfulness Skills (KIMS) (Baer, Smith, & Allen, 2004), the Five Facet Mindfulness Questionnaire (FFMQ) (Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006), and the Comprehensive Inventory of Mindfulness Experiences (CHIME) (Bergomi, Tschacher, & Kupper, 2014). The three studies that analysed the MAAS, the KIMS and the FFMQ have already been published (Medvedev et al., 2016a, Medvedev, Siegert, Kersten & Krägeloh, 2016b,c), and the CHIME study is currently under review.

Reliable improvement of measurement accuracy requires that all measures used in research are at least working at an interval level (Allen & Yen, 1979; Stucki, Daltroy, Katz, Johannesson, & Liang, 1996) meaning that psychometric properties of outcome measures used in mindfulness research also require comparable enhancement. Mindfulness was primarily applied for stress reduction (Kabat-Zinn, 1982, 1990) and improvement of psychological well-being (Bennet & Dorje, 2015; Josefsson et al., 2014), with further promising applications in rehabilitation medicine (Siegert, Rowland, & Theadom, 2016). Therefore, three other studies described in Chapters Seven to Nine used Rasch analysis to improve the precision of the most popular stress measure, the Perceived Stress Scale (PSS) (Cohen & Williamson, 1988), a widely used measure of psychological well-being - the Oxford Happiness Questionnaire (Hills & Argyle, 2002), and the Functional Assessment Measure UK FIM+FAM, which is widely used in rehabilitation medicine (Turner-Stokes, Nyein, Turner-Stokes & Gatehouse, 1999). The two studies, which investigated psychometric properties of the OHQ and the PSS, have already been published (Medvedev et al., 2016c, 2017b), and the UK FIM+FAM study is currently under review. These studies all used Rasch analysis to produce necessary psychometric modifications of the instruments and published ordinal-to-interval conversion algorithms that increase the precision of each measure without the need to modify their original response formats.

Chapter Ten of this thesis developed a novel technique to differentiate between state and trait variance components in a measure based on Generalizability Theory (GT) (Cronbach, Rajaratnam & Gleser, 1963). GT is an analytical technique that assesses numerous sources of variance associated with the main variable of interest (e.g. a mindfulness score) (Allal & Cardinet, 1976), which is a suitable method to differentiate between state and trait variance components in a measure. Its application is illustrated here with an empirical example using the Toronto Mindfulness Scale (TMS) (Lau et al., 2006). This study, described in Chapter Ten, has demonstrated that GT can be usefully applied to distinguish between state and trait components in a measure, and it is recommended as the most applicable psychometric method to validate state and trait measurement tools (Medvedev, Krägeloh, Narayanan, & Siegert, 2017c). Together, these findings have far-reaching implications to improve both reliability and validity of the investigated psychometric instruments and the accuracy of distinction between state and trait in mindfulness measurement and other areas of psychological assessment.

Chapter One. Mindfulness in Psychology

Definitions of Mindfulness

Recent years have seen a surge of interest in 'mindfulness', on the one hand as a component in clinical interventions for a wide range of psychological and health conditions (Chiesa & Serretti, 2010; Goldin & Gross, 2010; Zoogman, Goldberg, Hoyt, & Miller, 2014) and, on the other, as a trait or general capacity that is linked to optimal psychological wellbeing (Keng, Smoski, & Robin, 2011). Mindfulness has been described as "paying attention in a particular way: on purpose, in the present moment, and non-judgmentally" (Kabat-Zinn, 1994, p.4). As a broader disposition, mindfulness is the ability to be aware of external and internal experiences as phenomena without automatically using existing cognitive schemas that help construct our conceptual world and its objects (Olendzki, 2005).

Mindfulness is a translation of the word *sati* from the Pali language into English (Davids & Stede, 1921/2001). It also refers to remembering, awareness and attention and generally indicates presence of mind (Nyanaponika, 1973; Siegel, Germer, & Olendzki, 2009). Influential mindfulness definitions commonly used in psychology are presented in Table 1. One of the most cited definitions of mindfulness in the Western literature was proposed by Kabat-Zinn (1994). Recent definitions of mindfulness are used to cover various approaches and clinical interventions applied in psychology (Bishop et al., 2006; Hayes, Strosahl, & Wilson, 1999; Mace, 2008; Siegel et al., 2009). Therefore, additional elements and components such as *Describing*, *Acting With Awareness* and *Non-reacting* to inner experience were also added to the originally proposed definitions (Baer et al., 2004, 2006). For instance, 'paying attention in a particular way' (Kabat-Zinn, 1994) may include friendliness, acceptance, kindness, curiosity and allowing (Segal, Williams, & Teasdale, 2013). Alternatively, mindfulness can be operationalised as a state-like quality as defined by Lau et al. (2006) (Table 1).

Mindfulness involves awareness and attention, which can be considered as the basic components of consciousness. Awareness refers to the conscious recognition of external sensory objects and internal thoughts and sensations, and attention is explained as turning toward the stimulus that is sufficiently intensive to engage it (Brown et al., 2007). Attention is usually defined as focusing on specific stimuli while ignoring others, but, in the context of mindfulness attention is paid to the present moment experience while ignoring unrelated cognitions associated with future or past (Bishop et al., 2004).

Table 1. *Mindfulness definitions used in psychology listed by year of publication.*

Reference Source	Mindfulness Definition
Kabat-Zinn (1994)	"…paying attention in a particular way, on purpose, in the present moment, and non-judgmentally" (p.4).
Buchheld et al. (2001)	"…the dispassionate, non-manipulative participant-observation of ongoing mental states, without lapsing into conceptualizations about momentary mental content or becoming lost in emotional reactions…carried out with curiosity and without bias or expectation" (p.11).
Baer (2003)	"…the non-judgmental observation of the on-going stream of internal and external stimuli as they arise" (p. 125).
Brown & Ryan (2003)	"…attention to and awareness of whatever is occurring in the present" (p. 824).
Bishop et al. (2004)	"…self-regulation of attention so that it is maintained on immediate experience, thereby allowing for increased recognition of mental events in the present moment" (p. 232).
Lau et al. (2006)	"(a) the intentional self-regulation of attention to facilitate greater awareness of bodily sensations, thoughts, and emotions; and (b) a specific quality of attention characterized by endeavouring to connect with each object in one's awareness (e.g., each bodily sensation, thought, or emotion) with curiosity, acceptance, and openness to experience. Such a state involves an active process of relating openly with one's current experience by allowing current thoughts, feelings, and sensations." (p.1447).
Segal et al. (2013)	"…the awareness that emerges through paying attention on purpose in the present moment and non-judgmentally to things as they are" (p. 132).

It can be illustrated in the famous experiment on focused attention (Simons & Christopher, 1999) where participants were instructed to count ball passes in a basketball game and failed to notice a person wearing a gorilla suit cross the court. However, if the same scene is observed without first establishing a particular attentional focus, one will pay equal attention to all events in perceptual field and certainly will notice a gorilla-dressed person appearing on the game court. Ordinary perception is characterised by brief attention to sensory objects followed by emotional, cognitive and behavioural responses. These responses are often conditioned by previous experiences, involve self-related evaluation of perceived objects (e.g. 'like' or 'dislike') and can easily assimilate an experience into present mental schemas. Ordinary perception is often associated with automatic labelling, conceptualising and judging of sensory experiences (Bargh &

Chartrand, 1999) and is influenced by an individual's beliefs, ideas and mental schemas (Leary, 2004; Leary, Adams, & Tate, 2006).

Notwithstanding some adaptive benefits of everyday or ordinary perception, it can also impose filters of environmental conditioning and self-centred concepts on most sensory events resulting in distorted perceptions of environment (Brown et al., 2007). In contrast, mindfulness involves a receptive attitude associated with attention focused on pure registration of perceptual experiences in the present moment (Kabat-Zinn, 1994). Mindfulness involves the ability to be aware of external and internal experiences as *phenomena* without automatically using the existing cognitive schemas that help construct our conceptual world and its objects (Olendzki, 2005). It is our immediate, instant contact with reality that is not contaminated by habitual and conceptual cognitions, and which allows a greater clarity of consciousness and more objectively based responses (Brown et al., 2007). This theoretical view is consistent with the majority of contemporary approaches to mindfulness, though there are variations in training methods and practices (Germer et al., 2005; Kabat-Zinn, 2003; Langer, 1989; Marlatt & Kristeller, 1999; Olendzki, 2005). Mindfulness practices consistent with this theoretical view require a practitioner to take the position of an alert spectator of all external and internal phenomena in their perceptual field, while maintaining a non-judgmental attitude free from attachment, grasping and aversion (Kabat-Zinn, 2003; Olendzki, 2005).

Generally, meditation practices can be divided into two categories: mindfulness and concentration. Concentration-based meditation involves an individual's attention to be narrowed by voluntarily focusing on a single stimulus such as sounds, sensations of breathing, and visual or sensory stimuli (Baer, 2003; Ivanovski & Malhi, 2007). For instance, Shamata meditation is practised across different Buddhist traditions and can be performed by focusing one-pointedly on the physical sensations of breathing (Marlatt & Kristeller, 1999; Wallace, 1999). However, many meditation practices cannot be clearly categorised as either mindfulness or concentration because they may involve both, but to a different degree. Although, there might be an overlap between mindfulness and concentration based techniques, mindfulness can involve expanding awareness from a single focal object to the full perceptual field and is conceptually different from concentration-based meditation approaches (Baer, 2003; Ivanovski & Malhi, 2007). For instance, there is a mindfulness exercise where a practitioner is sitting relaxed but erect at the same time and watches the space in front between the body and other objects without focusing on anything (Dalai Lama, Baron, & Gaffiney, 2004). The focal object here is a space about one meter in front at the eyes level. Expanding awareness from this point refers to the process of effortlessly widening perception so that all objects in perceptual field can be perceived simultaneously without focusing on anything specific. This exercise is specially designed to expand awareness to the full field of perception.

Mindfulness-Based Interventions

Available methods of mindfulness practice range from ancient instructions to modern, therapy-focused techniques (Baer, 2003; Germer et al., 2005; Kabat-Zinn, 1994). Basically, all mindfulness practices aim to focus the individual's attention on the present moment (Germer et al., 2005). In the last three decades, various types of mindfulness practice have been integrated into psychological treatments resulting in specific treatment methods such as Mindfulness-Based Stress Reduction (MBSR) (Kabat-Zinn, 1982, 1990), Mindfulness-Based Cognitive Therapy (MBCT) (Segal, Williams, & Teasdale, 2002), Dialectical Behavioural Therapy (DBT) (Hayes, Follette, & Linehan, 2004) and Acceptance and Commitment Therapy (ACT) (Hayes, Strostahl, & Wilson, 1999).

The first developed and widely practiced mindfulness-based therapy is MBSR that is typically spread over eight weeks with one two-hour session per week and one full day meditation retreat (Kabat-Zinn, 1982, 1990). Participants of such programmes typically receive meditation instructions, in which they are advised to practice six days per week, for a minimum of 45 minutes per day. For instance, sitting meditation involves sitting in a wakeful but relaxed posture with closed eyes and focusing one's attention on the breath. Other exercises include various Hatha yoga postures aiming at observation of bodily sensations with mindfulness. Also, participants are usually instructed to direct attention to a specified target like walking or breathing and to maintain awareness of it moment by moment (Kabat-Zinn, 1982). MBSR aims to focus the individual's attention on the present moment, which is a common feature to all mindfulness-based treatments (Germer et al., 2005). All sensations, emotions and thoughts arising during mindfulness practice are observed with non-judgmental acceptance and without analysis of their contents. Also, participants are encouraged to practise mindfulness during everyday activities like standing, walking, sitting and eating (Kabat-Zinn, 1982). Consequently, mindfulness practice helps to realise the transitory nature of thoughts, sensations and emotions, appearing and disappearing "like waves in the sea" (Linehan, 1993b).

MBCT also integrates mindfulness practice into the treatment (Germer et al., 2005). The core exercise taught to patients is the three-minute 'breathing space' including three main components: awareness, gathering and expanding. ‘Awareness’ refers to bringing one’s awareness to the present moment by adopting an erect but relaxed body posture and asking “What is my experience right now…in thoughts…in feelings…and in bodily sensations?” (Segal, Williams, and Teasdale, 2002, p. 184). The next step is ‘gathering’, which is redirection of one’s full attention to the in- and out-breath as they follow naturally. Finally, ‘expanding’ refers to expanding one’s awareness beyond breathing and includes facial expression, body posture and feeling of one’s body as a whole (Segal et al., 2002). In contrast to traditional cognitive therapy, which aims to alter an individual’s cognitions, MBCT promotes exploration of one’s feelings and thoughts from the state of mindfulness. As a result, an individual is able to see that “thoughts are not facts” and one can let them appear and disappear again regardless of their content (Germer et. al., 2005, p. 125).

DBT treatment unifies mindfulness of non-judgmental observation derived from Zen Buddhism with the Western contemplative traditions that promote unlimited acceptance of life’s suffering (Hayes et al., 2004). At the beginning of the intervention, the goal is to develop individual skills of observing thoughts, emotions and external stimuli by describing them. Similarly, DBT emphasises acting with awareness as a skill by cultivating it through a series of exercises that develop a routine of focusing attention on activities. Non-judgemental acceptance is also a primary skill that is recognised as part of the therapeutic process. To foster this skill, patients are encouraged to accept their reality and tolerate any unwanted feelings or thoughts without judgement (Linehan, 1993a, 1993b).

ACT is grounded on Relational Framework Theory that merges behavioural principles with mindfulness, acceptance and reconsideration of values (Hayes, Strostahl, et al., 1999). The aim of ACT is to teach clients to accept inevitable life suffering that is beyond an individual’s control while committing to activities consistent with the individual’s primary values that would make a life worth living (Hayes, Luoma, Bond, Masuda, & Lillis, 2006). Typically, a patient receives encouraging instructions to accept immediate external and internal experiences without judgements and at the same time strives to achieve specific behavioural changes, which are also modulated through operant conditioning techniques (e.g. reinforcement) (Gaudiano & Herbert, 2006). In order to help clients reduce the impact of negative feelings and thoughts, ACT teaches mindfulness skills including: *diffusion* or letting go of distracting cognitions (e.g. memories, beliefs); *acceptance* of unpleasant sensations, drives and emotions; and *present moment focus* associated with alertness and openness.

Similar to ACT, a mindfulness component was integrated in another therapeutic approach, which primarily uses behavioural methods and refers to behavioural activation (BA) treatment. In BA, mindfulness is applied to address dysfunctional ruminations leading to pathological mental states such as depression. BA treatment does not involve altering the content of participants' thoughts, but instead patients are instructed to be mindful and, by noticing their own rumination, to switch their attention immediately towards external, environmental stimuli (Jacobson, Martell, & Dimidjian, 2001). This application of mindfulness, in combination with behavioural activation treatment, appears to have comparable efficacy to traditional cognitive therapy and medication for treatment of major depression (Dimidjan et al., 2006).

Mindfulness-integrated Cognitive Behaviour Therapy (MiCBT) is another example of employing mindfulness for therapeutic purposes that combines mindfulness practice with cognitive-behavioural methods to regulate attention and emotion. This approach emphasises neurophysiological similarities of reinforcement mechanisms involved in both mindfulness practice and operant conditioning and proposes a neurophenomenological reinforcement model that applies an extinction principle (Cayoun, 2011). According to this model, a perceived trigger is interpreted by higher cortical structures resulting in related body sensations, which in turn lead to 'mindless' or automatic responses. Here, reinforcement is understood as neurological dependence from learned responses associated with paired sensations and cognitions. It was shown that these conditioned responses related to any disorder can be extinguished if the patient remains in the state of mindfulness observing and accepting thoughts and sensations of the body without judgments and analysis (MiCBT Institute, 2011).

MiCBT employs a 4-stage therapeutic model to teach clients to achieve emotional stability and regulate attention. The first *personal stage* focuses on internalising attention through formal mindfulness training that aims at regulation of internal experiences, emotions and thoughts. The second *exposure stage* utilises behavioural exposure techniques to reduce reactivity and develop self-confidence while dealing with external life experiences. The third *interpersonal stage* focuses on interpersonal skills that involve expanding attention towards others and learning to inhibit emotional responses triggered by reactions of others. Finally, the fourth *empathic stage* aims at developing empathy and compassion based on immediate experience (Cayoun, 2011).

Generally, MBSR, MBCT and MiCBT emphasise the central role of mindfulness in the therapeutic process, but MBCT and MiCBT also add cognitive-behavioural elements to the treatment (Segal et al., 2002, 2013). However, in other approaches such as ACT, DBT and BA mindfulness is just a sub-component among other treatment tools, which is used in a comparatively limited way to increase sensory and perceptual awareness in normal, non-meditative circumstances (Hayes, Strosahl, et al., 1999; Linehan, 1993a).

There is growing evidence for the efficacy of mindfulness-based interventions including MBSR, MBCT, DBT and MiCBT in alleviating the symptoms and enhancing the coping abilities of people suffering from anxiety, stress, depression, emotional instability, substance abuse, post-traumatic stress disorder, borderline personality disorder, psychophysiological disorders and suicidal/self-harm behaviour (Chiesa & Serretti, 2010; Dimidjan et al., 2006; Hayes et al., 2006; Hofmann, Sawyer, Witt, & Oh, 2010; Ivanovski & Malhi, 2007). ACT treatment was also found effective for a number of clinical conditions such as obsessive compulsive disorder, depression, anxiety, anorexia, workplace stress, chronic pain, PTSD, psychosis and substance abuse (Bach & Hayes, 2002; Bond & Bunce, 2000; Branstetter, Wilson, Hildebrandt, & Mutch, 2004; Dahl, Wilson, & Nilsson, 2004; Twohig, Hayes, & Masuda, 2006; Zettle & Raines, 1989). BA treatment that integrated mindfulness component was found effective for treatment of major depression (Dimidjan et al., 2006). Also, the MBSR was shown to be effective in reducing sympathetic activity of fibromyalgia patients suffering from chronic pain as measured by skin conductance level during mindfulness practice before and after the treatment (Lush et al., 2009).

Even though mindfulness research with the general population remains limited, some validation studies report that mindfulness correlates positively with positive affect, well-being and openness and negatively with stress, anxiety, rumination, neuroticism and dissociation (Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006; Brown & Ryan, 2003; Carlson & Brown, 2005; Christopher & Gilbert, 2010; Frewen, Evans, Maraj, Dozois, & Partridge, 2008). However, most of these outcome studies used psychometric instruments to measure symptoms, but failed to measure/report expected changes in mindfulness levels and skills (Baer et al., 2004; Brown & Ryan, 2003; Cohen-Katz, Wiley, Capuano, Baker, & Shapiro, 2005; Hofmann et al., 2010). Research investigating the relationship between mindfulness and psychological health often employed ordinal scales (Hofmann et al., 2010; Christopher & Gilbert, 2010) that should not be used with parametric statistics without violating their fundamental assumptions (Stucki et al., 1996; Allen &

Yen, 1979). This could potentially result in misleading conclusions and implications of the findings linked to an individual's health and well-being. Therefore, there is a need for accurate measurement of mindfulness and related outcomes to assess temporary and enduring psychological changes associated with a specific type of mindfulness practice and the related therapeutic outcome.

Neurophysiological Research on Mindfulness

Neurophysiological studies on mindfulness mainly used Electroencephalogram (EEG), and a few studies used neuroimaging techniques such as Magnetic Resonance Imaging (MRI) and physiological measures (e.g. skin conductance level) (Cahn & Polich, 2006; Chiesa & Serretti, 2010; Coelho, Canter, & Ernst, 2007). Electroencephalogram (EEG) records postsynaptic electrical potentials generated by large populations of cortical neurons using electrodes attached to the scalp. These electrical potentials have waveforms of specific frequency and amplitude and correlate with the current psychological state of an individual. For instance, Gamma waves (30-100 Hz) are associated with cognitive activity or pain; Beta (12-30 Hz) are related to awake, alert, working; Alpha (8-12 Hz) are linked to a relaxed, reflective state; Theta (4-8 Hz) are related to the transition between sleep and wakefulness (e.g., meditation, drowsiness); and Delta (< 4 Hz) when we engage in deep sleep (Cahn & Polich, 2006; Gazzaniga, Ivry, Mangun, & Steven, 2009).

One pioneering study (Kasamatsu & Hirai, 1966) compared EEG recordings of Zen practitioners before, during and after meditation performed with open eyes with EEG recordings of non-matched controls. Overall, decrease of Alpha frequency within the Alpha band, increased frontal Alpha activity, and bursts of Theta, associated with high levels of meditation experience were reported during meditation compared to 'before meditation' condition and controls (Kasamatsu & Hirai, 1966; Murata, Koshino, & Omori, 1994). These findings were replicated by a later study (Takahashia et al., 2004). Also, Kasamatsu and Hirai (1966) reported a lack of Alpha-blocking habituation in response to repeated click sounds in Zen meditators, but not in controls. Alpha-blocking refers to a reduction of Alpha power after stimulus presentation compared to pre-stimulus EEG. This effect typically disappears after 10-20 repeated stimulus presentations, indicating habituation (Barlow, 1985). Lack of Alpha habituation in Zen practitioners may suggest alert attention to the present moment while in a relaxed state of meditation (Kasamatsu & Hirai, 1966). However, the difference in Alpha-blocking between Zen meditation and control groups was not found in the later study (Becker & Shapiro, 1981), possibly due to the differences in meditation experience between samples.

Dunn et al. (1999) compared the brain activity measured by EEG during mindfulness practice, focused meditation, and normal relaxation conditions of ten students after focused meditation and mindfulness training. Significantly higher mean amplitudes of frontal and posterior Delta, frontal Theta, central and posterior Alpha and Beta at all cortical areas were reported during mindfulness compared to both focused meditation and relaxation. The authors concluded that mindfulness meditation is a unique form of consciousness distinct from relaxation and focused meditation (Dunn, Hartigan, & Mikulas, 1999). However, in both conditions, relaxation and focused meditation, the participants were sitting with closed eyes in contrast to open eyes during mindfulness conditions. This is a limitation of Dunn et al. (1999) study, because the EEG based evidence shows that the eyes-open condition is associated with increased average of Alpha, Beta, Delta and Theta activity measured across scalp compared to eyes closed condition (Barry, Clark, Johnstone, Magee, & Rushby, 2007). Also, Dunn et al. (1999) compared relaxation, focused meditation and mindfulness using the same group of participants, who were first trained in focused meditation and then in mindfulness, assuming that they can easily switch from one state to another. For instance, an acquired habit to be mindful would naturally interfere with other experimental conditions as evidenced from efficiency of mindfulness based clinical interventions (Chiesa & Serretti, 2010). Considering that acquired mindfulness skills are likely to influence global neurophysiological functioning in daily life, Dunn et al. (1999) proposed that future studies should examine mindfulness in non-meditative conditions.

Davidson et al. (2003) compared the cortical activity of 25 participants before and after MBSR to 16 ‘wait list’ controls using EEG and psychometric measures of anxiety and affect. Significantly higher left-side anterior activation (C3/4) that is linked to positive affective style was reported for the mindfulness group compared to controls, but the band frequency was not specified. These changes in frontal Alpha asymmetry were not found by a later study in the treatment groups undergoing the MBCT (Keune, Bostanov, Hautzinger, & Kotchubey, 2011). Also, Davidson et al. (2003) found a significant decrease of negative affect, anxiety, and rise of antibody cells that were associated with mindfulness compared to the control group.

Another example is a recent EEG study which indicated that mindfulness practice is associated with reduced gamma power in frontal areas, which is linked to decreased self-referential processing, and increases of posterior gamma activity related to heightened sensory attention (Berkovich-Ohana, Glicksohn, & Goldstein, 2011). The authors claim that differences in gamma band (compared to controls) are found for mindfulness practitioners regardless of their experience level. However, Berkovich-Ohana et al. (2011) reported the results for closed-eyes conditions only, which limits generalisability of their findings to closed-eyes conditions.

One fMRI study reported that experienced Zen practitioners exhibit decreased duration of neural activity associated with conceptual automatic thinking compared to controls. The authors suggested that higher level of experience in meditation may facilitate voluntary regulation of mental flow (Pagnoni, Cekic, & Guo, 2008). Also, evidence obtained using MRI shows that grey matter volume correlates negatively with both attention task performance and age in the normal, non-meditative population but not in experienced Zen practitioners. These findings suggest that Zen meditation practice may prevent age-related cognitive deterioration by inhibiting reduction of grey matter volume (Pagnoni & Cekic, 2007). Consistent with these findings, a more recent study reported increased grey matter density in the brain regions involved in emotion regulation, learning, memory and self-related cognitions after MBSR training. These areas include hippocampus, cingulate cortex, cerebellum and temporo-parietal junction (Hölzel et al., 2011). One recent fMRI study reported that mindfulness practice produces significant signal changes in the brain regions involved in self-perception and regulation of emotion resulting in altered experiences of self (Ives-Deliperi, Solms, & Meintjes, 2010). The main methodological limitations of this study were lack of a control group and using mental task of generating numbers as a baseline for comparison (Ives-Deliperi et al., 2010).

Research investigating effects of mindfulness on autonomic function using acceptable methodology is lacking (Cahn & Polich, 2006; Chiesa & Serretti, 2010). Takahashi et al. (2004) compared heart rate variability (HRV) of 20 Zen practitioners during meditation to that of 20 naïve controls and found increase of the high frequency (HF) power indexing parasympathetic activity (Acharya, Joseph, Kannathal, Lim, & Suri, 2006) and reduction of the low frequency (LF) and LF/HF ratio, both associated with sympathetic activity (Thayer, Hansen, Saus-Rose, & Johnsen, 2009). This evidence suggests that mindfulness practice seems to play an important role in facilitating autonomic balance by reducing sympathetic and promoting parasympathetic activity.

Other neurophysiological studies that have methodological limitations such as using simplified mindfulness-related techniques (e.g. breathing exercise) or small samples (< 10) (Barnhofer, Chittka, Nightingale, Visser, & Crane, 2010) are not included here. The common limitations of neurophysiological studies include lack of control for mindfulness levels, invalid comparisons between groups (e.g. eyes open vs. eyes closed), and no matching between groups by demographic variables such as gender, age and ethnicity that all affect EEG data (Barry et al., 2007; Erwin, Mawhinney-Hee, Gur, & Gur, 1989).

Overall, neurophysiological studies indicate that mindfulness practice correlates with biological changes and the reduction of psychological symptoms. However, a general limitation of neurophysiological studies on mindfulness is the lack of precise control for individual mindfulness levels and clear distinction between state and trait mindfulness. There are mainly two reasons for that: 1) reliable and valid psychometric measures of mindfulness were not used; 2) applied (available) mindfulness measures have limited precision and may fail to distinguish clearly between state and trait mindfulness (Park, Reilly-Spong, & Gross, 2013; Chiesa & Serretti, 2010; Medvedev et al., 2016a; 2017c). Usually, participants were selected for a study based on their experience of mindfulness practice (e.g. years), which might not be accurate criteria because there are different mindfulness practices (e.g. Zen, MBSR, MBCT) that might have different efficiency over time. Also, individuals differ in capacities to acquire mindfulness skills and consequently different amount of practice time might be necessary for different individuals to achieve the same mindfulness level (Dalai Lama et al., 2004; Kabat-Zinn, 2000). If state and trait mindfulness levels cannot be reliably measured in neurophysiological studies, then there is nothing to compare/correlate with neurophysiological data at all. Specifically, the selection of the more experienced mindfulness practitioners is compromised because trait mindfulness levels of the prospective participants cannot be reliably assessed.

Thus, improving the precision of existing mindfulness instruments will be beneficial for neurophysiological studies (e.g. EEG) on mindfulness to control for both trait and state mindfulness levels. Enhancing psychometric properties of widely used mindfulness measures and their ability to distinguish clearly between state and trait mindfulness will be especially useful to assess temporary and enduring psychological changes associated with a type of mindfulness practice and any related therapeutic outcome.

Measurement Theories

Classical Test Theory (CTT) is the most prominent theory of measurement covering the construction and validation of psychometric instruments and dominated psychometric thinking and work in the 20th century. CTT foundations were established by psychologists such as C. Spearman, J. Cronbach and involved such psychometric luminaries as R. B. Cattel, L. Guttman, L. L. Thurstone, J Loevinger and others working in the field (Lord & Novick, 1968; Cohen & Swerdlik, 2010). CTT postulates that an observed score (*O*) consists of both a true score (*T*) and an error score (*E*) values expressed by the basic formula (Lord & Novick, 1968):

$$\boldsymbol{O = T + E} \qquad (1)$$

Here, true score refers to a mean score that would be obtained if a measure is applied countless number of times and related to consistency rather than to validity of the score (Steiner & Norman, 2008). Essentially, CTT uses correlational methods, which are inextricably linked to factor analytic techniques. Based on this approach, construction of a test is based on item-to-total correlations to establish the overall consistency of a scale and determine dimensionality within a scale by means of factor analysis. Modern factor analytic approaches typically involve exploring dimensionality of a scale if it is not yet established using exploratory factor analysis (EFA) and then testing psychometric properties of the factor structure of a scale by means of confirmatory factor analysis (CFA) (Nunnally & Bernstein, 1994).

The main advantages of CTT for developing a scale include popularity of its methods and ease of application with available statistical software (e.g. IBM SPSS). However, the application of CTT in test construction tend to result in scales with a large number of items. Also, psychometric properties of a scale tested by CTT methods largely depend on the sample used to construct a measure, and a scale developed in this way tends to capture differences mainly in the middle levels of the latent trait (DeVellis, 2006). This mainly results from treating all items as equal contributors in measuring the unobservable (latent) construct. However, item summary scores may not be an accurate estimate of the latent trait as different items may explain a different amount of information relevant to the latent trait, which is not considered if the total score is calculated (Stucki, Daltroy, Katz, Johannesson, & Liang, 1996; Allen and Yen 1979). Technically, ordinal scales cannot be used to calculate means and standard deviations because they do not support mathematical operations of adding, subtracting, dividing and multiplying (Merbitz, Morris, & Grip, 1989).

CTT summarises measurement error as a single variable, even though in fact it reflects variability due to different sources that affect observed scores (Bloch and Norman 2012). In naturally present environments, there are more factors including personal (e.g. personality, age) and situational (e.g. time of the day, room temperature) that might contribute to measurement error. Proponents of alternative approaches challenge CTT's basic assumption of a true score by arguing that administration of the same test to the same person countless times to obtain a true score and an error score appears unrealistic if not impossible (Borsboom, 2005). Thus, relying on CTT assumptions may lead to inaccurate conclusions about measurement instruments used to assess people (Hobart & Cano, 2009).

Item Response Theory and the Rasch Measurement Model

The Item Response Theory (IRT) approach was first published in 1968 and has since had a rapid gain in popularity for developing and evaluation of interval scales to measure abilities (Lord, Novick, 1968). Unlike CTT that uses correlations between items (or item-to-total) to select items measuring a latent construct, IRT methods apply a statistical model that predicts the precise mathematical relationship between an item and a latent trait to be measured. Proponents of IRT argue that, compared to the typically long scale with ordinal scores produced by CTT methodology, application of IRT results in a shorter, reliable measure with interval level scores (Embretson, 1996).

IRT models explain the relationship between an individual's ability on a latent trait (e.g. mindfulness) denoted as θ (theta) and response probability to the item. This relationship is presented graphically by an item characteristic curve (ICC), which shows item/person location on the scale of latent trait (x-axis) and the expected value for a person with specific ability on this item (y-axis). Figure 1 uses an example from the study described in Chapter Nine and shows ICC for two items from the functional assessment measure UK FIM+FAM (Turner-Stokes et al.,1999) measuring a patient's motor (physical) abilities using 7 response options (0-6). Y-axis indicates as a function of person ability (x-axis). Typical ICC is a monotonic (non-linear) function. The horizontal midline between 0 and 6 response categories used to determine location of the item at the intersection of the ICC with this line. There are three example patients with different levels of motor abilities (A, B and C) located on the x-axis. The ICCs for item 1 (eating) and item 2 (locomotion) can be explained as follows. Patients with the motor ability A (θ = -1) is expected to score lower (0.25) on the item 2 meaning that they are likely to score at 0. However, they are expected to score 1.20 and thus likely to score 1 on the item 1. Therefore, at this level of motor function ability to eat is less affected than ability to move. Patients with the ability B (θ = 0) are expected to score at 3 (the midpoint) on item 2 because this item has the same level of difficulty (θ = 0). Patients with this level of ability are expected to score 5 on item 1 because it is easier item compared to item 2. Accordingly, patients at this level of motor ability have moderate ability to move but fairly good ability to eat. Patients with the motor ability C ($\theta \approx 0.7$) are expected to score 5 on item 2 and 5.75 on item 1 and therefore likely to select option 6 on this item. At this level of motor ability patients have relatively similar good ability to eat and to move.

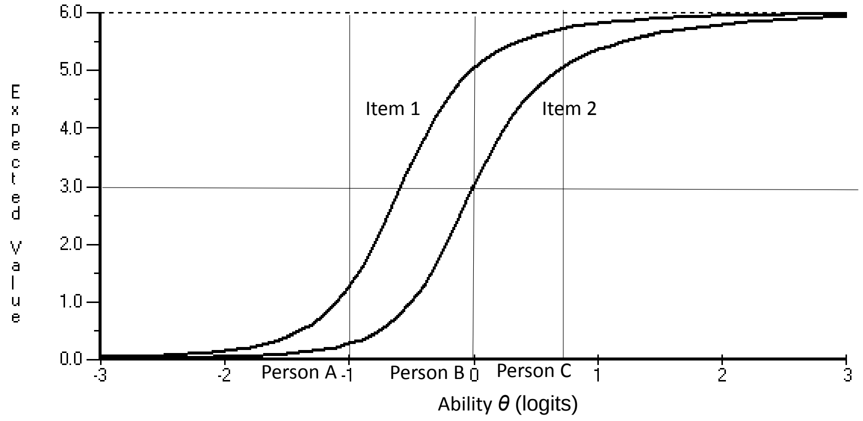


Figure 1. Item characteristic curve for item 1 (Eating) and 2 (Locomotion) of the functional assessment measure UK FIM+FAM.

The easier item is located towards the left hand-side and more difficult item on the right hand side of the scale (Figure 1). Similarly, patients with higher latent ability are located on the right-hand side and with lower on the left of the x-axis. In this example item 2 is more difficult compared to item 1 and patient C is more functional compared to patient A. Therefore, IRT models can accurately assess and relate both item location or difficulty and person ability on the latent variable. Consequently, persons with greater ability are more likely to endorse higher response options on both easy and difficult items while persons with low ability can only endorse higher response options on the easiest items. More complex IRT models include two or more parameters such as item discrimination and level of asymptote due to guessing. Two-parameter logistic model refers to 2PL and three-parameter model to 3PL.

Rasch model (Rasch, 1960, 1961) is a unidimensional, probabilistic, logistic model, which postulates that responding to a particular test item is influenced by just two variables - person ability (qualities of the person) and item difficulty (qualities of the item). Therefore, only one parameter - an ability - is estimated by this basic model. It was developed before the first IRT framework was published (Lord & Novik, 1968). In the literature Rasch model it is sometimes considered as a special case of IRT because it is mathematically similar to the one parameter IRT model, which also refers to simple logistic model (SLM) (Hobart & Cano, 2009). However, Rasch and IRT represent two different paradigms: IRT aims at finding the right model that best explains the data while Rasch model defines fundamental measurement (Thurstone, 1931) and the data should fit the model to meet its requirements (Tennant & Conaghan 2007; Hobart & Cano, 2009). Fundamental measurement criteria were formulated by Thurstone (1931), which are similar to the laws of physics (Rasch, 1960). A measurement instrument should 'transcend' sample groups meaning that a scale must work equally well for every person regardless of personal factors (e.g. gender). The measurement should estimate only one parameter of the measurement object, which is a universal criteria for all measurement and refers to unidimensionality in the Rasch model. A measurement unit should be the same at every part of the scale continuum, which refers to *additivity criterion* required for an interval level measurement. The Rasch model conforms to these criteria because it requires unidimensionality and scale/items invariance across sample groups, and it produces measurement units in logit values used to locate items and persons on the same continuum of the latent trait (Rasch, 1960, 1961; Brogden, 1977).

One distinct advantage of Rasch analysis over classical psychometric methods is in examining internal construct validity where Rasch requires unidimensionality, a monotonic relationship between item responses and the latent variable, local independence, and no item bias among sample subgroups (Tennant & Conaghan, 2007). The other key advantages include estimating difficulty (location) of every item, testing appropriate ordering of response options of polytomous items and finally transformation from an ordinal to an interval measure (Rasch, 1960; Wilson, 2005; Wright & Stone, 1979; Hobart & Cano, 2009). Thus, Rasch analysis can generate questionnaire scores, which are based on a genuine interval scale (Bond & Fox, 2007). This is different from the raw, ordinal scores that are available from the original scale and can differentiate only in terms of rank order amongst response options. In contrast, Rasch analysis accurately estimates thresholds between response options of an item by accounting for both item difficulty and sample abilities. Threshold refers to the level of a latent trait when the probability of choosing one of two subsequent response options is the same (Andrich 1978). Based on estimated item thresholds and sample abilities, the Rasch model generates a template for transformation of ordinal scale responses into interval-level data, given that a scale is unidimensional (Tennant & Conaghan 2007).

When data fit the Rasch model, the interval-transformed scores will accurately reflect changes on a latent trait similar to any other interval measure such as length or temperature. Thus, the Rasch model produces a scale that is superior to the ordinal version derived from CTT because of its conformity to the principles of fundamental measurement and production of a linear interval scale (Rasch, 1960; Bond & Fox, 2007). Its value has been demonstrated at the group level for such scales where scores are summed up and compared across different subgroups (Khan, Chien, & Brauer, 2013; Lundgren Nilsson et al., 2005), as well as at the individual level in terms of patient responsiveness (Hobart, Cano, & Thompson, 2010).

The dichotomous Rasch model was developed first, which is expressed by the following formula (Rasch, 1960):

$$p_i\,(\theta) = \frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}} \qquad (2)$$

It refers to a simple logistic model, where $p_i\,(\theta)$ is the probability that a respondent with ability θ will respond positively to an item, and δ_i is an item difficulty parameter, which is the only relevant parameter in the Rasch model. If $\delta = \theta$, the probability to answer an item positively or negatively is equal ($p_i\,(\theta) = 0.5$). The denominator in the formula (2) functions as a normalising factor to warrant the probability for positive response ranging from 0 to 1.

Following this, two parameterisations were developed for polytomous items (with three or more response options) including the Partial Credit Model (Masters, 1982) and the Rating Scale Model (Andrich, 1978). The polytomous models routinely estimate a threshold for each response option, which refers to the level of a latent trait when the probability to choose any of two subsequent response options is the same. Both polytomous Rasch models assume that differences between thresholds of individual items vary. However, the Rating Scale Model (Andrich, 1978) assumes that these variations are uniform across all items and can be expressed as follows:

$$\ln\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - \delta_i - \tau_j \qquad (3)$$

This formula estimates probability of a person n (θ_n) to respond to each response option (j) of an item (i) by including a threshold parameter for each response option (τ_j). In contrast, the Partial Credit Model (Masters, 1982) allows thresholds distances to vary across items, so that every item has individual rating scale parameters that can be expressed as:

$$\ln\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - \delta_i - \tau_{ij} \qquad (4)$$

This formula estimates threshold parameters for each individual item (τ_{ij}) independently. Thus, both polytomous Rasch models estimate a response probability for each category of each polytomous item and are widely used in health measurement to evaluate and improve psychometric properties of ordinal scales (Lundgren Nilsson & Tennant, 2011; Hobart & Cano, 2009). The decision which polytomous model to use is based on the likelihood-ratio test conducted before analysis, which compares threshold distances between individual items. If these distances are significantly different, the unrestricted Partial Credit model will be used (Tennant & Conaghan, 2007).

The Rasch model involves the testing of several attributes: (i) ordering of response option thresholds in polytomous items, (ii) item-trait interaction, (iii) unidimensionality, (iv) local independence assumptions, and (v) potential item bias. Applying Rasch analysis to mindfulness and outcome measures would be beneficial through the identification of item bias (DIF) when respondents from different groups (e.g. meditators vs non-meditators) with the same level of latent trait respond differently to an item. When data fit the Rasch model, these parameters meet the model expectations so the items can be ordered by their difficulty and the participants by their ability on the latent trait (e.g. mindfulness) using the same log-odds interval scale. One important end product of Rasch analysis is an algorithm to transform scores from an ordinal to an interval scale to increase the precision of measurement (Brogden 1977; Rasch 1961), which has been demonstrated empirically (Norquist et al. 2004).

Taken together, IRT and Rasch measurement models are both theoretically and practically advanced compared to traditional CTT methods. In particular, benefits of Rasch analysis over traditional psychometric methods for investigating and improving psychometric properties of ordinal scales such as enhanced precision of measurement and better targeting of sample abilities have been demonstrated empirically by a number of studies (Hobart & Cano, 2009; Norquist et al. 2004). Therefore, there are considerable advantages in using Rasch analysis in both mindfulness and health-related outcome measurement.

Generalisability Theory and the State versus Trait Distinction

A trait refers to a relatively stable characteristic or enduring behavioural pattern displayed by a person, while a state represents an individual’s experience in a given moment, situation or condition (Hamaker et al. 2007; Spielberger et al. 1970). For example, a student who is generally fairly relaxed and non-anxious (trait) can become quite anxious and tense just before or during an important final exam (state). Essentially, a state is determined by interaction between person and occasion and reflects an individual’s unique adaptation to the present moment and environment (Buss 1989; Epstein 1984). Both state and trait and their interactions are considered important to understand variability and steadiness of an individual’s functioning (Buss, 1989; Epstein, 1984). However, reliability and validity of psychological measurements such as mindfulness may be compromised through confounding of mindfulness as a state and a trait. It is important to develop and apply reliable methods for distinguishing between the two, otherwise therapeutic interventions, for example, cannot be properly assessed for their effectiveness over time. Mindfulness-based interventions aim at lasting or trait changes, and if only state changes are achieved during treatment, relapse is inevitable. This is because state can be explained as more short-term experience (e.g. immediately after a session), whereas trait refers to a pattern established over the longer term (e.g. lasting beyond completion of a mindfulness programme).

Generalisability Theory (GT) is an analytical technique for data acquired using psychometric instruments (e.g. rating scales, performance tests). It is named GT because it estimates the extent to which the influence of any specific source of error variance can be generalised to all possible situations and contexts as opposed to only a limited amount of data obtained from a specific testing situation (Cronbach et al. 1963). GT assesses numerous sources of variance contributing to the measurement error associated with the main variable of interest (e.g. a mindfulness score) (Allal & Cardinet, 1976). It represents an extension of classical test theory (CTT), based on the idea that every score consists of both true and error values, but it goes beyond its limited assumption considering error variance as a single factor (Allen & Yen, 1979). In naturally occurring environments, there are more factors including personal (e.g. personality), methodological (e.g. psychometric characteristics of the measure used) and situational (e.g. time of the day) that might each independently contribute to measurement error. GT provides an advanced method for assessing these factors and their interactions thus contributing to the improvement of methodology and precision of an assessment instrument.

GT employs repeated-measures factorial analysis of variance (ANOVA) to estimate the relative contribution of different sources of variability to the overall measurement error, which is also referred to as ‘noise’ (Brennan, 2001). Every such contribution can be expressed as an intra-class correlation coefficient (ICC) ranging from 0 to 1, similar to other reliability coefficients. For instance, the amount of variance between mindfulness scores that is explained by differences between the participants can be represented as an ICC that reflects the discriminative ability of the mindfulness questionnaire as follows (Bloch & Norman, 2012):

$$\mathbf{ICC} = \frac{\boldsymbol{variance}\ (\boldsymbol{participants})}{\boldsymbol{variance}\ (\boldsymbol{participants}) + \boldsymbol{variance}\ (\boldsymbol{error})} \qquad (5)$$

Here, ICC depends on two factors: the actual ability of an instrument to discriminate between participants and amount of noise due to other influencing factors. ICC is a reliability coefficient that expresses the ratio between the amount of variance in scores attributed to the primary variable being measured and the total amount of observed variance. ICC was originally introduced in CTT, represented by a slightly different but essentially similar formula using the concept of ‘signal-to-noise ratio’ (SNR) (Fisher, 1925, 2006). SNR is mathematically equal to the square of the effect size (ES^2), which could be extracted from any ANOVA analysis and represents a ratio between consistent change (variance) in the X variable that refers to ΔX and total variance (σ^2) in the data (Bloch & Norman, 2012):

$$\mathrm{SNR} = \mathrm{ES}^2 = \frac{\Delta \mathrm{X}^2}{\sigma^2} \qquad (6)$$

Therefore, ICC based on SNR definition is expressed by the following formula:

$$\mathrm{ICC} = \frac{\mathrm{SNR}}{1+\mathrm{SNR}} \qquad (7)$$

The larger the amount of variance in a variable of interest (signal) compared to noise, the better are the chances to detect these changes reliably. An ICC close to 1 would indicate that there is mainly a real difference related to signal and relatively low amount of noise, and an ICC close to 0 would indicate that there was mainly noise or error in the data. ICC refers to a G-coefficient in GT terminology and similarly expresses the ratio of the observed (true) variance due to the object of measurement (σ_p^2) and the total variance of universe scores including the observed (true) variance and the error variance (σ_{error}^2) (Brennan 1992; Shavelson et al. 1989):

$$G_p = \frac{\sigma_p^2}{\sigma_p^2+\sigma_{error}^2} \qquad (8)$$

Here, p refers to a person effect because person is commonly an object of psychometric measurement. A G-coefficient is normally computed for the variable of interest (e.g. trait mindfulness) but can also be computed for every factor contributing to error variance, given that a research design provides relevant data to assess variability due to these contributions (Bloch and Norman 2012). In this case, the G-coefficient expresses the generalisability of influence attributed to specific factors to all possible situations and contexts.

GT can be used to identify and compare the amount of variance uniquely explained by the person, the item and the occasion plus their respective interactions (Brennan 2001; Bloch and Norman 2012). The variance due to person-occasion interaction is a direct reflection of the 'stateness' of a latent construct, while person variance alone is representative of a trait (Buss 1989; Chaplin et al. 1988; Epstein 1984). Importantly, GT permits this analysis for the total test, subscales and even individual items. In other words, true 'state items' can be distinguished from items that are not truly sensitive to occasion. Estimation of variance associated with the object of measurement (e.g. persons) and influencing facets (e.g. occasions) is conducted in a G-study (generalisability study). Variance components are estimated based on observed values obtained from the universe of all possible (hypothetical) observations. Scales and individual items measuring state are expected to reflect a higher amount of variance attributed to person-occasion interaction and low generalisability across occasions (e.g. G<.70) as opposed to reliable trait measures, which are expected to have a G of .80 or higher (Arterberry et al. 2014; Gardinet et al. 2009). However, traits are the basic determinants of states through interaction with situational factors for the same latent construct, and a precise distinction between state and trait can only be estimated based on their variance components (Hamaker et al. 2007; Geiser et al. 2015). To date, there are no commonly accepted benchmarks for the relative proportions between state and trait components in a valid state measure.

Arguably the best validated state and trait self-report measure is the State and Trait Anxiety Inventory (STAI) (Spielberger et al., 1970). The STAI is a 40-item questionnaire split into state and trait subscales of 20 items each. It was proposed as a measure of anxiety as both a trait, related to general perception of the environment as dangerous, and as a state that refers to the experience of anxiety at the present moment only. Validation studies of STAI consistently report lower test-retest scores obtained for the state subscale (.16 – .57) compared to the trait subscale (.78 – .83) over various time-intervals that confirm the expectations of the state-trait relationship (Ramanaiah, Franzen, & Schill, 1983; Spielberger, 1999). State and trait subscales correlate with each other in the range between .70 and .80 (Ramanaiah et al., 1983). Other examples of state and trait measures are the State-Trait Anger Expression Inventory-2 (Spielberger, 1999) and the Positive And Negative Affect Scale (PANAS) (Watson, Clark, & Tellegen, 1988). All three mentioned measures include instructions clarifying temporal aspects of responses for a participant. For the state subscale, a participant has to respond how they feel 'right now, at this moment' and for the trait subscale, how they feel 'generally'.

The traditional method for demonstrating distinct state and trait components in a scale has been to examine test-retest reliability coefficients, which are expected to be lower for a valid measure of state (e.g. <.60) and higher for a trait measure (e.g. >.70) (Ramanaiah et al. 1983; Spielberger 1970, 1999). The main limitation of this method that it is based entirely on the total score correlations at Time 1 and Time 2. If relationships and distinctions between trait and state are to be given a systematic and robust foundation, there is a need to properly understand the different contributions made by individual item effects, scale effects, person effects and occasion effects to changes in trait and state. Identifying such effects will require a much deeper analysis of variances found in the different dimensions of the research study so that such variances can be identified and isolated if necessary to provide greater control in future experimental studies. Most importantly, the test-retest coefficient fails to account for variability due to interaction between person and occasion, which is an essential determinant of state changes in an individual (Buss 1989; Chaplin et al. 1988; Epstein 1984). In other words, we do not expect trait scores to vary substantially across situations. In contrast, the interaction between the person and the occasion is state virtually by definition.

To date, the exploration of state and trait variability is limited to structural equation modelling (SEM) approaches (e.g. Hamaker et al. 2007; Geiser et al. 2015; Kenny and Zautra 2001). Various analytical models using SEM were proposed to investigate differences between trait and state variability (Hamaker et al., 2007; Kenny & Zautra, 1995; Steyer, Ferring, & Schmitt, 1992). However, none of the proposed SEM methods account for all the important sources of variance (e.g. individual items and person x item interactions) contributing to the measurement error associated with state and trait variability, which limits their applicability for validation of state and trait measures. Such differences in variability require a more detailed or micro-level study of how factors or components that can affect state and trait, including person and situation, person x occasion interaction, and the ‘stateness’ or ‘traitness’ of the individual items, can be quantified so that changes in state and trait can be predicted by knowing of changes in person and situation, which is a true generalisability, in other words. Therefore, further exploration is necessary to reliably differentiate between state and trait variance components in a measure. While GT has been previously applied to assess reliability of trait measures (e.g. Arterberry et al. 2014), to date there are no studies that have used GT methods to distinguish between state and trait components in a state measure.

Psychometric Measures of Mindfulness

A number of self-report questionnaires are available to assess individual levels of mindfulness. Table 2 provides a list of commonly cited instruments with their demonstrated psychometric properties (Baer, Smith, & Allen, 2004; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006; Brown & Ryan, 2003; Cardaciotto, Herbert, Forman, Moitra, & Farrow, 2008; Chadwick et al., 2008; Feldman, Hayes, Kumar, Greeson, & Laurenceau, 2007; Haigh, Moore, Kashdan, & Fresco, 2011; Lau et al., 2006; Walach, Buchheld, Buttenmuller, Kleinknecht, & Schmidt, 2006; Bergomi, Tschacher & Kupper, 2014). Measurement of mindfulness is a relatively new research area, with the first self-report mindfulness measure, the Freiburg Mindfulness Inventory (FMI), published in 2001 (Buchheld, Grossman, & Walach, 2001; Walach et al., 2006). The main purpose of the FMI was to provide a quantitative assessment of an individual’s mindfulness and to monitor changes associated with meditation practice.

Table 2. *Properties of commonly used mindfulness scales and number of citations in Google Scholar and Web of Science (20th of January 2017).*

Scale	Reference (original)	Subscales (Items)	Cronbach's Alpha	Reliability Test-retest	Validity of Construct	Google Scholar	Web of Science
MAAS	Brown & Ryan (2003)	1 (15)	α = .78 - .92	ICC*= .81; p<.001	FMI, KIMS, MMS, CAMS-R, SMQ: r = .14 - .51; p<.05	5341	1910
FFMQ	Baer et al. (2006)	5 (39)	α = .67 - .93	not reported	Based on FMI, KIMS, MMS, SMQ, and MAAS items.	2815	1141
KIMS	Baer et al. (2004)	4 (39)	α = .72 - .97	r = .81 -.86, Observe r = .65	FMI, SMQ, CAMS r = .51 - .67	1525	535
TMS	Lau et al. (2006)	2 (13)	α = .85 - .91	not reported	KIMS, MAAS,FFMQ CAMS-R, SMQ, FMI r = .10 - .74; p<.05	693	243
FMI	Buchheld et al. (2001)	1 (30)	α = .80 - .94	not reported	KIMS, MAAS, SMQ, and CAM-R: r = .31 - .60	611	257
	Walach et al. (2006)	1 (14)	α = .86	not reported			
CAMSR	Feldman et al. (2007)	1 (12)	α = .61 - .81	not reported	KIMS, MAAS, FMI and SMQ: r = .51 - .67; p<.05	566	209
PHLMS	Cardaciotto et al. (2008)	2 (20)	α = .75 - .91	not reported	KIMS, MAAS: r = .38 - .61; p<.05	436	150
SMQ	Chadwick et al. (2008)	1 (16)	α = .82 - .89	not reported	KIMS, MAAS, FMI: r = .38 - .61; p<.05	318	92
MMS	Haigh et al. (2011)	4 (21)	α = .45 - .86	not reported	not reported	50	22
SMS	Tanay & Bernstain (2013)	2 (23)	α = .95	r = .64 -.65	TMS, FFMQ: r = .31 -.47 MAAS: r = .00 -.07	13	2
CHIME	Bergomi et al. (2014)	8 (37)	α = .70-.90	.70-.90	FFMQ: r =.85	11	4

Note. *ICC – inter-class correlation coefficient; MAAS=Mindful Attention Awareness Scale; FFMQ=Five Facets Mindfulness Questionnaire; KIMS=Kentucky Inventory of Mindfulness Skills; TMS= the Toronto Mindfulness Scale; FMI= the Freiburg Mindfulness Inventory; CAMS-R= the Cognitive and Affective Mindfulness Scale-Revised; PHLMS= the Philadelphia Mindfulness Scale; SMQ= the Southampton Mindfulness Questionnaire; MMS= the Mindfulness/Mindlessness Scale; SMS= the State Mindfulness Scale; CHIME=Comprehensive Inventory of Mindfulness Experiences.

Of all mindfulness instruments, the Mindful Attention Awareness Scale (MAAS) (Brown & Ryan, 2003) is the most cited, and its psychometric properties are supported by a larger number of studies than for any other instrument (Park, Reilly-Spong, & Gross, 2013). The second most cited instrument, the Five Facets Mindfulness Questionnaire (FFMQ) (Baer et al., 2006), followed by the Kentucky Inventory of Mindfulness Skills (KIMS) (Baer et al., 2004) both proposed a multidimensional profile of mindfulness skills (Table 2).Temporal reliability (test-retest) has only been reported for the MAAS (Brown &

Ryan, 2003), the Kentucky Inventory of Mindfulness Skills (KIMS) (Baer et al., 2004), the State Mindfulness Scale (SMS) (Tanay & Bernstein, 2013) and the Comprehensive Inventory of Mindfulness Experiences (CHIME) (Bergomi, Tschacher & Kupper, 2014).

Test-retest reliability scores (Table 2) suggest that the MAAS, the CHIME and the KIMS (with the exception of its Observe subscale) are all trait measures (Ramanaiah et al. 1983; Spielberger, 1970, 1999). The SMS was proposed as a state measure (Tanay & Bernstein, 2013) , which is reflected by the expected test-retest score below .70 (Ramanaiah et al. 1983; Spielberger, 1999). Given that test-retest reliability is the only psychometric criteria currently used to distinguish between state and trait measures, we cannot be certain whether scales are measuring state or trait mindfulness.

The MAAS was constructed to assess attention and awareness to present-moment experiences, which may vary among individuals and can be developed as a result of practice (Brown & Ryan, 2003). This scale emphasises the presence or absence of awareness and attention in relation to the immediate experience of an individual. The MAAS is a 15-item self-report measure that uses a 6-point Likert-scale response format (1 = almost always to 6 = almost never). It is suitable to measure mindfulness in both clinical and general populations regardless of meditation experience. The MAAS has demonstrated good internal reliability and satisfactory external reliability (test-retest) over a four-week interval (Table 2). Unidimensionality of the MAAS was supported by a number of studies (Brown & Ryan, 2003; Carlson & Brown, 2005; Christopher, Charoensuk, Gilbert, Neary, & Pearce, 2009; MacKillop & Anderson, 2007). Convergent validity of the MAAS was tested by comparing it with other mindfulness measures including FMI, KIMS, CAMS-R, MMS and SMQ and showed positive correlations ranging from weak to moderate (Baer et al., 2006; Brown & Ryan, 2003; Christopher & Gilbert, 2010). Positive correlations found between the MAAS and measures of well-being, positive affect and openness and negative correlations with stress, anxiety, rumination and neuroticism support the construct validity of MAAS (Baer et al., 2006; Brown & Ryan, 2003; Carlson & Brown, 2005; Christopher & Gilbert, 2010; Frewen et al., 2008). Cordon & Finney (2008) found that experienced meditators score significantly higher on the MAAS compared to non-meditators, which is in line with expectations for a valid mindfulness measure.

The KIMS was constructed as a multi-dimensional self-report measure to assess specific traits or skills associated with mindfulness interventions introduced in DBT (Baer et al., 2004; Dimidjian & Linehan, 2003). KIMS includes 39 items, which were developed to capture the four fundamental mindfulness skills used in mindfulness-based treatment represented by four subscales labelled as Accept Non-Judgementally, Observe, Act With Awareness, and Describe (Linehan, 1993a; Segal et al., 2002). The items are presented in a 5-point Likert scale format with responses ranging from 'never' to 'almost always'. The Accept Non-Judgementally subscale assesses individual self-criticism and judging behaviour reflected in the common definitions of mindfulness (Table 1). The Observe subscale measures the degree of an individual's attention to present moment experiences. The Act With Awareness subscale assesses an individual's ability to be fully aware of any activities one is performing (e.g. driving, walking, conversing). Both subscales appear consistent with the common mindfulness definitions included in Table 1. Describe is a subscale that measures an individual's predisposition to describe their external and internal experiences. However, the 'describing' element is found in neither psychological definitions of mindfulness (Table 1); nor in traditional concepts of mindfulness (Dalai Lama et al., 2004; Gunaratana, 2002). Describing of internal and external experiences may have a therapeutic value but it is apparently unrelated to both psychological and traditional mindfulness concepts (Baer, 2003; Gunaratana, 2002) possibly because linguistic processing reduces an individual's capacity to pay attention to the present moment (Nickerson, 1978).

The total KIMS scale and its subscales have showed acceptable internal consistency and good external reliability for all but the Observe subscale (Table 2). The four factor KIMS structure has emerged from EFA because four factors explained together 43% of variance in the data and this structure was confirmed by CFA. However, the overarching mindfulness trait was not supported by CFA raising concerns about the validity of the total KIMS score (Baer et al., 2004). There is good supporting evidence for the construct validity for Accept Without Judgement and Act With Awareness facets, but poor support for Describe and Observe facets (Baer et al., 2004; Christopher & Gilbert, 2010; Frewen et al., 2008). The main limitations of KIMS include low correlations among subscales resulting in failure to provide the total mindfulness score and concerns related to the content validity (e.g. the Describe subscale). Also, the reported test-retest reliability of .65 for the Observe subscale (Table 2) would seem more indicative of a state, rather than a trait (Barker et al., 1976; Spielberger, 1999). Before the present work commenced, there were no studies that used methods other than CTT (e.g. IRT, Rasch, G Theory) to investigate psychometric properties of the KIMS.

The FFMQ was constructed from 112 combined items of the five scales including FMI, MAAS, KIMS, SMQ and CAMS-R using factor analysis (Baer et al., 2006). In total, there are five subscales: Four of them are similar to KIMS subscales including Non-judging Inner Experience, Act With Awareness, Observe and Describe and one new subscale called Non-reactivity to Inner Experience which was identified by EFA and confirmed by CFA (Baer et al., 2006). Psychometric properties of the FFMQ are represented in Table 2 showing good internal consistency but test-retest reliability was not reported. Similar to KIMS, FFMQ has 39 items derived from the initial item pool of 112 items meaning that a large number of items was excluded to fit the five-factor model, which was confirmed by CFA (Baer et al., 2006, 2008). The main limitations of the FFMQ seem to be an inability to assess both state and trait mindfulness and to produce an interpretable total mindfulness score. Similar to the KIMS, psychometric properties of the FFMQ were mainly investigated using more traditional (CTT) methods with the exception of two studies that used IRT methods to investigate differential item functioning (DIF) of the FFMQ items with meditator and non-meditator samples (Van Dam et al. 2009; Baer 2010). However, the reported DIF findings were contradictory and neither study provided a practical solution to improve the psychometric properties of the FFMQ meaning that further research is necessary using IRT and Rasch methods in particular.

The Toronto Mindfulness Scale (TMS) (Lau et al., 2006) is the first and the most cited instrument designed exclusively to assess state mindfulness. While developing the TMS, the authors defined mindfulness as a state-like quality (Bishop et al., 2006) having two components: “(a) the intentional self-regulation of attention to facilitate greater awareness of bodily sensations, thoughts, and emotions; and (b) a specific quality of attention characterised by endeavoring to connect with each object in one’s awareness (e.g., each bodily sensation, thought, or emotion) with curiosity, acceptance, and openness to experience. Such a state involves an active process of relating openly with one’s current experience by allowing current thoughts, feelings, and sensations.” (Hayes et al., 1999) (p. 1447). The TMS includes two subscales (Curiosity and Decentering) derived from exploratory factor analysis and supported by confirmatory factor analysis (Lau et al., 2006). Both subscales have demonstrated good internal consistency with Cronbach’s alpha ranging from .86 to .91 for Curiosity and .85 to .87 for Decentering (see Park et al., 2013). The total score is not reported due to modest correlation (*r* = .42) between the subscales, which supports a two-dimensional structure for the TMS (Lau et al., 2006).

The Decentering subscale of the TMS showed higher correlations *(r* range: .20 to .74) compared to the Curiosity subscale (*r* range: .10 to .54) (Davis et al., 2009) with other mindfulness measures including: the Freiburg Mindfulness Inventory (Walach et al., 2006), MAAS (Brown & Rayan, 2003), the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007), the Kentucky Inventory of Mindfulness Skills (Baer et al., 2004), the Five Facets Mindfulness Questionnaire (Baer et al., 2006) and the Southampton Mindfulness Questionnaire (Chadwick et al., 2008). Both TMS subscales showed positive correlations with the Reflection subscale of the Rumination-Reflection Questionnaire (Trapnell & Campbell, 1999), the Psychological Mindedness Scale (Conte et al. 1990), the Tellegen Absorption Scale (Tellegen, 1982) and the Surroundings subscale of the Situational Self-Awareness Scale (Govern & Marsch, 2001; Lau et al., 2006). As predicted by the authors, self-consciousness and internal states of awareness correlated significantly with the Curiosity subscale only (*r* = .31 - .41) and only the Decentering subscale correlated significantly with openness (*r* = .23) and cognitive failures (*r* = -.16) (Lau et al., 2006). Meditators scored higher on both TMS subscales compared to those without meditation experience, and Decentering scores were shown to reflect meditation experience (Davis et al., 2009) and changes in psychological symptoms (Lau et al., 2006). Both TMS subscales displayed increased scores after mindfulness training, which provide support for their construct validity. However, no test-retest reliability scores were reported, which is the only conventional psychometric criterion used to distinguish between state and trait scales. Therefore, the ability of the TMS to distinguish clearly between state and trait should be investigated using appropriate methods such GT.

Currently, there is no agreement on the dimensionality of mindfulness with the number of factors ranging from one to five for commonly used measures (Table 2). However, recent analysis of validated mindfulness measures identified an even wider range of aspects underpinning the construct (Bergomi, Tschacher, & Kupper, 2013). In an attempt to cover the eight mindfulness aspects identified across currently available mindfulness measures (Bergomi et al., 2013), the 37-item Comprehensive Inventory of Mindfulness Experiences (CHIME) (Bergomi, Tschacher, & Kupper, 2014) was developed in the German language. Accordingly, this measure includes eight subscales measuring Awareness of Internal Experiences, Awareness of External Experiences, Acting With Awareness, Accepting Nonjudgmental Attitude, Nonreactive Decentering, Openness to Experience, Awareness of Thoughts' Relativity, and Insightful Understanding. The total

CHIME scale and all subscales had demonstrated good internal consistency (α range .70 to .90) as well as adequate test-retest reliability (*r* range .70 to .90) in the initial validation study (Bergomi et al., 2014). The proposed eight-factor CHIME structure was confirmed with a different sample (*n*=202) (Bergomi et al., 2014). Strong correlations (*r*=.85) found between the CHIME and the FFMQ (Baer et al., 2006) total scores as well as between conceptually similar subscale scores (e.g., act with awareness, *r*=.63) support the construct validity of the CHIME. The CHIME total score correlates with measures of wellbeing (.40), depression (-.46), and anxiety (-.39) at a moderate level and in the expected directions (Bergomi et al., 2014). While the instrument has demonstrated acceptable psychometric properties according to classical test theory approaches, its ability to discriminate precisely across individual mindfulness levels has not yet been rigorously investigated with modern IRT approaches.

This literature review shows that the MAAS, the FFMQ, and the KIMS were the most frequently used, evaluated and cited scales by researchers (Table 2). Also, the CHIME represents a new and very promising multidimensional mindfulness measure, because it incorporates the most relevant aspects of mindfulness captured by the currently available measures (Bergomi et al., 2013, Park et al, 2013). However, these measures were developed and validated using more traditional CTT methodology, the limitations of which were clearly outlined earlier. In particular, the ability of these instruments to accurately discriminate between individual mindfulness levels and their internal construct validity have not been thoroughly examined using suitable methodology such as Rasch analysis. To date, only a limited number of Rasch analyses focused on mindfulness measures have been reported (Goh et al. 2015; Inchausti et al. 2014; Medvedev et al. 2016a,b; Sauer et al. 2013). While generally these studies communicated useful psychometric information about the measures, only three of these previous studies have published ordinal-to-interval transformation tables for the three widely used mindfulness measures: the MAAS (Medvedev et al., 2016a), the subscales of KIMS (Medvedev et al. 2016b), and the FFMQ (Medvedev et al., 2017a).

TMS is the most cited and first developed state measure of mindfulness. However, its ability to distinguish between state and trait mindfulness has not been tested using suitable methodology such as GT. Currently, the only psychometric criteria to distinguish between state and trait measures is test-retest reliability, which was only reported for four out of the eleven mindfulness measures included in the most recent review (Park et al., 2013). Consequently, development and application of reliable GT based methodology is necessary to evaluate the ability of TMS and other measures to distinguish clearly between state and trait mindfulness.

Outcome Measures for Mindfulness Research

In mindfulness research, the relationships between mindfulness and related outcomes (e.g. wellbeing, mood, anxiety) need to be thoroughly investigated using reliable and valid measurement tools. In particular, comparisons between mindfulness and outcomes using parametric statistical tests such as ANOVA require that all involved measures are at least at interval level of measurement (Allen & Yen, 1979; Stucki et al., 1996). Therefore, the psychometric properties of those outcome measures used in mindfulness research also require comparable enhancement. The present work selected three main areas implicated in mindfulness research, namely stress (Kabat-Zinn, 1982, 1990), well-being (Bennet & Dorje, 2015; Josefsson et al., 2014) and functional independence (in rehabilitation) (Siegert et al., 2016) and investigated one widely used measure from each of these three areas.

Stress has become a ubiquitous term used in both everyday language and scientific research to describe heightened emotional states associated with physiological changes that affect social and occupational functioning (McEwen & Stellar, 1993; Helton & Näswall, 2015; Balducci et al., 2015). From an evolutionary perspective, the stress response is adaptive as it provides an individual with the necessary biological resources to deal with a potentially life-threatening situation (Korte, Koolhaas, Wingfield, & McEwen, 2005). However, much of the research in health psychology has focused on the negative effects of stress that become more averse with extended exposure (Cohen et al., 1998; Hillhouse, Kiecolt-Glaser, & Glaser, 1991; Syvalahti, 1987).

Given the aversive effects of stress, mindfulness was first applied and widely used for stress reduction using the MBSR programme (Kabat-Zinn, 1982, 1990). To evaluate the effectiveness of such and similar programmes, an accurate assessment of both mindfulness and stress levels pre/post and during the intervention is required. Therefore, accurate measurement of stress has become an important research issue. In particular, precise assessment of perceived stress is critical because it reflects the subjective evaluation of environmental events (Bloch, Neeleman, & Aleamoni, 2004), which in turn influence physiological responses (LeDoux, 2000; Medvedev, Shepherd, & Hautus, 2015).

The Perceived Stress Scale (PSS) (Cohen & Williamson, 1988) is a widely used measure of perceived stress, approaching 12,000 citations by the beginning of 2017, according to Google Scholar. The PSS was first developed as a 14 item-scale (Cohen & Williamson, 1988) but four items displayed poor loadings on the first principal component (.11 - .39), and were removed leaving the popular PSS-10 version. The PSS-10 has been translated into 25 different languages and validated cross-culturally (Cohen, 2013). While the PSS-10 has demonstrated acceptable psychometric properties, its accuracy in discriminating between individual stress levels has not previously been rigorously investigated using modern IRT approaches and Rasch modelling in particular.

Evidence shows that MBIs increase psychological well-being (Bennet and Dorje, 2015; Josefsson et al., 2014) and various models of psychological well-being include trait mindfulness as a major predictor (Brown & Kasser, 2005; Pearson et al., 2015). Given the demonstrated contribution of MBIs to individuals' health and well-being, reliable evaluation of outcomes of MBIs requires an accurate assessment of pre- and post-treatment levels of both well-being and trait mindfulness (Visted et al. 2015). Thus, precise well-being and mindfulness instruments with robust psychometric properties are required for accurate assessment of psychological and cognitive changes in individuals undergoing MBIs. To increase reliability of comparisons between mindfulness and well-being measures both measures should be of interval level of measurement, which can be achieved using Rasch analysis (Rasch, 1961; Tennant & Conaghan, 2007).

Hills and Argyle (2002) considered limitations of earlier happiness measurements in constructing their Oxford Happiness Questionnaire (OHQ). The authors used terms such as well-being, subjective well-being, and psychological well-being as synonymous to happiness in describing the OHQ to cover various definitions of happiness. The scale is based on research findings indicating a single happiness dimension and contains items assessing positive and negative affect, and cognitive evaluations including life satisfaction and happy traits, which are the main components of subjective well-being (Andrews & McKennell, 1980; Argyle, 2001; Diener, 1984; Diener, Suh, Lucas, & Smith, 1999; Hills & Argyle, 2002). Also, there are items reflecting further cognitive components and traits found within the single happiness factor labeled as sociability, sense of control, physical fitness, positive cognition, mental alertness, self-esteem, cheerfulness, optimism and empathy (Hills & Argyle, 1998, 2002). This instrument is a new version of the Oxford Happiness Inventory (Argyle, 2001), and both scales were widely used in Oxford for assessment of personal happiness and shown to have satisfactory psychometric properties (Hills & Argyle, 2002). However, at the time of commencing this thesis, psychometric properties of this measure had not been tested using IRT and specifically Rasch methodology.

The OHQ is an ordinal scale, which technically is not suitable for parametric statistics and hence for comparative analyses with interval level data such as neurophysiological recordings and interval-transformed mindfulness scores. Psychometric properties of the OHQ can be, thus, enhanced up to an interval-level measure using Rasch analysis (Tennant & Conaghan, 2007; Hobart & Cano, 2009).

Application of mindfulness techniques in rehabilitation medicine are becoming increasingly popular, but their thorough evaluation requires accuracy of both functional levels and mindfulness assessment (Siegert et al., 2016). Therefore, both measures should have enhanced psychometric properties that produce an interval-level score, which again can be achieved through application of Rasch analysis (Tennant & Conaghan, 2007). The Functional Independence Measure (FIM) is one of the most popular outcome measures in rehabilitation world-wide, including 13 ‘motor’ and 5 ‘cognitive’ items (Keith, Granger, Hamilton & Sherwin, 1987; Hamilton, Granger, Sherwin, Zielezny & Tashman, 1987). The Functional Assessment Measure (FAM) was constructed in the United States to extend coverage of cognitive and psychosocial function of the FIM by adding additional 12 items (Hall, Hamilton, Gordon & Zasler, 1993). This extension was particularly important for use in patients with more complex disabilities following acquired brain injury. The 30-item UK version of this combined measure (UKFIM+FAM) was first published in 1999 (Turner-Stokes, Nyein, Turner-Stokes & Gatehouse, 1999). The UKFIM+FAM assesses physical, cognitive, communicative and psychosocial function. Therefore, this measure was selected as a suitable candidate for enhancement through application of Rasch analysis, which has potential benefit for both mindfulness research and rehabilitation medicine.

Conclusion

Evidence from clinical outcome studies suggests that mindfulness practice leads to psychological and cognitive changes with positive outcomes, which requires accurate measurement of both mindfulness and the related outcomes. The MAAS, KIMS, FFMQ and CHIME are the best-validated mindfulness measures to date, however, similar to other mindfulness and outcome measures, they mostly employ only an ordinal level of measurement, which does not satisfy fundamental assumptions of parametric statistical tests such as ANOVA and hence limits their application in research. Similarly, the widely used outcome measures of perceived stress (PSS-10), subjective well-being (OHQ) and functional assessment of dependency/independence (UK FIM+FAM) are ordinal measures that lack precision. Usage of these ordinal scales in modern mindfulness and health-outcome research may compromise the validity of comparisons with neurophysiological (e.g. EEG, heart rate) and biological (e.g. cortisol level) data. Therefore, it is necessary to investigate the psychometric properties of these scales in order to improve their precision up to an interval-level measure. Such an investigation can be conducted using Rasch analysis, a technique that is particularly suited for this purpose (Rasch, 1961; Masters, 1982). Moreover, Rasch analysis can be used to test the internal construct validity of these measures, which is important due to lack of consensus regarding the construct of some instruments (Cohen & Williamson, 1988; Taylor, 2015).

Accurate measurement of psychological and physiological changes of people undergoing MBIs is only possible when using reliable and valid mindfulness and outcome measures of both state and trait. Some of these changes manifest as traits on the level of normal (non-meditative) everyday functioning, while a relative or dynamic mindfulness level (state) may depend on the time of the day, physical conditions and environmental variables. For instance, measuring a state after a specific mindfulness exercise will indicate to what extent an exercise induces a state of elevated mindfulness. Also, both state and trait measurements of mindfulness are important before and after mindfulness-based treatment to evaluate the effect of both mindfulness level and specific mindfulness skills on a therapeutic outcome. For instance, state can be explained as more short-term experience (e.g. immediately after a session), whereas trait refers to a pattern established over the longer term (e.g. well after completion of a programme). Efficiency of any mindfulness-based treatment depends on expected long-term changes of trait mindfulness necessary to prevent a relapse. However, effectiveness of a specific mindfulness exercise/meditation can only be evaluated by measuring state changes.

Currently, distinction between state and trait scales is merely based on a single correlation between test scores at two different occasions (test-retest), which are expected to be lower for a state measure compared to a trait. Ultimately, identifying and comparing variance components contributed by person x occasion interaction (state) and by person (trait) in a given measure will give a more accurate estimation of whether it captures state or trait. Therefore, methodology for making a clear distinction between state and trait measures should be developed based on GT (Bloch & Norman, 2012), which is arguably the most suitable measurement theory for this purpose.

Taken together, this work applied Rasch analysis to improve psychometric properties of the four trait mindfulness measures and three related outcome measures and developed a GT based methodology to more accurately distinguish between state and trait measures. This methodology was applied to evaluate TMS – a state mindfulness measure, which was used as an example to illustrate this novel GT-based method.

Chapter Two. Rasch Analysis to Enhance the Psychometric Properties of Scales

Chapter One outlined the common limitations of ordinal measures such as lack of precision and incompatibility with parametric statistics and neurophysiological measures, which can all be addressed using Rasch analysis. This chapter describes the general methodology of applying Rasch analysis to evaluate and enhance the psychometric properties of mindfulness and outcome measures.

Rasch analysis can be performed using RUMM2030 software developed by Andrich, Sheridan & Luo (2009), although other packages such as Winsteps and “R” are also available (Linacre, 2011). This work used RUMM2030 because of its specific advantages such as interactive access to Rasch model fit statistics and graphs, and useful extensions for examining threshold parameters and invariance across individual items. Prior to Rasch analysis, data needs to be formatted using IBM SPSS v.23 and saved as an ASCII file to be imported into the software RUMM2030. The Likelihood-ratio test is computed first on the initial output of analysis for each measurement instrument. Rating scale (Andrich, 1978) is only used if differences between item thresholds are uniform across individual items while items can vary by their difficulty, which is an average of the item thresholds. However, if differences between thresholds vary significantly between items (p<.05), the unrestricted Partial Credit model will be used (Masters, 1982). Generally, Rasch analysis includes the following sequential steps (Siegert et al., 2010):

1. A test for overall data fit to the Rasch model.
2. Identifying items with disordered thresholds and rescoring them.
3. Deletion of items with poor fit to the Rasch model.
4. Re-testing individual item fits and overall fit to the Rasch model.
5. Analysis of Differential Item Functioning (DIF) for gender, age, sample population, and other personal factors.
6. Unidimensionality test.
7. Examination of local dependency based on the residual correlation matrix.
8. Distribution analysis of the participant-item thresholds.

9. Comparison of Rasch results with more traditional psychometric tests (e.g. item-to-total correlations and factor analysis).

Rasch analysis is an iterative procedure that is completed upon meeting the following criteria: Overall model and individual item fit are both satisfactory, and unidimensionality of the scale is clearly evident (Tennant & Conaghan, 2007).

Rasch Model Fit Criteria

The Rasch model fit is evaluated by the mean item and person locations, individual item fit residuals, the overall item-trait interaction and DIF using the following interpretations (Tennant and Conaghan, 2007; Gustafsson, 1980):

a) The item location mean is used as a base and set to zero.

b) A person location mean in the range from -0.50 to + 0.50 indicates a good coverage of a sample by a scale.

c) In case of an overall excellent fit, both item and person fit residuals values approximate 0.00 (SD=1.00).

d) Individual items fit residuals should be within the range of -2.50 to +2.50.

e) The overall and individual item-trait interaction chi-square should not be significant (*p* >.05).

f) No significant DIF should be evident by personal factors (e.g. gender, age).

Item-trait interaction(item e in the above list) is an index of consistency for the item parameters across a range of individual trait levels. Item-trait interaction is reflected by an overall and individual item chi-square fit statistic, which depends on sample size and number of class intervals and should be non-significant if data fit the model. It is assessed at the beginning of each analysis and after every modification of a scale (Tennant & Conaghan, 2007). Calculation of the chi-square fit statistic requires Bonferroni adjustment for the number of tests, and involves calculating the adjusted *p*-value by dividing the conventional *p*-value of .05 by the number of tests conducted. Only chi-square *p*-values below the adjusted *p*-value are considered as statistically significant.

Disordered Thresholds

Rasch analysis starts with testing the overall fit to the model. The threshold map of the software output is used to identify any items showing disordered thresholds. A threshold is disordered when participants' higher capacity on a construct (e.g. mindfulness) is not consistently reflected in progressively higher scores on the ordinal-scale response options for that specific item (Andrich, 1978). Figure 2 is an example taken from the study described in Chapter Four and shows item 29 of the KIMS (Likert scale from 0 to 4) with disordered thresholds (top panel) and ordered thresholds of the same item using the method of collapsing disordered categories (bottom panel).

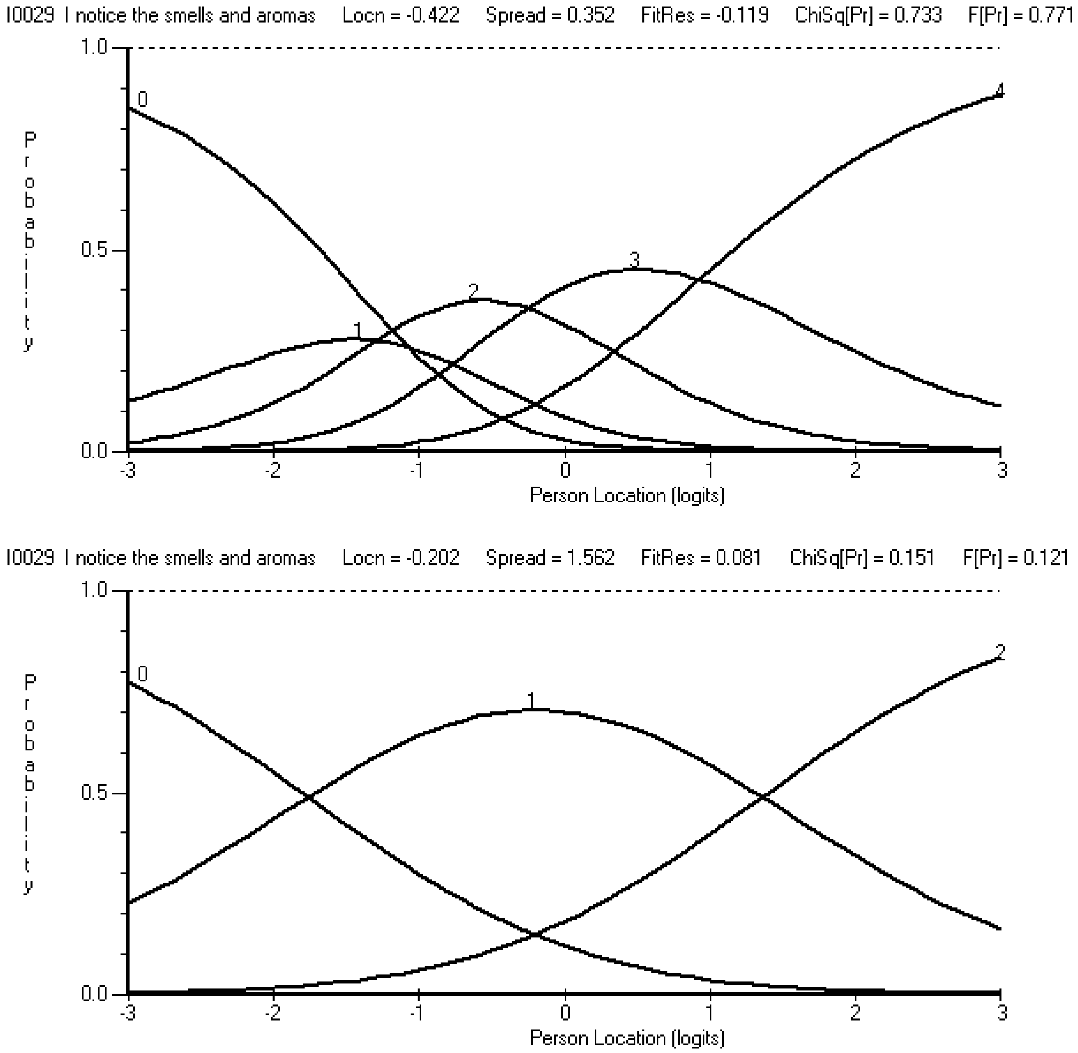


Figure 2. Item category probability curves illustrating disordered thresholds (top panel) and orderly thresholds after rescoring (bottom panel).

It can be seen on the top panel (Figure 2) that the probability to select response option 1 after 0 is lower than to select response option 2, which refers to a disordered threshold. In other words, if a latent trait level increases a participant would more likely select the response option 2 after 0 rather than 1. Disordered thresholds are usually corrected by collapsing relevant response options, which refers to rescoring of an item (Tennant & Conaghan, 2007). In the example (Figure 2) disordered threshold was corrected by rescoring options 0,1,2,3,4 (top panel) as 0,1,1,1,2 (bottom panel). In this case three response categories (1, 2 and 3) were collapsed because collapsing only two categories (1 and 2) did not result in expected order of thresholds. Usually, one or two items with disordered thresholds are re-scored at a time, and, at each step, goodness of fit to the model is re-tested. In the same way, after thresholds are satisfactorily adjusted, poorly fitting items are removed one at a time and the overall fit is re-calculated.

Differential Item Functioning (DIF)

DIF refers to the situation where participants with the same ability on the latent construct, but from different groups (e.g. males and females), respond differently to an item (Andrich & Hagquist, 2013). To investigate DIF in Rasch analysis, the sample is divided by class intervals according to different levels of the latent trait, and mean scores are calculated for each class interval separately for each sample sub-group (e.g. males versus females) under investigation. It is important that class intervals contain approximately equal-sized groups of participants, which may be difficult to achieve if a large number of class intervals is used. Therefore, the number of class intervals needs to be adjusted accordingly. DIF analysis is conducted by comparing the distributions of individual scores aggregated by class interval mean scores between groups of each person factor (e.g. age, gender, and ethnicity) and for each individual item using ANOVA. If the effect of a person factor is significant for an item, it is followed by visual examination of the relevant item characteristic curve (ICC) with class interval means for all groups plotted on the ICC. If mean differences are not consistent across observed class intervals (e.g. with at least one shared class interval point), DIF is considered as non-uniform. On the other hand, consistent differences refer to a uniform DIF (Andrich & Hagquist, 2013).

Figure 3 illustrates an example of uniform DIF by sample (students vs general population) from the FFMQ study described in Chapter Five, where participants drawn from the general population with the same level of the latent trait score consistently higher on this particular item. It can be seen that the sample is divided into four class intervals, and mean scores for each class interval are represented as ‘red crosses’ for the general population and as ‘blue circles’ for the student samples. If a uniform DIF for a specific personal factor is identified in one or more items, the item(s) concerned can be split into relevant categories, so that the same item(s) measure(s) different groups independently, without the need to delete it/item (Wainer & Kiely, 1987). This is done to obtain an unbiased measure, that ‘transcends’ group differences and satisfies the criteria of fundamental measurement (Thurstone, 1931).

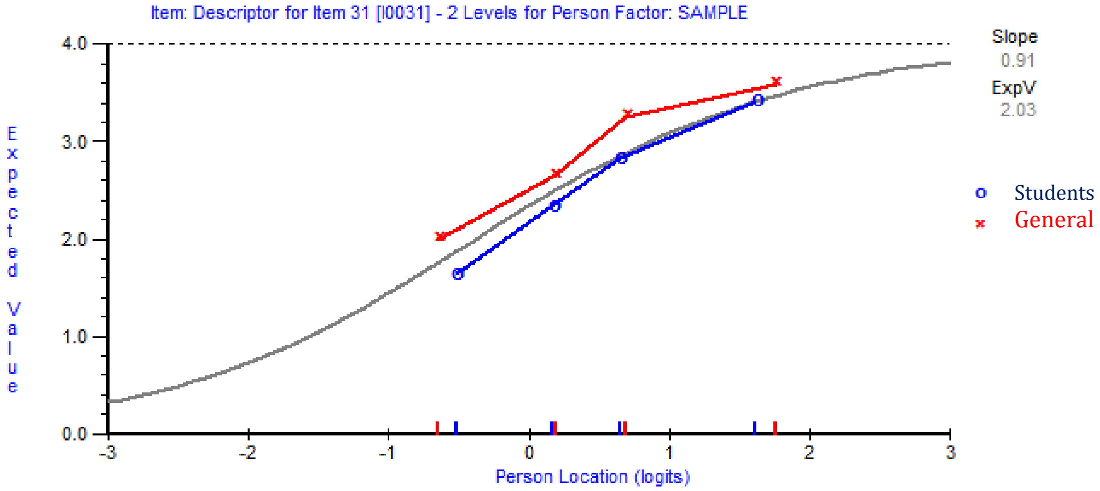


Figure 3. Example of uniform DIF in Rasch analysis of the FFMQ (Chapter Five).

Testing Unidimensionality

The Rasch model requires unidimensionality of a measure, which is normally tested using the method proposed by Smith (2000). This method employs an independent-samples *t*-test to compare person-estimates for two item groups with the highest positive and the highest negative factor loadings on the first principal component of the residuals, after the latent factor is removed. Unidimensionality is tested for each subscale individually because of multidimensionality and weak relationships between subscales of the KIMS. For instance, if a subscale has ten items then the items, with the highest positive factor loadings above .20 will form one set (e.g. three items) and the same number of items with the highest negative loadings (in this example, also three items) will form another set. Then, the estimates of each individual on these two sets of items will be compared by a paired-samples *t* tests. The percentage of significant *t* tests will be computed together with the +/- 95% binominal confidence interval. If the lower bound of the binominal confidence intervals computed for this percentage of significant *t*-tests is smaller than 5% (e.g. 4%), then, based on a statistical convention of alpha .05, we do not have any reason to believe that there is a real difference between the estimates, and unidimensionality is therefore accepted (Tennant & Pallant 2006).

Local Dependency

Both the overall and the individual item fit to the Rasch model could be affected by local dependency between items, which refers to a situation when two or more items are strongly associated in some way apart from the latent variable/trait of interest. For example, if one item in a questionnaire about negative emotions asked about whether a person had recently 'been upset' and another whether the person had recently 'been angered', these items might be locally dependent due to their shared meaning. Such a relationship violates the local independency assumption, compromises estimation of model parameters and inflates reliability (Wright, 1996).

There should be no local dependency evident between individual items in a subscale, which is examined using a residual correlation matrix. A residual correlation with a magnitude more than .20 compared to the mean of all residual correlations is regarded as a sign of local dependency (Christensen, Makransky, & Horton, 2016; Marais & Andrich, 2008). Instead of removing locally dependent items, these items can be simply added together into a subtest or testlet to solve local dependency issues (Wainer and Kiely, 1987; Lundgren-Nilsson et al., 2013). Using the example with negative emotions the subtest would be similar to one item measuring both aspects (get upset and angered). Subtests in Rasch analysis are analogous to item parcels in confirmatory factor analysis, and their advantages have been well documented elsewhere (Little et al., 2002; Rushton, Brainerd, & Pressley, 1983). Combined items show higher reliability compared to individual items and more scale points contributing to accuracy of measurement. Moreover, more accurate estimates of latent structures can be obtained using item parcels compared to individual items because combining items measuring the same construct reduces the amount of measurement error due to an individual item.

Reliability – Person Separation Index (PSI)

In Rasch analysis, reliability of subscales is estimated by the person separation index (PSI), which is not an index of Rasch model fit *per se* but indicates how accurately individuals are spread along the measurement construct as defined by the items (Fisher, 1992). PSI is similar to Cronbach's alpha numerically and measures how well the measure differentiates among people at different levels of the latent construct of interest. When distribution of persons is well covered by distribution of item thresholds and there are no extremes and missing values, then the two coefficients have comparable values (Fisher, 1992). Unlike Cronbach's alpha, PSI calculation involves non-linear transformation of the raw scores and can be performed with random missing data, which is an advantage of Rasch analysis.

Person-Item Threshold Distribution and Ordinal-to-Interval Conversion

When the basic criteria for fit to the Rasch model are satisfied, person abilities of the sample can be plotted against the items' thresholds on the same Rasch-derived interval-level scale. This plot is called a person-item threshold distribution and allows us to determine how well the range of individual abilities on a latent trait is covered by the range of item difficulties represented by individual item thresholds (Tennant & Conaghan, 2007). Figure 4 illustrates person-item threshold distributions for the Rasch analysis of the 30-item functional assessment measure UK FIM+FAM with a neurological sample of right stroke patients. It can be seen that above 90% of the sample functionality levels (top pannel) are well covered by item thresholds (bottom pannel) of the scale. However, there are signs of both floor and ceiling effects indicating that there are patients with a high level of disability on the right hand side of the graph and with a low level of disability on the left uncovered by the scale range.

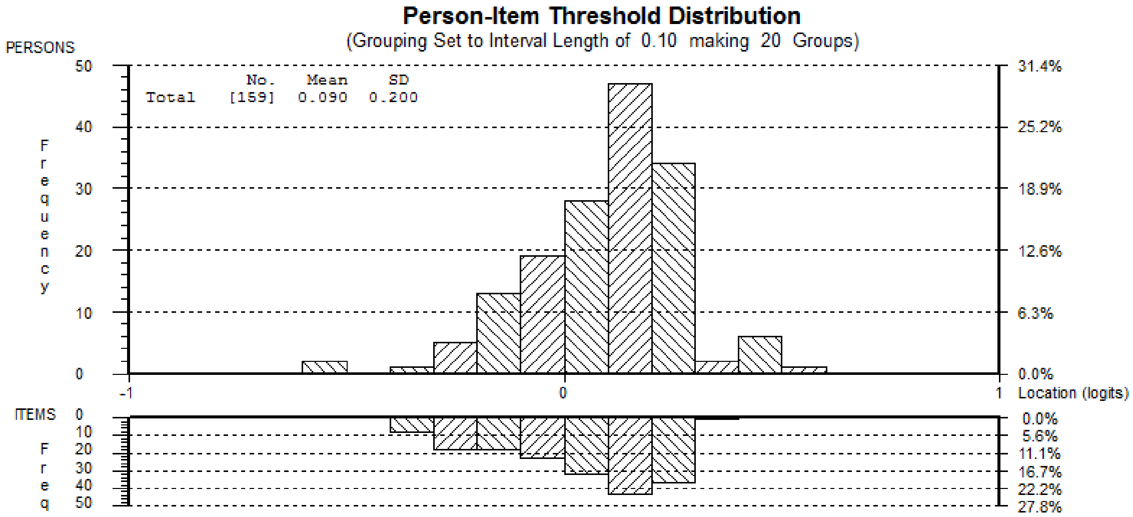


Figure 4. Person-item threshold distributions for the UK FIM+FAM right stroke patients. The distribution of persons is negatively skewed indicating the prevalence of patients with high levels of disabilities in this sample.

When items thresholds representing the latent construct adequately cover sample abilities ordinal-to-interval transformation scores can be computed that allow users to transform ordinal responses to an interval level data. Interval scores are originally estimated in logit units used as a universal metric for both person ability and item/threshold difficulty. If uniform DIF is found and one or more items are split for DIF the interval scores are computed individually for each sample group responding differently to these items (Wainer & Kiely, 1987).

Chapter Three. Improving The Precision of the MAAS using a Rasch model

Introduction

The MAAS is the widely used mindfulness scale, with validity studies indicating acceptable reliability and convergent validity. Perhaps the attractiveness of the MAAS (Brown & Ryan, 2003) is related to its simple unidimensional structure and relative brevity. The MAAS is a 15-item self-report questionnaire of trait mindfulness that uses a six-point Likert-scale response format (1 = "almost always" to 6 = "almost never"). A total score is calculated as the mean of responses to all items, with a higher score corresponding to a greater mindfulness level. Example items are: "I rush through activities without being really attentive to them" and "I find myself preoccupied with the future or the past". The MAAS is not fully consistent with mindfulness definitions used in psychology as it focuses on attention/awareness to the present moment but lacks items distinctly measuring a non-judgemental attitude (Kabat-Zinn, 1994; Bergomi, Tschacher & Kupper, 2013). However, based on the assumption that mindless states are more common (Brown & Ryan, 2003), MAAS items ask individuals about lack of mindfulness, which means that the instrument may serve as an indirect assessment of self-criticism (Bergomi et al., 2013).

Converging evidence supports good internal reliability and satisfactory external reliability over a four-week interval (Table 2). Tests of convergent validity of the MAAS by comparing it with a number of other mindfulness measures showed positive correlations in the range from weak to moderate (Baer et al., 2006; Brown & Ryan, 2003; Christopher & Gilbert, 2010). Construct validity of the MAAS was supported by its positive correlations with measures of positive affect, well-being and openness, and negative correlations with stress, anxiety, rumination, and neuroticism (Baer et al., 2006; Brown & Ryan, 2003; Carlson & Brown, 2005; Christopher & Gilbert, 2010; Frewen et al., 2008). Consistent with to expectations, significantly higher MAAS scores were found for experienced meditators compared to novices (Cordon & Finney, 2008).

Generally, evidence supports the proposed unidimensional structure of the MAAS (Brown & Ryan, 2003; Carlson & Brown, 2005; Christopher et al., 2009; MacKillop & Anderson, 2007). However, the factor loadings of some MAAS items (Items 5, 6, and 13) on the first principle component were occasionally reported to be below 0.30 (Table 3). Specific investigations of the performance of individual MAAS items also revealed some issues. In a study using Item Response Theory, Van Dam et al., (2010) used category response curves (CRC) to demonstrate functioning of individual MAAS items that included thresholds between pairs of adjacent response options for each item. CRC shows the probability of a person selecting a specific response category based on estimation of their latent trait (i.e., mindfulness). The findings showed that only 6 out of 15 items (Items 4, 7, 8, 9, 10, and 14) have equally distributed thresholds indicating that only these items are able to adequately discriminate between different mindfulness levels across the available response options. In addition, the relative contribution or ability of each item to assess a latent trait has been examined. Only five items (Items 7, 8, 9, 10, and 14) together explain about 66% of the information related to the latent variable mindfulness (Van Dam et al., 2010). Table 3 shows that items with lower loadings on the first principal component (e.g. Item 6) also explain a relatively lower amount of information related to the latent trait. These findings suggest that further research is necessary to investigate the functioning of individual items, which can be conducted using Rasch analysis, a technique that is particularly suited for this purpose (Tennant & Conaghan, 2007).

Table 3. *Loadings on the first principal component and the total amount of information explained by each MAAS item on the latent variable mindfulness.*

Item		**Range of factor loading**	***(%) Total Information**
1	experiencing some emotion and not be conscious of it	0.43 - 0.50	2.20
2	break or spill things because of carelessness, not paying attention	0.36 - 0.57	2.41
3	difficult to stay focused on what’s happening in the present	0.51 - 0.67	5.88
4	tend to walk quickly without paying attention to what I experience	0.41 - 0.62	3.84
5	tend not to notice feelings of physical tension or discomfort	0.27 - 0.51	4.63
6	forget a person’s name almost as soon as I’ve been told it	0.26 - 0.49	1.37
7	‘‘running on automatic,” without much awareness of what I’m doing	0.59 - 0.80	13.18
8	rush through activities without being really attentive to them	0.68 - 0.76	17.60
9	focused on the goal I want to achieve that I lose touch with what I’m doing	0.38 - 0.72	9.85
10	do jobs or tasks automatically, without being aware of what I’m doing	0.69 - 0.74	13.92
11	listening to someone with one ear, doing something else at the same time	0.45 - 0.56	3.01
12	drive places on ‘automatic pilot’ and then wonder why I went there	0.46 - 0.62	4.63
13	find myself preoccupied with the future or the past	0.28 - 0.54	2.42
14	find myself doing things without paying attention	0.71 - 0.78	11.77
15	snack without being aware that I’m eating.	0.36 - 0.62	3.29

Note: *The range of factor loading presented here is based on a systematic review by Park et al. (2013). Total information in % on the latent trait (mindfulness) measured by each MAAS item is based on the Item Response Theory analysis conducted by Van Dam et al. (2010).

The distinct advantages of Rasch analysis over classical psychometric methods have been discussed in detail in Chapters One and Two and also extensively argued elsewhere (Rasch, 1960; Wilson, 2005; Wright & Stone, 1979). Essentially, Rasch analysis provides a template to convert ordinal-level data to interval level, which improves precision of measurement, provided a measure is unidimensional (Rasch, 1961).

In summary, the MAAS is the commonly used mindfulness scale, perhaps to a large extent because it is a brief, well-validated instrument with good psychometric properties that can be applied to a wide range of clinical and non-clinical populations. However, recent evidence suggests that only a small subset of MAAS items adequately discriminate between mindfulness levels (Van Dam et al., 2010). Rasch analysis is a suitable method to investigate the performance of individual items to discriminate on their overarching construct, but to date Rasch analysis has only been applied to the Spanish version (Inchausti et al., 2013) and neither to an English-language version nor to a non-clinical sample. The aim of the present study is to apply Rasch analysis to explore and to improve the psychometric properties of the MAAS.

Method

Participants

The present study analysed data from 250 participants, based on the recommended optimal sample size estimates for the purposes of Rasch analysis (Linacre, 1994). For the present Rasch analysis, the sample included a randomly selected sub-set of 125 from a sample of New Zealand university students and a randomly selected sub-set of 125 participants from a New Zealand-wide postal survey to examine for DIF effects. Also, we aimed to make the results applicable to both students and general adult populations in line with the original validation study (Brown & Ryan, 2003). The total sample size of the university student sample was 253 (79.1% females, 19.0% males, 2.0% missing gender identification). Ages ranged from 18 to 59 with a mean age of 23.33 (SD=7.73). Ethnic groups include 51.8% Caucasians, 5.5% Māori, 7.1% Pasifika, 16.5% Asian, and 17.9% of other unspecified ethnicities. The total sample size of the national general population survey was 436, of whom 155 (35.6%) indicated that they were male and 280 (64.2%) that they were female. Ages ranged from 18 to 91, with a mean of 52.87 and a standard deviation of 17.05. The majority self-classified as Caucasian (81.9%), 8.5% as Māori, 2.8% as Pasifika, 2.8% as Asian, and the remainder as other ethnicities. After merging the two sub-sets of 125 respondents, the mean age was 38.20 years (SD=20.01). To investigate DIF, three age categories were created: 18-21 (*n*=86), 22-50 (*n*=83), and 51-88 (*n*=76). The number of males was 63 and number of females 185 (missing gender data

= 5), and those regular engaging in mindfulness practice were 84, as opposed to 154 not engaging in regular practice.

Procedure

The present study collected responses from university students as well as from the general population. For the student sample, potential participants were approached in class with permission of a paper coordinator and invited to complete the survey and to hand the survey back to the researchers or submit it to a locked collection box at their respective faculty. Students completed the questionnaire in class before the lecture or during the break. To obtain a sample from the general population, a questionnaire was posted to a sample of 4,000 individuals randomly selected from the national electoral roll. Respondents returned completed questionnaires using an enclosed self-addressed pre-paid return envelope. The response rate was 11%. The authors' institutional ethics committee approved this study.

Measures

The MAAS (Brown & Ryan, 2003) is a 15-item self-report questionnaire that has been described as measuring trait mindfulness (Siegling & Petrides, 2014). Sample questions include "I find it difficult to stay focused on what's happening in the present" and "I do jobs or tasks automatically, without being aware of what I'm doing" (Appendix C1). All items use a 6-point Likert-scale response format (1=almost always to 6=almost never). Internal consistency (Cronbach's alpha) in the current data set was .87.

Data Analysis

Descriptive statistics and reliability analyses of the MAAS were conducted using IBM SPSS v.22. Data were then formatted and saved as an ASCII file to be imported into the software RUMM2030 for Rasch analysis (Andrich et al., 2009). The likelihood-ratio test was conducted on the initial output analysis and should indicate appropriateness of the unrestricted (Partial-Credit) version of the model. Rasch analysis includes the sequential steps described elsewhere (Siegert et al., 2010) and outlined in Chapter Two.

Results

Preliminary test of the overall fit to the Rasch model

The person separation index (PSI) of .88 indicated good reliability. However, unsatisfactory overall fit to the model was evident ($\chi 2$ (45)=146.71, *p*<.001), and Items 2, 5, 12 and 15 displayed clearly disordered thresholds (Table 4, Test 1). Therefore, rescoring of the MAAS items was conducted prior to any further analyses.

Table 4. *Summary of fit statistics for the original MAAS version (1), after uniform item rescoring (2), after removing Items 6 and 15 (3), and after combining Items 7 and 8 into subtest (4).*

	Item residual		Person residual		Goodness of fit		PSI	Independent *t*-test	
Tests	**Value**	**SD**	**Value**	**SD**	**χ2 (df)**	***p***		**%**	**95% CI**
1	**0.44**	**2.26**	**-0.21**	**1.50**	**147 (45)**	**<.001**	**0.88**	**6.36**	**0.03 - 0.09**
2	**0.16**	**1.90**	**-0.29**	**1.42**	**118 (45)**	**<.001**	**0.86**	**7.20**	**0.04 - 0.10**
3	**0.11**	**1.38**	**-0.36**	**1.42**	**61 (39)**	**<.01**	**0.87**	**7.20**	**0.04 - 0.10**
4	**0.11**	**1.30**	**-0.36**	**1.38**	**47 (36)**	**.11**	**0.87**	**7.20**	**0.04 - 0.10**

Rescoring of MAAS items

Iterative rescoring of the MAAS items showed that optimal ordering of thresholds and goodness of fit could be achieved using uniform rescoring of all the 6-point Likert scale items by collapsing response Category 2 (very frequently) with 3 (somewhat frequently), and Category 4 (somewhat infrequently) with 5 (very infrequently). Figure 5 shows an example of the effect of rescoring Item 2 on the category response probability curves. All disordered thresholds became ordered after uniform rescoring, and the overall fit to the model was also improved, although still not at acceptable levels ($\chi 2$ (45)=118.14, *p*<.001, Table 4, Test 2). Therefore, it was decided to remove items with poorest fit (highest fit residuals) one at a time, with subsequent tests of fit to the model.

Fit residuals for all 15 items were analysed after uniform rescoring all items. Items 6 (forgetting names) and 15 (snacking without awareness) displayed the largest fit residuals and highest chi-square values, indicating poor fit to the Rasch model, and were removed before the analyses continued. Table 5 includes the chi-square values for all MAAS items from the initial test before rescoring (1) and after rescoring and removing non-fitting Items 6 and 15 (Test 3). Large chi-square values are associated with poor fit to the Rasch model. Also, Table 5 includes the location of each item in log units of probability, or logits, that indicates the relative difficulty of each item on the Rasch scale. Higher positive values signify difficult items (e.g. Item 13 preoccupied with future or past) meaning that few individuals obtain higher scores, and negative values correspond to less difficult items with more people having higher scores (e.g. Item 2 not paying attention).

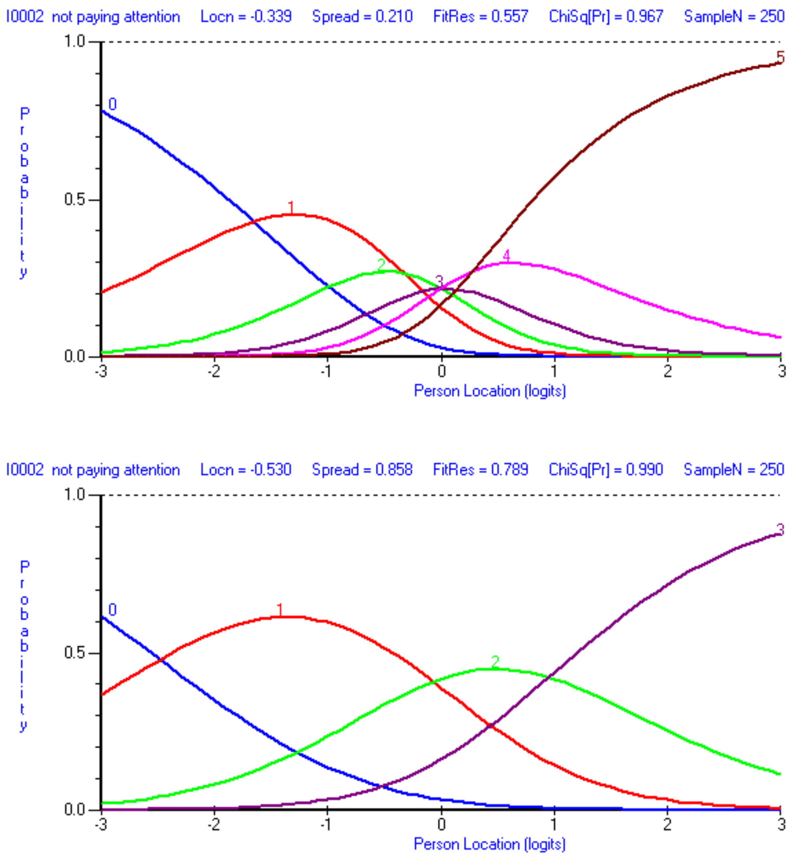


Figure 5. Item category probability curves for MAAS Item 2 before rescoring (top panel) and after rescoring (bottom panel).

Removing non-fitting items

Substantial improvement of fit was noted after removing Items 6 and 15, which both had the lowest loading on the first principal component and item-to-total correlations (Table 4). However, chi-square for overall person-item interaction was still significant (χ2 (39)=61.01, *p*<.01, Table 4, Test 3). At this stage, all the remaining 13 items had satisfactory model fit, with fit residuals below 2.50. Therefore, local dependency between items was investigated because it affects estimations of both discrimination parameters and test information.

Local dependency

The residual correlations between items were analysed, and the highest correlation was found between Items 7 and 8 (.28). Any residual correlations above .20 are generally considered as indicating local dependency between items. To confirm this observation, the correlation matrix between all items was also examined and showed the highest correlation between Items 7 and 8 (.66). Together, these observations confirmed local dependency, and Items 7 and 8 were therefore combined into a single testlet. This solution provided a desirable alternative to achieve a good fit to the Rasch model ($\chi 2$ (36)=46.79, p>.05, Table 4, Test 4) without excluding further MAAS items. This solution was replicated with the full sample (*n* = 689), showing identical issues with non-fitting items and local dependency.

Table 5. *Corrected item-to-total correlation and loadings on the first principal component for MAAS items together with Rasch model fit statistics: item location, fit residuals and chi-square from the initial analysis (1), and chi-square after rescoring and removing Items 6 and 15 (3).*

	Item	Item-to-total correlation	Item Loadings 1st PC	Item difficulty (location)	Item-fit residual	Chi-square (1)	Chi-square (3)
1	not conscious of emotions	0.42	0.50	-0.33	1.68	1.31	1.97
2	not paying attention	0.47	0.55	-0.34	0.56	0.26	1.55
3	difficult to focus on present	0.59	0.70	-0.14	-1.04	11.36	6.44
4	walk without paying attention	0.57	0.66	0.34	0.24	1.05	3.97
5	not notice physical tension	0.51	0.59	-0.07	0.27	0.36	5.44
6	forgetting names	**0.27**	**0.34**	0.69	**6.12**	48.16	-
7	running on automatic	0.70	0.77	0.11	-1.56	12.47	7.92
8	rush not attentive to activities	0.72	0.80	-0.11	-2.55	21.45	6.40
9	focused on goal achievement	0.60	0.67	0.07	-0.07	2.65	4.68
10	do tasks automatically	0.54	0.62	-0.17	0.01	1.86	0.79
11	listening and doing something	0.42	0.50	0.28	2.58	9.48	7.85
12	drive on 'automatic pilot'	0.58	0.66	-0.35	-1.51	4.57	0.99
13	preoccupied with future or past	0.52	0.59	0.38	1.24	1.47	2.76
14	doing things without attention	0.70	0.76	-0.16	-2.30	16.96	10.25
15	snack without awareness	**0.34**	**0.41**	-0.21	**2.86**	13.29	-

Test for unidimensionality

The set of person estimates from the three items with the highest positive loadings on the first principal component were compared with the set of estimates from the three items with the highest negative loadings. Out of 250 t-test comparisons between both sets calibrated to the same metric, 17 tests (6.8%) were significant. A binominal test was conducted to estimate the exact amount of acceptable deviations based on sample size. The calculated value of the binominal 95% confidence interval (CI) for the observed proportion overlapped 5% on the lower bound and thus confirmed unidimensionality of the current solution (Table 4).

Differential item functioning (DIF)

DIF was analysed by controlling for gender, age, sample (students versus general population), and practice factors. Significant DIF effect was found between students and general population responses to Items 3 (difficult to focus on present) (*F*(1,249)=11.29, *p*<.001) and 5 (not notice physical tension) (*F*(1,249)=12.25, *p*<.001), Bonferroni adjusted. However, graphical examination showed that for Item 3, the differences between samples were not consistent across observed confidence intervals. Therefore, only Item 5 was split for sample DIF resulting in the same item measuring each population independently. Also, the effect of age on DIF was significant for Item 5 (*F*(2,249)=7.90, *p*<.001), Bonferroni adjusted. However, graphical examination revealed that the respective observed means are not systematically different across observed confidence intervals for any of the age groups. No other significant effects on DIF were observed for other person factors.

Item-person threshold distribution

Figure 6 shows the person-item threshold distribution where person ability and item difficulty are plotted on the same logit scale for the final solution (Table 4, Test 4). Ability refers to mindfulness that is the latent trait measured by the MAAS.

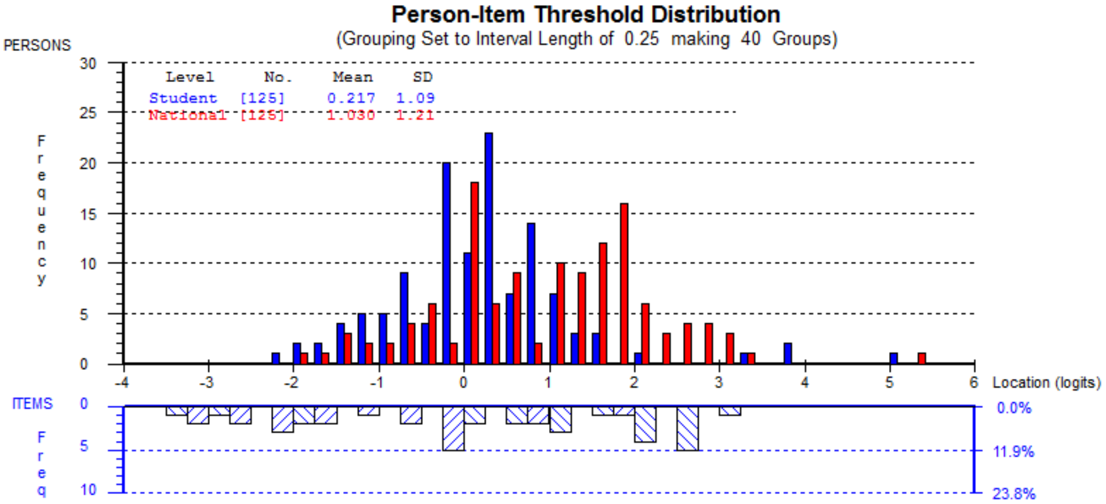


Figure 6. Person-item threshold distribution for the final 13-item MAAS solution (*n*=250).

Person-threshold distribution is close to normal, with evidence of a small ceiling effect indicating limited ability of the MAAS to discriminate between higher mindfulness levels. However, the item-threshold distribution satisfactorily covers most people's abilities for both students and the national sample on the latent trait, and there was no evidence of a floor effect.

Equating test

A paired-samples t-test was conducted to compare the means of person estimates from the full 15-item MAAS and the 13-item version. A significant difference was found between the person estimates of the two versions ($t(250)=1.96$, $p<.01$), indicating significant change in the ability of the 13-item version to discriminate between individual mindfulness levels compared to the original 15-item version. This confirms that the implemented modifications led to an improved solution for the MAAS.

Item-to-total correlations and loadings on the first principal component

Item-to-total correlations and loadings on the first principal component for all the original MAAS items were computed in IBM SPSS to permit a comparison with Rasch results and are included in Table 5. It shows that the excluded Items 6 and 15 have the lowest values for both parameters, confirming that these items are less consistent with the latent construct represented by the remaining items.

Ordinal-to-interval conversion table

Table 6 shows how raw scores can be converted from an ordinal to an interval scale. The raw scores shown here are after Items 6 and 15 have been removed and after response categories 2 and 3 as well as 4 and 5 have been merged. Researchers who have already used the MAAS to collect data can apply the results of this study as follows: Drop Items 6 and 15 and recode the response categories almost always as 0, very frequently and somewhat frequently as 1, somewhat infrequently and very infrequently as 2, and almost never as 3. Then, sum the 13 item responses (range of scores 0 to 39). Next, use Table 5 to convert these scores to means on a 1-to-6 scale similar to the original MAAS scoring system. By using the conversion table provided here, users are able to increase the reliability of the MAAS. Considering the above-reported DIF by sample, separate conversion tables are presented for use with student and with general population samples. Note that the ordinal-to-interval scale conversion proposed here does not require altering the response format of the scale, but only involves a different scoring algorithm. These conversion tables were replicated with the full sample size (n = 689), showing almost identical results.

Table 6. *Converting from a uniformly rescored 13-item MAAS raw score (0 to 39) to an interval scale in logit units and in mean scores.*

Raw score*	**Interval measure**				**Raw score***	**Interval measure**			
	Students sample		**National sample**			**Students sample**		**National sample**	
	Logit	**Mean**	**Logit**	**Mean**		**Logit**	**Mean**	**Logit**	**Mean**
0	-5.45	1.00	-5.42	1.00	20	0.13	3.61	0.19	3.63
1	-4.59	1.40	-4.55	1.40	21	0.30	3.69	0.36	3.71
2	-3.97	1.69	-3.94	1.69	22	0.47	3.77	0.54	3.79
3	-3.53	1.90	-3.49	1.90	23	0.64	3.85	0.71	3.87
4	-3.17	2.07	-3.13	2.07	24	0.81	3.94	0.88	3.95
5	-2.86	2.21	-2.82	2.22	25	0.99	4.02	1.06	4.03
6	-2.59	2.34	-2.55	2.34	26	1.16	4.10	1.23	4.12
7	-2.34	2.46	-2.30	2.46	27	1.34	4.18	1.41	4.20
8	-2.11	2.57	-2.07	2.57	28	1.53	4.27	1.59	4.28
9	-1.89	2.67	-1.86	2.67	29	1.71	4.36	1.78	4.37
10	-1.69	2.76	-1.65	2.76	30	1.91	4.45	1.97	4.46
11	-1.49	2.85	-1.45	2.86	31	2.11	4.54	2.17	4.55
12	-1.30	2.94	-1.26	2.95	32	2.32	4.64	2.38	4.65
13	-1.12	3.03	-1.07	3.04	33	2.55	4.75	2.60	4.76
14	-0.93	3.12	-0.89	3.12	34	2.80	4.87	2.85	4.87
15	-0.75	3.20	-0.70	3.21	35	3.08	5.00	3.12	5.00
16	-0.57	3.29	-0.52	3.29	36	3.40	5.15	3.45	5.15
17	-0.39	3.37	-0.34	3.38	37	3.81	5.34	3.85	5.34
18	-0.22	3.45	-0.16	3.46	38	4.39	5.61	4.43	5.61
19	-0.05	3.53	0.01	3.54	39	5.22	6.00	5.25	6.00

Note: The following uniform rescoring of response options for all 13 items is required before converting into an interval scale: 1 to 0; 2 to 1; 3 to 1; 4 to 2; 5 to 2; 6 to 3. The 13-item raw score is calculated as the sum of rescored values from all MAAS items except for Items 6 and 15.

Discussion

The MAAS (Brown & Ryan, 2003) is the widely used scale to measure trait mindfulness, despite the fact that its psychometric properties have not been fully clarified. Recently, Van Dam et al. (2010) reported results that challenged the ability of the MAAS to adequately discriminate between mindfulness levels, as only 6 out of the total 15 items had ordered thresholds, and five of these items represented approximately two thirds of the total information of the scale. The present Rasch analysis added to the limited number of studies that have investigated the performance of individual MAAS items in detail. While two items (Item 6 forgetting names and Item 15 snacking without awareness) had to be removed, the functioning of the remaining 13 items could be improved substantially by uniform rescoring. The psychometric properties of the MAAS following these adjustments are thus robust, and the precision of the scale can be further improved by using the ordinal-to-interval conversion algorithm in Table 6. This increased precision is not only desirable for studies that investigate the effects of clinical interventions on trait mindfulness but it also means that parametric statistics may now be legitimately used to analyse MAAS data.

Compared to Van Dam et al. (2010), who reported disordered thresholds for Items 1, 2, 3, 5, 6, 11, 12, 13, and 15, we only found clearly disordered thresholds for Items 2, 5, 12, and 15. However, the thresholds for Items 6, 11, and 13 were only marginally acceptable. For that reason, we decided to use uniform rescoring of all items. This improved the thresholds of all items and goodness of fit to the Rasch model. This solution also ensures that our proposed rescoring algorithm is easy to apply and suitable for users unfamiliar with Rasch analysis.

The removed items (forgetting names and snacking without awareness) were found to have item-to-total correlations and factor loadings that were clearly lower than those of other items and were also found to explain only a relatively small amount of information on the latent trait (Table 3). Possible reasons are that forgetting names may be more related to assessment of verbal memory and might only indirectly refer to mindfulness. This item may also be biased as the effort to remember a difficult (e.g. foreign) name might interfere with attention available to the present moment (Nickerson, 1978). The item “I snack without being aware that I’m eating” may not work well because it appears relatively unlikely that people are unaware of a whole episode of snacking, even though they may be absentminded during some periods during their snacking. Also, in a multicultural sample such as the present one, snacking habits may be very varied, and some people may prefer ‘eating’ to ‘snacking’, as the latter may be associated with unhealthy food. Future research may investigate to what extent re-wording of these items may improve the psychometric properties of the 15-item MAAS. However, until this work has been conducted, we recommend using the 13-item version with our proposed scoring algorithm.

Van Dam et al. (2010) argued that the ability of the MAAS to discriminate between mindfulness levels is impaired due to application of items measuring mindlessness because an individual without special training is not capable of accurately registering mindless states. However, Brown & Ryan (2003) insisted that mindlessness is more common among the general population and hence they should have the ability to adequately report it. The results of the present study support the construct validity of the MAAS in the 13-item format and suggest that items measuring lack of mindfulness might be adequate to reliably assess the construct if uniform item rescoring is applied. Moreover, the ordinal MAAS scale can be converted to an interval scale without changing the response format (Table 6), which accounts for DIF between students and general adult populations and provides interval level scores for each sample.

Limitations and Conclusions

The following limitations need to be acknowledged. Even though the sample reflects New Zealand's diversity of ethnic groups, no efforts were made to purposively sample underrepresented groups. The response rate of 11% for the national electoral roll sample was very low, which may reflect self-selection biases. However, such response rates are not uncommon for research of this nature in New Zealand (Hill, Billington, & Krägeloh, 2014; Krägeloh et al., 2013), and the above DIF analyses explored in detail any effects by demographic variables. Additionally, achieving a suitable fit to the Rasch model required rescoring of all items, which makes scoring the scale somewhat more complex. Nonetheless, converging from ordinal to an interval level scale can be conducted for both students and general population in logit units and in mean values using the same metric as the original scale for easy comparison. The readers are advised to refer to the present version of the MAAS as 'the 13-item version' to differentiate it from the original 15-item version.

Trait mindfulness has emerged as an important contributor to health and well-being, and its accurate measurement represents an-ongoing challenge. The current study used Rasch analysis to address previously reported limitations of the widely-used MAAS trait mindfulness instrument. We demonstrated that item functioning and precision of the MAAS can be enhanced to satisfy the expectations of a unidimensional Rasch model using uniform rescoring of item response categories. Two items significantly affected individual estimates and appeared less consistent with the latent trait. The precision of the MAAS can be optimized by discarding these two items and by using the ordinal-to-interval conversion tables published here.

Chapter Four. Rasch Analysis of the Kentucky Inventory of Mindfulness Skills

Introduction

The KIMS was developed as a multi-dimensional self-report measure of mindfulness-related skills introduced in the context of Dialectical Behavior Therapy (DBT) (Baer, Smith, & Allen, 2004). DBT treatment unifies mindfulness of non-judgmental observation derived from Zen Buddhism with Western contemplative traditions (Hayes, Follette, & Linehan, 2004). At the beginning of therapy, the goal is to develop individual skills of observing thoughts, emotions, and external stimuli by describing them. DBT emphasizes acting with awareness as a skill by cultivating it through a series of exercises that develop a routine of focusing attention on activities. Non-judgemental acceptance is also a primary skill that is recognized as part of the therapeutic process. To foster this skill, patients are encouraged to accept their reality and tolerate any unwanted feelings or thoughts without judgement (Linehan, 1993a, 1993b). It was proposed that mindfulness-related facets measured by the KIMS have utility in therapeutic contexts because they allows professionals to separate areas of skill development and, accordingly, assist individuals in strengthening specific skills (Baer et al., 2004).

It should be noted that unlike mindfulness-based stress reduction (Kabat-Zinn, 1982, 1990) and mindfulness-based cognitive therapy (Segal, Williams, & Teasdale, 2002), which emphasize the central role of mindfulness in the therapeutic process, DBT includes mindfulness as a sub-component among other treatment tools to increase sensory and perceptual awareness in normal, non-meditative circumstances (Hayes, Strosahl, & Wilson, 1999; Linehan, 1993a). For example, the Describe subscale would not be consistent with the most cited mindfulness definitions used in psychology to design mindfulness measures (Bishop et al., 2006; Kabat-Zinn, 1994; Segal et al., 2013), which limits comparisons of the KIMS with other mindfulness measures. The exception is the Five Facets Mindfulness Questionnaire (FFMQ) (Baer et al., 2006), which has some structural similarities with the KIMS and was constructed from the combined items of five mindfulness scales including the KIMS (Baer et al., 2004), the Mindful Attention and Awareness Scale (MAAS) (Brown & Ryan, 2003), the Freiburg Mindfulness Inventory (Buchheld et al., 2001; Walach et al., 2006), the Southampton Mindfulness Questionnaire (Chadwick et al., 2008), and the Cognitive and Affective Mindfulness Scale (Feldman et al., 2007). However, the FFMQ has five subscales compared to the four of KIMS and includes items that are not specifically designed for the measurement of mindfulness skills utilized in a DBT context.

The KIMS includes 39 items divided into four subscales: Accept Non-Judgementally (Accept), Observe, Act With Awareness (Act), and Describe (Baer et al., 2004). Accept is a subscale of the KIMS that measures the judging behaviour present in individuals, such as self-criticism. Observe is a subscale that measures the degree of attention an individual pays to both external events and internal emotions, sensations, and cognitions. Act assesses the individual's ability to be fully attentive to the present moment. Describe is a subscale that measures an individual's predisposition to describe or label their external and internal experiences. The items are presented in a 5-point Likert scale format, with responses ranging from 'Never or very rarely true' = 1 to 'Very often or always true' = 5. Examples of typical items reflecting the four skills include: "I tell myself I shouldn't be feeling the way I'm feeling" (Accept)), "I notice when my mood changes" (Observe), "I tend to do several things at once" (Act), and "I find words to describe my feelings" (Describe). Evidence indicates acceptable internal consistency for the total KIMS scale and all subscales, and good test-retest reliability for all subscales (r =.81 to .86), with the exception of Observe (r =.65). In an exploratory factor analysis, 43% of variance in the data was explained by four factors, which was interpreted as support for the four-factor model of the KIMS (Baer et al., 2004). Confirmatory factor analysis supported the four-factor model but failed to confirm an overarching second-order mindfulness factor (Baer et al., 2004; Baum et al., 2010).

Convergent and divergent validity of the total KIMS and its subscales Non-Judgementally and Act With Awareness were supported by positive correlations with self-compassion, openness, and emotional intelligence and negative correlations with mindlessness, neuroticism, and dissociation (Baer et al., 2006). However, these relationships appear less consistent for the Observe and the Describe subscales (Baer et al., 2004; Christopher & Gilbert, 2010; Frewen et al., 2008). Describing internal and external experiences may have a therapeutic value although it does not feature in most psychological definitions of mindfulness (Baer, 2003; Kabat-Zinn, 1994; Segal et al., 2013). The main limitations of the KIMS include relatively low correlations (ranging from 0.09 to 0.34) between subscales (Baer et al., 2004) and concerns related to the content validity (e.g. the Describe subscale) (Park et al., 2013).

To date, no reports are available about investigations into the psychometric properties of the KIMS using modern item-response theory and in particular the Rasch model. The ultimate goal of a Rasch analysis is conversion from ordinal-level data to interval level, which increases measurement precision and permits parametric statistical analyses without violation of their fundamental assumptions (Brogden, 1977; Rasch, 1961). Generally, only few reported studies so far have subjected mindfulness measures to Rasch analysis (Goh et al., 2015; Inchausti et al., 2014; Medvedev et al., 2016a; Sauer, Ziegler, Danay, Ives, & Kohls, 2013). Rasch analysis can be beneficial to improve the precision of the instrument, given its distinct advantages over classical psychometric methods which have been well argued elsewhere (Rasch, 1960; Wilson, 2005; Wright & Stone, 1979). Rasch analysis involves a unidimensional measurement model (Rasch, 1961) and in the case of the KIMS, Rasch analysis will be applied to each of the four KIMS subscales individually due to multidimensionality, low correlation between the subscales, and lack of support for an overarching mindfulness factor (Baer et al., 2004; Baum et al., 2010).

The KIMS (Baer et al., 2004) is a widely-used multidimensional measure of four mindfulness traits with generally accepted psychometric properties. The main purpose of the KIMS was to be used in mindfulness-based treatment and studies, and it is critical to establish precision of its subscales. However, the ability of the KIMS subscales to precisely discriminate between trait levels and the functioning of its 39 individual items has not been investigated rigorously. Rasch analysis is a suitable method to investigate the performance of individual items to discriminate on their overarching trait, but to date Rasch analysis has not been used to study the psychometric properties of the KIMS. The aim of this study is to apply Rasch analysis to investigate the psychometric properties of the KIMS and to explore strategies to improve precision and item functioning of its subscales.

Method

Participants

This study analyzed data from a sample of 287 New Zealand university students (78.7% females, 19.2% males, 2.1% missing gender). The mean age was 23.05 (SD=7.64), with ages ranging from 18 to 59. Ethnicities included 51.3% Caucasians, 8.7% Māori, 7.7% Pasifika, 19.7% Asians, and 12.2% of unspecified others. The sample size met recommended optimal sample size estimates for Rasch analysis (Linacre, 1994). To investigate DIF, two age categories were created: 18-20 (n=149) and 21-59 (n=127). There were 43 (15%) individuals regularly engaging in mindfulness practice, as opposed to 240 not engaging in regular practice, and 4 individuals with data missing. Therefore, DIF was tested for the person factors including gender, ethnic group, age, and engagement in meditation and relaxation practices, where meditation practice refers to regularly performing formal meditation exercises and relaxation practices refers to regular exercises such as yoga or progressive muscle relaxation.

Procedure

Potential participants were approached in lectures and invited to complete the survey and to hand the survey back to the researchers or submit it to a locked collection box at their respective faculty. Students completed the questionnaire in class before the lecture or during a break. The authors' university ethics committee approved this study.

Measures

The Kentucky Inventory of Mindfulness Skills (KIMS) is a 39-item self-report questionnaire developed to capture the four mindfulness skills acquired in DBT treatment including Accept, Observe, Act and Describe (Baer et al., 2004). There are 16 negatively worded items measuring absence of mindfulness skills including items 3, 4, 8, 11, 14, 16, 18, 20, 22, 24, 27, 28, 31, 32, 35 and 36. These items were reverse coded prior to statistical analysis.

Data Analysis

Prior to Rasch analysis, basic psychometric properties including reliability and factor structure of the KIMS were tested. Descriptive statistics, reliability analysis, and exploratory factor analysis were completed using IBM SPSS v.22, and Rasch analysis was completed using the software RUMM2030 (Andrich et al., 2009). Rasch analysis is a unidimensional measurement model that involves testing of unidimensionality along with other psychometric criteria. Therefore, unidimensionality of the full KIMS was tested by treating subscales as subtests in the Rasch model (Lundgren Nilsson et al., 2013). Unidimensionality of the KIMS subscales and their fit to the Rasch model were analyzed separately for each subscale including: Accept, Observe, Describe, and Act. First, the likelihood-ratio test was conducted on the initial analysis output for each subscale to confirm appropriateness of the unrestricted (Partial-Credit) version of the model. Rasch analyses followed ten main steps described elsewhere (Siegert et al., 2010) and outlined in Chapter Two.

Results

Exploratory factor analysis (EFA) using the principal axis factoring extracted ten factors with eigenvalues above 1.00. However, a large amount of variance in the data (44%) was explained by just four factors and supported by a clear cut-off point on the scree-plot, which was consistent with the original report (Baer et al., 2004). For that reason, the number of extracted factors was fixed to four. Applying Varimax rotation with Kaiser normalization yielded the factor loadings presented in Table 7, with items generally following the factor structure of the original study (Baer et al., 2004). The internal consistency of the full 39-item scale was satisfactory with Cronbach's alpha of .82. However, some individual item-to-total correlations for the full scale were low, ranging from -.19 to .50 (mean r = .29), with two items falling below 0.10 (Item 8, "I tend to evaluate whether my perceptions are right or wrong", r = -.19, and Item 19, "When I do things, I get totally wrapped up in them and don't think about anything else", r = .03). The low correlations (.10 to .30) found between the subscales were consistent with the original validation report (Baer et al., 2004) and provide additional evidence for multidimensionality of the KIMS. Attempts to fit the full scale to the Rasch model by combining items of each subscale into subtests using methodology of Lundgren Nilsson et al. (2013) was unsuccessful with clear signs of multidimensionality (>10% of significant t-tests) and the overall poor fit with significant item-trait interaction (p<.001).

Table 8 shows means and standard deviations for meditators and non-meditators together with Cronbach's alpha coefficients for each subscale of the KIMS. Cronbach's alpha for the subscales was in the acceptable range with the exception of the Act With Awareness subscale (α =.65). According to expectations, significantly higher mean values were observed for meditators compared to non-meditators, with the exception of Accept subscale as evidenced by subsequent *t* tests (Observe: *t* (276)= -4.49, *p* < .001; and also Describe: *t* (274)= -2.42, *p* = .016; Act: *t* (274)= -2.26, *p* = .024).

Table 7. *Initial item location, fit residual, corrected item-to-total correlation and factor loadings for Accept, Observe, Describe and Act subscale items of KIMS.*

N	Subscale/ Item	Item Location	Item Fit Residual	Item-Total	Factor Loading
	Accept				
4	I criticize myself for having irrational or inappropriate emotions.^{R}	0.29	-1.21	0.68	0.77
8	I tend to evaluate whether my perceptions are right or wrong.^{R}	0.46	8.08	0.21	0.27
12	I tell myself that I shouldn't be feeling the way I'm feeling.^{R}	0.09	-0.50	0.67	0.76
16	I believe some of my thoughts are abnormal or bad.^{R}	-0.20	-0.86	0.68	0.76
20	I make judgments about whether my thoughts are good or bad.^{R}	0.14	-1.27	0.70	0.77
24	I tend to make judgments about my experiences.^{R}	0.02	2.16	0.56	0.62
28	I tell myself that I shouldn't be thinking the way I'm thinking.^{R}	-0.17	-2.37	0.74	0.83
32	I think some of my emotions are bad or inappropriate.^{R}	-0.42	-2.47	0.73	0.82
36	I disapprove of myself when I have irrational ideas.^{R}	-0.21	-0.45	0.68	0.74
	Observe				
1	I notice changes in my body, breathing slows down or speeds up.	-0.22	1.10	0.42	0.49
5	I pay attention to whether my muscles are tense or relaxed.	0.62	0.07	0.49	0.53
9	When I'm walking, I deliberately notice the sensations	0.72	0.60	0.47	0.60
13	When I take a shower or a bath, I stay alert to the sensations.	0.17	1.23	0.45	0.56
17	I notice how foods/drinks affect my thoughts and emotions.	0.19	0.87	0.43	0.50
21	I pay attention to sensations, such as the wind in my hair.	0.11	-1.37	0.58	0.71
25	I pay attention to sounds, such as clocks ticking, birds chirping	0.07	1.06	0.45	0.61
29	I notice the smells and aromas of things.	-0.42	-0.12	0.49	0.63
30	I intentionally stay aware of my feelings.	-0.10	-0.63	0.54	0.57
33	I notice visual elements in art or nature, such as colors, shapes	-0.22	1.53	0.43	0.54
37	I pay attention to how emotions affect my thoughts and behavior.	-0.43	0.12	0.48	0.49
39	I notice when my moods begin to change.	-0.50	1.75	0.37	0.37
	Describe				
2	I'm good at finding the words to describe my feelings.	-0.04	-1.33	0.65	0.76
6	I can easily put my beliefs, opinions, and expectations into words.	-0.41	-1.34	0.68	0.73
10	I'm good at thinking of words to express my perceptions	0.18	1.36	0.56	0.55
14	It's hard for me to find the words to describe what I'm thinking.^{R}	-0.01	-2.24	0.69	0.78
18	I have trouble thinking of the right words to express how I feel.^{R}	-0.02	-2.16	0.72	0.79
22	When I have a sensation in my body, it's difficult to describe it.^{R}	-0.42	2.25	0.44	0.59
26	Even when I'm feeling terribly upset, I can put it into words.	0.37	3.64	0.51	0.60
34	My natural tendency is to put my experiences into words.	0.36	3.10	0.48	0.56
	Act				
3	When I do things, my mind wanders off and I'm distracted.^{R}	0.22	-0.77	0.43	0.54
7	When I'm doing something, I'm only focused on what I'm doing.	-0.12	-0.15	0.40	0.55
11	I drive on "automatic pilot" without paying attention.^{R}	-0.48	1.81	0.24	0.34
15	When I'm reading, I focus all my attention on what I'm reading.	-0.60	1.23	0.28	0.47
19	When I do things, I get totally wrapped up in them	-0.20	1.56	0.16	0.44
23	I don't pay attention what I'm doing because I'm daydreaming ^{R}	-0.12	-0.27	0.39	0.37
27	When I'm doing chores, I tend to daydream or think ^{R}	0.59	0.93	0.24	0.42
31	I tend to do several things at once rather than focusing on one ^{R}	0.21	0.42	0.31	0.49
35	When I'm working on something, part of my mind is occupied ^{R}	0.52	-1.64	0.49	0.62
38	I get completely absorbed in what I'm doing	-0.02	0.70	0.26	0.47

Note: $^{\mathbf{R}}$ reverse-scored item.

The initial model fit statistics for the 9-item Accept subscale are presented in Table 9. PSI of .88 confirmed satisfactory reliability of the subscale, and none of the items displayed disordered thresholds. However, the overall model fit was poor (χ^2(36)=157.43, *p* <.001), and Item 8 displayed an extremely high fit residual of 8.08, well above the 2.50 cut-off point (Table 7). Table 7 shows the location or difficulty of each item on the Rasch scale in probability units or logits. Deletion of item 8 resulted in a satisfactory overall model fit (χ^2(40)=45.43, *p* >.05) and acceptable reliability (PSI= .89) of the subscale.

Table 8. *Means and standard deviations (SD) for meditators and non-meditators, and Cronbach's alpha coefficients for the KIMS subscales.*

Subscale	Meditators (n=42)		Non-Meditators (n=221)		Cronbach's alpha
	Mean	SD	Mean	SD	
Accept	25.40	7.74	25.70	7.37	0.88
Observe	43.19	6.80	37.76	7.31	0.82
Describe	26.90	4.91	26.44	5.77	0.85
Act	29.29	5.60	27.42	4.63	0.65

Table 9. *Summary of fit statistics for the initial and the final Rasch analyses of the four KIMS subscales.*

	Item fit residual		Person fit residual		Goodness of fit		PSI	Independent *t*-test
Analyses	Value / SD		Value / SD		χ2 (df)	*p*		%LB^a
Accept								
Initial	0.12	3.27	-0.44	1.54	157.43 (36)	< 0.001	0.88	6.20
Final	0.31	1.21	-0.53	1.38	20.97 (25)	0.690	0.85	4.40
Observe								
Initial	0.52	0.93	-0.29	1.40	63.02 (48)	0.070	0.82	7.20
Final	0.55	0.62	-0.28	1.30	40.50 (40)	0.450	0.80	4.80
Describe								
Initial	0.41	2.44	-0.50	1.60	54.37 (32)	0.008	0.85	7.60
Final	0.62	1.62	-0.50	1.36	22.52 (20)	0.310	0.78	3.40
Act								
Initial	0.38	1.09	-0.38	1.44	67.39 (40)	0.004	0.67	8.30
Final	0.45	0.85	-0.38	1.36	38.40 (32)	0.202	0.60	4.40

Note: ^{a}LB = lower bound of the 95-% confidence interval.

Following the deletion of Item 8, Item 24 also exhibited an unacceptably high fit-residual of 3.64 and was therefore also removed before the analysis continued. The overall model fit improved after the deletion of item 24 (χ^2(35)=34.89, *p* =.47), and no other misfitting items were identified. At this stage, the residual correlation matrix was examined, and local dependencies were found between items 4 and 12, and between items 16 and 32, as evidenced by residual correlations exceeding the .20 limit above the mean of all residual correlations. After combining two pairs of locally dependent items into two subtests, the overall good fit to the Rasch model was further improved ($\chi^2(25)=20.97$, *p*=.69, Table 9, Accept, Final). At this stage, all individual items had acceptable fit to the model, and no other locally dependent items could be identified.

To test unidimensionality, the set of person estimates from the items with the highest positive loadings on the first principal component were compared with the set of estimates from the items with the highest negative loadings by an independent-samples t test. After calibrating t tests between both sets of estimates to the same metric, 20 t test comparisons out of 287 (6.97%) were significant. A binominal test was used to calculate the precise amount of acceptable deviations for the current sample. Unidimensionality of the final solution was confirmed by the overlap of the 5 % cutoff point on the lower bound surrounding *t* test (Table 9, 'Accept, Final). No DIF was found for person factors including gender, ethnic group, age, and engagement in meditation and relaxation practices.

The person-item threshold distribution plot for the Accept Non-Judgementally subscale (Final Analysis) is presented in Figure 8. The plot represents the relationship between distribution of item difficulty and person ability on the latent trait (e.g. Accept) converted to the same metric in logit units. Distribution of person thresholds is close to normal with some signs of ceiling and floor effects. However, over 90% of the sample were adequately covered by the items of the modified subscale.

Initial analysis conducted for the 12-item Observe subscale yielded acceptable chi-square ($\chi^2(48)=63.02$, *p* =.07) and reliability (PSI = .82) values. However, *t*-test comparisons between two sets of estimates with highest and lowest loadings on the first principal component after removing the latent trait component failed to confirm unidimensionality, with 8.71% of significant *t* tests and lower bound overlap above 5% (Table 9). Also, Item 29 displayed disordered thresholds and needed to be rescored before the analysis continued. After collapsing response options "Never or very rarely true" and "Rarely true", and "Sometimes true" and "Often true", thresholds of item 29 were precisely ordered. The overall model fit was slightly improved after rescoring Item 29 ($\chi^2(48)=62.12$, *p*=.09), and all individual items showed a good fit to the Rasch model (Table 9, Observe). However, unidimensionality of the subscale was not confirmed. Unidimensionality of a scale can be compromised by locally dependent items. Therefore, the residual correlation matrix was examined, and local dependency was found between items 21 and 25, and items 30 and 37, which were then combined into two subtests. This minor modification produced the final solution for this subscale, with overall good fit to the model (χ^2(40)=40.50, *p* =.45) and acceptable reliability (PSI = .80) (Table 9). A binominal test to test unidimensionality of the final solution indicated overlap on the lower bound surrounding *t*-test with the 5 % cutoff point, which confirmed unidimensionality (Table 9, Observe, Final). No DIF was noted for personal factors, such as gender, ethnic group, age, and meditation and relaxation practices. Figure 7 shows the item-person threshold distribution for the final solution of the Observe subscale. Overall, person thresholds are distributed close to normal and well targeted by item threshholds, but there are some signs of a small ceiling effect. In this analysis, a good fit of the Observe subscale to the Rasch model was achieved with the minor modifications of rescoring one item and creating two subtests, without a need to remove any misfitting items.

Initial analysis of the 8-item Describe subscale indicated acceptable reliability (PSI = .85) but an overall lack of fit to the model (χ^2(32)=54.40, p=.008) and lack of evidence for unidimensionality (Table 9). At this stage no items displayed unacceptably high fit residuals. Therefore, the residual correlation matrix was examined, indicating local dependency between items 14, 18, and 22, which were then combined into a subtest before analysis continued. After creating the subtest, the chi-square had a lower but still significant value (χ^2(24)=39.12, p =.03). However, at the individual item level, Item 6 displayed a high fit residual of 2.81, above the 2.50 cut-off point and was removed, which resulted in a good overall model fit (χ^2(20)=22.52, p=.31) and continued acceptable reliability (PSI = .78) (Table 9, Describe, Final). At the individual item level, all items showed acceptable fit to the model. Unidimensionality of the final solution was confirmed by the binominal test indicating overlap on the lower bound surrounding t test with the 5 % cut-off point. No significant DIFs were found for personal factors.

The person-item threshold distribution for the final analysis of the Describe subscale shows acceptable targeting of the person locations by the item thresholds (Figure 7). However, a slight ceiling effect was apparent indicating some limitation of the subscale in measuring higher personal abilities on Describe. Thus, satisfactory fit to the model was evident after few modifications that involved combining locally dependent items into a subtest and removal of one non-fitting item.

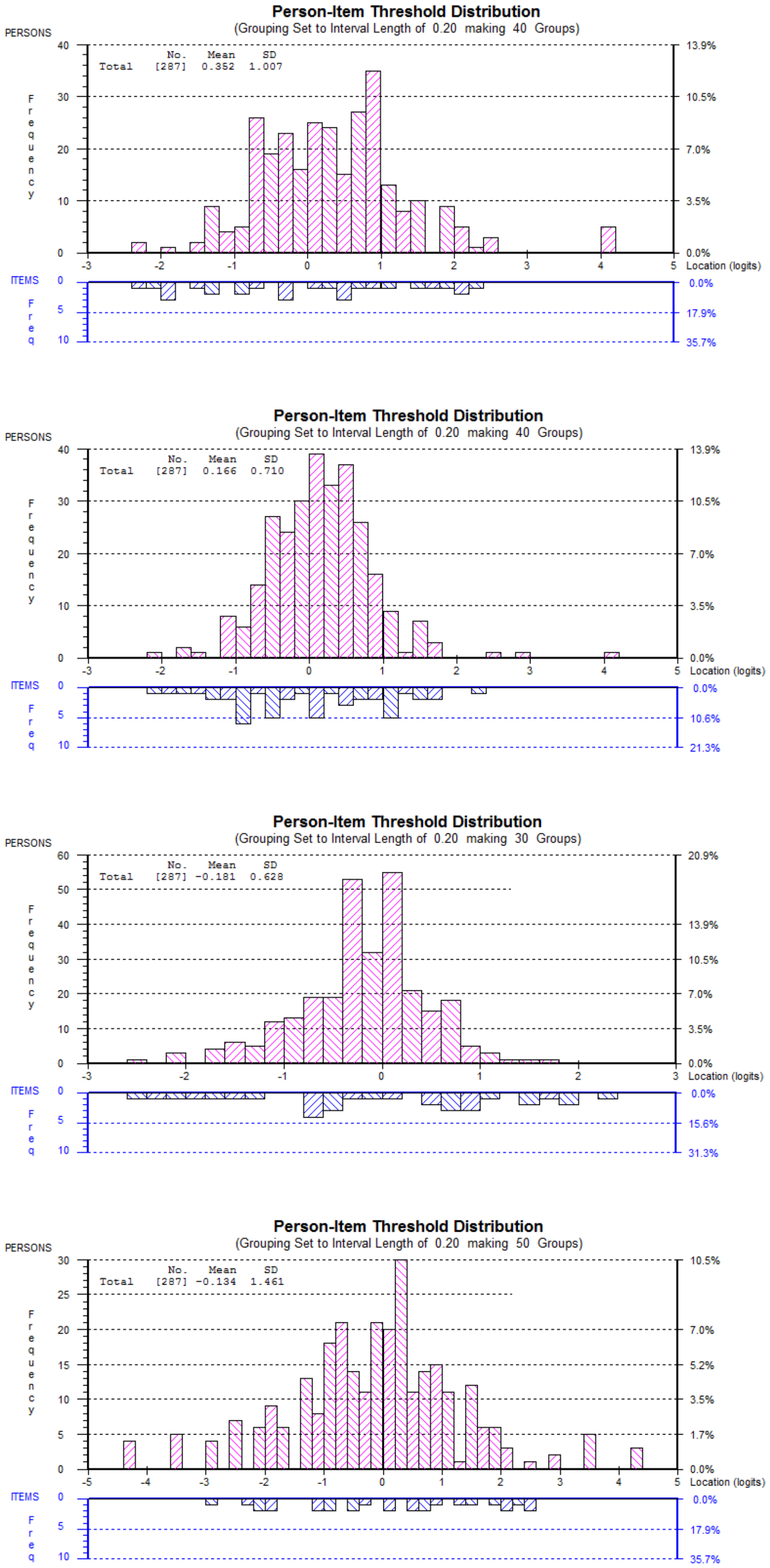


Figure 7. Person-item threshold distribution for modified KIMS subscales from top to bottom including Accept, Observe, Describe and Act (*n*=287).

Initial testing of the 10-item Act subscale of the KIMS revealed an overall lack of fit to the model with a significant chi-square for overall person-trait interaction ($\chi^2(40)=67.39$, *p*=.004) and reliability (PSI) of .66 (Table 9, Act, Initial). The assumption of unidimensionality was violated as indicated by the binominal test not overlapping the 5 % cut-off point on the lower bound. Consequently, individual item fit statistics were examined, indicating that Item 35 had a significantly high fit residual ($\chi^2(4)=15.75$, *p* <.001). Item 35 was thus deleted, resulting in acceptable overall model fit, with a non-significant chi-square value ($\chi^2(36)=41.20$, *p*=.25) but a slight reduction in reliability (PSI=.61). Additionally, evidence for unidimensionality was lacking as the binominal test indicated no overlap with the 5% cut-off point on the lower bound.

To test for local dependency, the residual correlation matrix was examined showing that Item 19 had high residual correlations with items 7 and 38 that exceeded the .20 cut-off point above the mean of all residual correlations. Item 19 also had the lowest item-to-total correlation of .19 (Table 9) and so was removed resulting in the final solution with an acceptable chi-square value ($\chi^2(4)=38.40$, *p* =.202) and evidence for unidimensionality (Table 9, Act, Final). However, the reliability of the Act subscale decreased slightly after these modifications (PSI=.60). At this stage, all individual items were showing good fit to the model, and no DIF was identified for any personal factors. The bottom panel of Figure 7 shows the item-person thresholds distribution for the final solution of the Act subscale, with fairly good coverage of the range of individuals' locations by the subscale items thresholds. Thus, the Act subscale modified by removing two misfitting items satisfied all but reliability criteria for fitness to the Rasch model.

Table 10 includes conversion scores from an ordinal-to-interval level scale for all four subscales of the KIMS. For convenience, all the scores are adjusted to the scoring algorithm of each original subscale, and the total ordinal score is calculated by adding the scores of all the individual items included in a final version of a subscale after negatively worded items are reverse coded. Also, item 29 has to be rescored according to the algorithm at the bottom of Table 10, before calculating the total score for the Observe subscale. Therefore, ordinal scores are represented on the left-hand side and corresponding Rasch interval-level scores on the right-hand side (Table 10). The conversion table provided here allows users to increase precision of the KIMS subscales without the need to modify the original response format of the scale. These conversions can only be used when there are no missing data.

Table 10. *Converting from ordinal to interval level scores for the subscales of the 34-item KIMS version.*

Accept Scores		Observe Scores				Describe Scores		Act Scores	
Ordinal	Interval	Ordinal	Interval	Ordinal	Interval	Ordinal	Interval	Ordinal	Interval
7	**7.00**	**10**	**10.00**	**43**	**37.67**	**7**	**7.00**	**8**	**8.00**
8	**9.72**	**11**	**14.49**	**44**	**38.22**	**8**	**9.84**	**9**	**11.19**
9	**11.65**	**12**	**17.46**	**45**	**38.80**	**9**	**11.72**	**10**	**13.43**
10	**13.02**	**13**	**19.43**	**46**	**39.40**	**10**	**12.96**	**11**	**15.01**
11	**14.12**	**14**	**20.92**	**47**	**40.04**	**11**	**13.90**	**12**	**16.25**
12	**15.06**	**15**	**22.11**	**48**	**40.74**	**12**	**14.70**	**13**	**17.29**
13	**15.90**	**16**	**23.12**	**49**	**41.48**	**13**	**15.41**	**14**	**18.21**
14	**16.66**	**17**	**24.00**	**50**	**42.31**	**14**	**16.07**	**15**	**19.02**
15	**17.38**	**18**	**24.79**	**51**	**43.25**	**15**	**16.72**	**16**	**19.77**
16	**18.06**	**19**	**25.50**	**52**	**44.32**	**16**	**17.36**	**17**	**20.46**
17	**18.71**	**20**	**26.17**	**53**	**45.59**	**17**	**17.99**	**18**	**21.11**
18	**19.34**	**21**	**26.78**	**54**	**47.17**	**18**	**18.61**	**19**	**21.73**
19	**19.96**	**22**	**27.37**	**55**	**49.24**	**19**	**19.24**	**20**	**22.32**
20	**20.57**	**23**	**27.92**	**56**	**52.36**	**20**	**19.86**	**21**	**22.89**
21	**21.18**	**24**	**28.45**	**57**	**57.00**	**21**	**20.48**	**22**	**23.45**
22	**21.79**	**25**	**28.97**			**22**	**21.10**	**23**	**23.99**
23	**22.41**	**26**	**29.47**			**23**	**21.72**	**24**	**24.53**
24	**23.03**	**27**	**29.97**			**24**	**22.34**	**25**	**25.06**
25	**23.66**	**28**	**30.44**			**25**	**22.96**	**26**	**25.60**
26	**24.30**	**29**	**30.92**			**26**	**23.58**	**27**	**26.13**
27	**24.96**	**30**	**31.40**			**27**	**24.21**	**28**	**26.68**
28	**25.66**	**31**	**31.87**			**28**	**24.85**	**29**	**27.24**
29	**26.39**	**32**	**32.33**			**29**	**25.53**	**30**	**27.81**
30	**27.19**	**33**	**32.80**			**30**	**26.29**	**31**	**28.41**
31	**28.08**	**34**	**33.27**			**31**	**27.17**	**32**	**29.05**
32	**29.13**	**35**	**33.73**			**32**	**28.26**	**33**	**29.73**
33	**30.44**	**36**	**34.20**			**33**	**29.71**	**34**	**30.47**
34	**32.32**	**37**	**34.68**			**34**	**31.85**	**35**	**31.30**
35	**35.00**	**38**	**35.16**			**35**	**35.00**	**36**	**32.25**
		39	**35.64**					**37**	**33.39**
		40	**36.13**					**38**	**34.84**
		41	**36.63**					**39**	**36.95**
		42	**37.14**					**40**	**40.00**

Note: Item 29 from the *Observe* subscale needs to be rescored before calculating the ordinal scores for this subscale as follows: 1=1, 2=1, 3=2, 4=2, 5=3. Negatively worded items 3, 4, 8, 11, 14, 16, 18, 20, 22, 24, 27, 28, 31, 32, 35 and 36 have to be reversed coded prior calculating the total score. For the *Accept* subscale, add items 4, 12, 16, 20, 28, 32, and 36. For the *Observe* subscale, add items 1, 5, 9, 13, 17, 21, 25, 29, 30, 33, 37, and 39. For the *Describe* subscale, add items 2, 6, 10, 14, 18, 22, 26, and 34. For the *Act* subscale, add items 3, 7, 11, 15, 19, 23, 27, 31, and 38. For each subscale sum score, find the equivalent interval-level score in the above conversion table. This table cannot be used for respondents with missing data.

Discussion

The KIMS (Baer et al., 2004) is a multidimensional measure of mindfulness widely used to assess four mindfulness-related skills: observing, describing, acting with awareness, and accepting non-judgementally. These mindfulness skills are linked to therapeutic outcomes of MBIs and especially in the context of DBT (Baer et al., 2004; Dimidjian & Linehan, 2003). Our intention was not to assess the overall construct validity of the KIMS as such reports are already available (Baer et al., 2004; Park et al., 2013) but rather to perform investigations and fine-tuning for each individual subscale of the KIMS and test their structural validity using Rasch analysis. Thus, the aim of the current study was to use strategies of Rasch analysis to improve precision and item functioning of the KIMS subscales.

Given the fact that the subscales of the KIMS are commonly found only to be loosely related to each other (Baum et al., 2010), the finding confirmed in the present study, we focused our analyses on the subscales. Accuracy of any ordinal scale is limited, but can be improved up to an interval level using the Rasch model (Tennant & Conaghan, 2007). The results of this analysis show that precision and item functioning of the KIMS subscales can be improved substantially using the Rasch model. Satisfactory fit to the Rasch model was achieved by a few modifications of the KIMS subscales that involved rescoring one item, deleting 5 non-fitting items and combining locally dependent items into subtests. Generally, the results support the structural validity and reliability of the modified KIMS subscales to measure mindfulness skills in the sample population. However, initially low reliability of the Act subscale remained at the level of .60, indicating that the subscale is unable to distinguish between two strata with different ability levels (Fisher, 1992), which limits its applicability. The Rasch analysis confirms unidimensionality of all four modified KIMS subscales meaning that raw scores can be readily transformed into interval level scores using the same metric as long as there are no missing data. Researchers and practitioners may use the conversion algorithms of Table 10 to transform ordinal data into interval-level scores to investigate precise effects of MBIs on specific mindfulness traits.

The modification of the KIMS subscales involved rescoring of only one item (Item 29) that displayed disordered thresholds in the Observe subscale, which indicates that the response options selected by the authors (Baer et al., 2004) were overall appropriate. Also, only five items needed to be removed as they did not fit the Rasch model. No items were removed from the Observe subscale, and only Item 6 (“I can easily put my beliefs, opinions, and expectations into words”) was removed from the Describe subscale due to its poor fit. This item might not fit well to the model because it is not focused on describing experiences, which is the primary target of the Describe subscale. Also, focusing on describing beliefs, opinions, and expectations could be seen as moving away from the present moment and thus associated with less mindfulness.

Two semantically close non-fitting items were deleted from the Accept subscale, namely Item 8 (“I tend to evaluate whether my perceptions are right or wrong”) and Item 24 (“I tend to make judgments about how worthwhile or worthless my experiences are”). Item 8 had the lowest item-to-total correlation for both the full scale (*r*= -.19) and its subscale (*r*= .21), and it is the most difficult item where just a few individuals scored high. Item 24 also seems to measure dichotomous judgement. These items might not work well with the current sample because they focus on extreme dichotomous judgement attitudes perhaps more common among people with borderline personality disorder (Linehan, 1993a, 1993b).

Finally, two non-fitting items were removed from the Act With Awareness subscale: Item 19 (“When I do things, I get totally wrapped up in them and don’t think about anything else”) and Item 35 (“When I’m working on something, part of my mind is occupied with other topics, such as what I'll be doing later, or things I’d rather be doing”). Item 19 had very low item-to-total correlations for the full scale (*r*= .03) and its subscale (*r*= .16) and seems to measure one-pointed concentration rather than mindfulness (Brown, Ryan, & Creswell, 2007; Olendzki, 2005). Item 35 seems to be a negatively worded counterpart of item 19 and it might also lead to bias due to its relatively complex wording. Future studies might examine if rewording of these items can improve the psychometric properties of the KIMS subscales. However, until such investigation has been completed, it is recommended to use the proposed subscale versions together with the ordinal-to-interval scoring algorithm.

Removing non-fitting items resulted in a 34-item KIMS version that included the following modified subscales: the 7-item Accept, the 7-item Describe and the 8-item Act. The original 12-item Observe subscale retained all its items after modification. Combining locally dependent items into subtests was beneficial in achieving satisfactory model fit for the subscales Observe, Describe, and Accept, without the need to discard more items. The psychometric properties of the modified KIMS subscales are, therefore, improved substantially with the exception of the Act subscale, which had initially low internal consistency that could not be improved by the current modifications.

Unlike in the Rasch analysis of the MAAS (Medvedev et al., 2016a), where uniform rescoring of all items was conducted to correct disordered thresholds, the KIMS items displayed no disordered thresholds, with the exception of Item 29, which supports the psychometric properties of the KIMS. Similarly to the MAAS, no more than two items were removed per subscale to achieve a satisfactory fit to the Rasch model, thus providing support for overall good structural validity of both scales. Also, four items (6, 8, 24 and 35) out of the five non-fitting items identified through Rasch analysis of the KIMS were included in the FFMQ constructed through factor analysis of the available mindfulness questionnaires (Baer et al., 2006), and it might be worthwhile to also investigate the functioning of those items in the FFMQ using Rasch analysis. In addition, consistent with the earlier reports (Park et al., 2013), significantly higher mean scores were found for meditators compared to non-meditators for all but Accept subscale of KIMS, which supports construct validity of these subscales.

Limitations and Conclusions

The following limitations should be noted. The study was conducted with a single sample of university students and should be replicated with more diverse samples including clinical and general populations. The analyses might also have been affected to some degree by missing data as well as the uneven distribution of gender, age, and formal meditation experience in the sample. Although the sample reflects New Zealand's diversity of ethnic groups (Statistics New Zealand, 2013), no efforts were made to purposively sample underrepresented groups. Even though a satisfactory fit to the Rasch model was achieved for all subscales of the KIMS, the reliability of the Act subscale could not be improved, and item 29 required rescoring before computing a total score of the Observe subscale. However, if one has complete data, conversion from ordinal to an interval level scale can be conducted simply by adding responses on each modified subscale version and selecting a corresponding interval score in the right column (Table 10). The benefits of this conversion certainly outweigh inconvenience, and the author can be contacted if assistance with data conversion is necessary (Medvedev et al., 2016b).

The current study reported Rasch analysis conducted to advance psychometric properties of the widely-used KIMS, a multidimensional measure of four mindfulness traits. It has been demonstrated that the KIMS subscales are structurally (or internally) valid after modifications that involved rescoring Item 29, removing misfitting items 6, 8, 19, 24, and 35, as well as combining locally dependent items into subtests. Precision of the KIMS can be improved substantially by using the proposed 34-item version of the instrument together with the ordinal-to-interval conversion table presented here (Table 10), without any need to modify the original response format. These findings will be of interest for clinicians applying mindfulness-based interventions and researchers investigating neurophysiological and psychological correlates of trait mindfulness.

Chapter Five. Improving The Precision of the FFMQ Using a Rasch Approach

Introduction

The Five Facet Mindfulness Questionnaire (FFMQ) (Baer et al., 2006) is the most widely used multidimensional measure of mindfulness skills including: Observing (Observe), Act With Awareness (Act), Non-Judging (Nonjudge), Describing (Describe), and Non-reacting (Nonreact) to inner experience. The FFMQ (Baer et al., 2006) includes 39 items. After reverse coding of the 19 negatively worded items, higher FFMQ scores denote greater mindfulness. The FFMQ was originally constructed by combining the 112 items from five available mindfulness scales: the Freiburg Mindfulness Inventory (FMI) (Walach et al., 2006), the Mindful Attention Awareness Scale (MAAS) (Brown & Ryan, 2003), the Kentucky Inventory of Mindfulness Skills (KIMS) (Baer et al., 2004), the Cognitive Affective Mindfulness Scale (Hayes et al., 2004), and the Mindfulness Questionnaire (Chadwick et al. 2008). Using principal axis factor analysis with oblique rotation Baer et al. (2006) extracted five factors out of the 112-item pool, of which four (Observe, Describe, Act and Nonjudge) were labeled in a similar way to the KIMS subscales. One additional extracted component was called Nonreact to inner experience. Initially, all items with factor loadings below .40 were excluded in this analysis, resulting in 64 items representing the five mindfulness facets. Finally, items with the lowest loadings on their factor and various cross-loadings were also excluded. Confirmatory factor analysis provided similarly acceptable fit indices for both the five- and the four-factors solution in the student sample but the five-factor model had better fit with the meditator sample and represents the final five FFMQ subscales (Baer et al., 2006). The five-factor model was also confirmed by Christopher, Neuser, Michael & Baitmangalkar (2012) using a sample of both meditators and non-meditators. Multidimensionality of the FFMQ was also supported by weak correlations found between the subscales ranging from -.07 to .34 (Baer et al., 2006), though slightly higher correlation coefficients were reported in a later study with experienced meditators (Baer et al., 2008).

The five FFMQ subscales have shown good internal consistency with Cronbach’s alphas in the range between .67 and .93 (Park et al., 2013). Meditation experience has been shown to influence the relationship between the overarching construct of mindfulness and the observe facet, reflected by the differences in factor loadings between meditator and non-meditator samples (Baer et al., 2006). The FFMQ subscales and its total score correlate positively with well-being, self-compassion, openness and emotional intelligence and negatively with depression, anxiety, neuroticism, alexithymia and dissociation, which supports the validity of the construct (Baer et al., 2006; Cash & Whittingham, 2010; Fisak & von Lehe, 2012). However, Baer et al. (2006) reported that out of the five FFMQ subscales only three (Act, Nonjudge, and Nonreact) were valid predictors of psychological symptoms.

Despite the popularity of the FFMQ, most psychometric evaluations of the instrument to date have employed classical test theory approaches such as exploratory and confirmatory factor analysis. Only two studies so far using Item Response Theory (IRT) methods have investigated DIF of the FFMQ items with meditator and non-meditator samples (Van Dam et al., 2009; Baer et al., 2010). DIF refers to the case where respondents with the same level on the latent variable (i.e. mindfulness), but from different groups (e.g. meditators and non-meditators), respond systematically differently to an item. Van Dam et al. (2009) investigated item DIF and found differences between demographically unmatched samples in responding to 18 out of 39 FFMQ items. In particular, six items (7, 8, 24, 27, 37 and 38) showed DIF greater than 0.64, which was described as large based on Penfield's criterion (Penfield, 2007). Baer et al.'s (2010) study replicated Van Dam et al. (2009) with demographically matched meditators and non-meditators and found that DIF effect by sample was only significant for four items (1, 11, 18 and 23) out of those identified by Van Dam et al. (2009). However, when Baer et al. (2010) conducted DIF analysis by grouping conceptually related items (e.g. Act items), they found no DIF between meditators and non-meditators. These contradictory findings suggest that further investigation of the FFMQ items DIF is necessary. Further research should also assess the psychometric properties of the FFMQ in order to improve the precision of the instrument up to an interval measure. Such investigation can be carry out using Rasch analysis, which employs a probabilistic logistic model and is specifically suited for this purpose (Tennant and Conaghan 2007; Rasch 1961).

The aim of the present study was to use Rasch analysis to assess the psychometric properties of the FFMQ, which are widely used in mindfulness research, with the intention to improve their precision by generating ordinal-to-interval transformation algorithms.

Method

Participants

The data for the current study were collected from 296 participants in New Zealand and included 200 university students (68%) studying health sciences and 96 individuals sampled from the general population (32%). Participants did not receive any tangible benefit for their participation such as class credit or monetary reward. Sample size for Rasch analysis with polytomous items should be a minimum 20 cases per item in the largest subscale (Lundgren Nilsson & Tennant 2011; Linacre 1994), whichever is greater here because the maximum number of items in the FFMQ subscales is 12. The sample comprised of 229 females (77%), 62 males (21%) and 5 participants with missing gender information. Participants' ages ranged from 17 to 84 years with a mean of 33.09 (SD=18.83). Ethnic groups included 60% Caucasians, 7% Māori, 11% Pasifika, 6% Asian, and 14% of unspecified others. The sample included 36 (12%) participants regularly practicing meditation, as opposed to 259 not practicing meditation, and 1 participant with missing data. The effects of all personal factors on functioning of individual items (DIF) were examined including gender, age, ethnic group, sample (students vs general population) and engagement in meditation and relaxation practices on a regular basis. For this purpose, three approximately equal-sized age categories were created as follows: 17-20 (n=101), 21-35 (n=89) and 36-84 (n=86), where 20 participants' age data was missing.

Procedure

Student participants were invited to complete the survey in lectures and to hand the survey back to the researchers or submit it to a locked collection box at their respective faculty. Students completed the questionnaire in class before the lecture or during a break. General population participants received questionnaires that were equally distributed by the researcher into post-boxes across five main Auckland regions. Homes were selected by their closeness to randomly selected public post-boxes in each region. About 50 surveys were distributed into houses located around each selected public post-box and response rate was about 12%. Auckland is the largest city in New Zealand with a population of 1,333,000 people, which is about 30% of New Zealand population. Participants posted completed questionnaires back to the researchers using a self-addressed, pre-paid envelope. The participants did not received any monetary reward for participation in this research. The authors' university ethics committee approved this study.

Measures

The FFMQ (Baer et al., 2006) uses a 5-point Likert scale format, with responses ranging from ‘Never or very rarely true’ = 1 to ‘Very often or always true’ = 5. Items 3, 5, 8, 10, 12, 13, 14, 16, 17, 18, 22, 23, 25, 28, 30, 34, 35, 38 and 39 are negatively worded and were reversed coded prior to data analysis.

Data Analysis

Descriptive statistics and reliability analysis were conducted using IBM SPSS v.23, and Rasch analysis was performed using RUMM2030 software (Andrich et al., 2009). Rasch analyses were performed for each FFMQ subscale and the total scale, and involved several main steps (Siegert et al., 2010) described in Chapter Two.

Results

The full 39-item scale shows satisfactory internal reliability, with a Cronbach’s alpha value of .89. However, item-to-total correlations for the full scale were low, ranging from .10 to .60 (mean *r*= .39). Consistent with the original validation report (Baer et al., 2006), associations between individual subscales were in the range from non significant correlations to significant correlation coefficients as high as .45.

Prior to the main analysis, the likelihood-ratio test was computed on the initial analysis output for each FFMQ subscale and the total scale, which supported suitability of the unrestricted Partial Credit version of the model (*p* = .001). Fit statistics for each individual item including item locations, fit residuals and Chi-square for the initial analysis of the FFMQ including Observe, Act, Nonjudge, Describe, and Nonreact are included in Table 11. Overall the FFMQ items displayed acceptable fit indices with the exception of Item 32 in the Describe subscale. Item location indicates the difficulty level of an item with higher scores corresponding to more difficult items.

Table 11. *Rasch model item fit statistics for the initial analysis of the FFMQ subscales Observe, Describe, Act, Nonjudge and Nonreact.*

Items / Factors	Item Location	Item-fit Residual	Chi-square
Observe			
15 I pay attention to sensations	-0.06	-0.85	5.24
31 I notice visual elements in art or nature	-0.35	-1.17	5.07
20 I pay attention to sounds	0.01	-0.15	5.01
26 I notice the smells and aromas of things	-0.89	-0.66	3.21
6 I stay alert to the sensations of water	0.40	0.58	4.26
1 I notice the sensations of my body moving	0.73	0.30	1.07
11 I notice how foods and drinks affect thoughts	0.43	1.56	3.93
36 I notice how emotions affect thoughts and behaviour	-0.28	1.70	5.87
Act			
38 doing things without paying attention ^{R}	-0.04	-1.06	6.98
13 I am easily distracted ^{R}	0.48	-1.69	3.94
5 my mind wanders off and I'm easily distracted ^{R}	1.06	-1.15	4.37
8 I don't pay attention to what I'm doing ^{R}	-0.33	-0.83	1.64
34 I do jobs or tasks automatically ^{R}	-0.01	1.17	7.23
18 I find it difficult to stay focused ^{R}	-0.56	0.08	9.28
28 I rush through activities without being attentive ^{R}	-0.09	1.55	12.71
23 I am "running on automatic" ^{R}	-0.51	1.71	3.88
Nonjudge			
25 I shouldn't be thinking the way I'm thinking ^{R}	-0.17	-0.78	2.54
35 I judge myself as good or bad ^{R}	-0.08	0.13	4.46
17 I make judgments about my thoughts ^{R}	0.39	-0.50	2.12
30 I think my emotions are bad or inappropriate ^{R}	-0.63	-0.70	3.84
14 I believe my thoughts are abnormal or bad ^{R}	-0.49	-0.77	5.06
10 I shouldn't be feeling the way I'm feeling ^{R}	0.52	0.85	3.87
39 I disapprove of myself ^{R}	-0.28	1.17	6.07
3 I criticize myself for inappropriate emotions ^{R}	0.73	2.11	5.29
Describe			
37 I can usually describe how I feel at the moment	0.11	-1.21	7.54
2 I'm good at finding words to describe my feelings	-0.08	-1.46	5.05
12 It's hard for me to find the words to describe ^{R}	0.14	-1.79	10.77
16 I have trouble thinking of the right words ^{R}	0.13	-0.81	9.43
7 I can easily put my thoughts into words ^{R}	-0.34	-0.60	0.97
27 when upset, I can find a way to put it into words	0.20	0.34	7.22
32 tendency is to put experiences into words	-0.01	4.57*	15.88
22 I can't find the right words to describe sensation ^{R}	-0.15	2.61	8.34
Nonreact			
33 I just notice distressing things and let them go	0.15	-0.69	3.14
29 notice distressing things without reacting	-0.03	-1.00	10.20
24 I feel calm soon after distressing things	0.43	1.34	2.39
9 I watch my feelings without getting lost in them	0.12	0.01	3.27
19 I am aware of distressing thought or image	-0.16	0.21	2.04
21 I can pause without immediately reacting	-0.51	1.25	5.37
4 I perceive my emotions without reacting to them	0.00	0.81	2.74

Note: $^{\mathbf{R}}$ reverse-scored item. * Significant misfit (p<0.01).

A summary of the model fit statistics for the 8-item Observe subscale is presented in Table 12. The model fit was satisfactory from the beginning, with none of the items displaying disordered thresholds. In line with model expectations, the overall item-trait interaction was not significant (χ^2(32)=33.65, *p* < .01), with a PSI of .76. All individual items demonstrated acceptable model fit, and there were no signs of local dependency between them. Fewer than 5% of *t* tests (4.41%) were significant and this confirmed strict unidimensionality of the Observe subscale (Table 12, Observe, Final).

Table 12. *Rasch model fit statistics for the initial and the final analyses of the FFMQ.*

	Item fit residual		Person fit residual		Goodness of fit		PSI	Significant *t*-tests	
Analyses	**Value / SD**		**Value / SD**		**χ2 (df)**	***p***		**%**	**Lower bound**
Observe									
Final	0.16	1.07	-0.34	1.24	33.65(32)	.39	.76	4.41	1.92
Act									
Initial	-0.03	1.35	-0.48	1.44	50.02 (32)	.02	.86	7.09	4.63
Final	-0.59	1.45	-0.46	1.21	13.70 (20)	.85	.81	5.76	3.28
Nonjudge									
Final	0.19	1.08	-0.43	1.34	33.25 (32)	.41	.89	4.41	1.92
Describe									
Initial	0.21	2.25	-0.56	1.71	65.19 (32)	.01	.89	7.77	5.31
Final	0.07	1.78	-0.62	1.65	33.56 (32)	.39	.88	7.09	4.63
Nonreact									
Initial	0.27	0.91	-0.49	1.54	29.14 (28)	.41	.79	7.80	5.31
Final	0.16	0.75	-0.59	1.58	20.14 (24)	.69	.76	6.78	4.29
FFMQ									
Initial	0.32	1.39	-0.43	2.29	370.52(156)	.01	.90	>10	>10
Final	0.17	0.98	-0.42	1.14	14.28(20)	.82	.82	7.19	4.69

Significant uniform DIF by sample (student vs general population) was found for items 31 (*F*(1,294)= 13.85, *p*<0.001) and 36 (*F*(1,294)= 20.10, *p*<.001). To resolve DIF issue, items 31 and 36 were split for DIF by sample, meaning that these items are now measuring students and general population independently. No DIF was found for any other person factors. Figure 8 shows the person-item threshold distribution plot for the Observe subscale (top panel). This plot illustrates the relationship between distribution of item difficulty and person ability on the latent trait (in this case, Observe), presented after conversion to the same metric, namely logit units. The distribution of person thresholds approximates a normal distribution with some signs of ceiling and floor effects. However, the range of item thresholds of the subscale adequately covered 97% of the sample abilities on the latent trait.

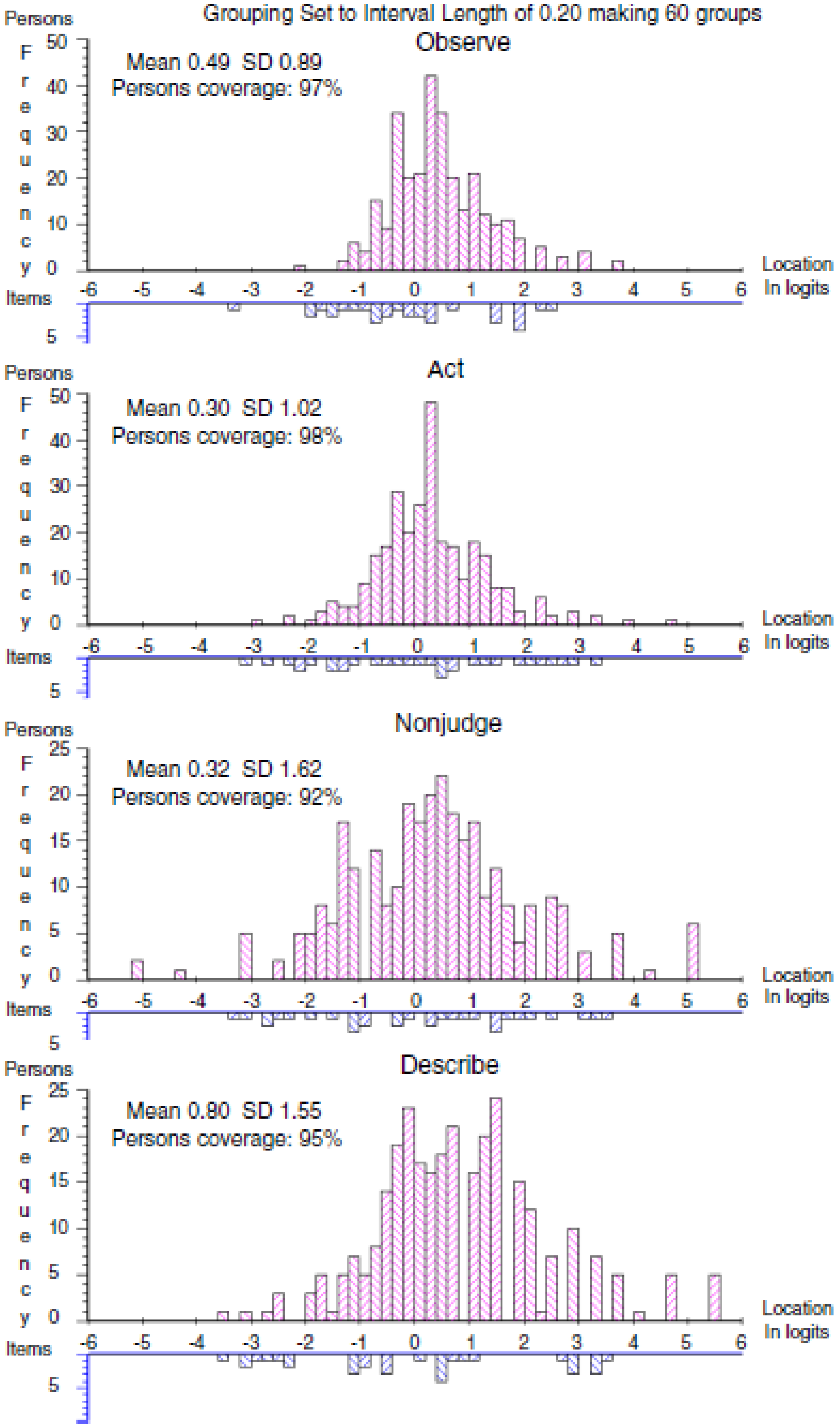


Figure 8. Person-item threshold distributions for the FFMQ subscales Observe, Act, Describe, and Nonjudge.

Initial analysis of the 8-item Act subscale showed acceptable reliability (PSI = .86) and unidimensionality, and all fit residuals were within the acceptable range and had ordered thresholds. However, the overall fit to the model was not satisfactory ($\chi^2(32)=50.02$, $p <$ .05) and local dependency was found between items 5, 8, 13 and 18 as well as between items 34 and 38. These items were subjected to principle component analysis (PCA) to clarify this observation. The first group of four items (5, 8, 13 and 18) clearly loaded on one factor, with factor loadings ranging from .73 to .85. The second group of two items (34 and 38) loaded on the second factor (loadings from .74 to.86), and items 28 and 23 loaded on two distinct factors, with coefficients of .89 and .90, respectively. Therefore, dependent items in the first and the second group were combined into subtests.

An improved model fit was achieved after creating the two subtests, as evidenced by a decreased and no longer significant chi-square ($\chi^2(20)=13.70$, p=.85) and an acceptable PSI value (.81). Unidimensionality of this final model was confirmed by independent samples t-test as described in data analysis section (Table 12). No significant DIF was found for person factors. The person-item threshold distribution for the final analysis of the Act subscale shows that 98% of the person locations are well covered by the item thresholds (Figure 8, Act). Thus, satisfactory fit to the model was evident after combining two sets of locally dependent items into subtests.

Initial analysis of the 8-item Nonjudge subscale of the FFMQ indicated an overall good fit to the model with non-significant chi-square for overall person-trait interaction ($\chi^2(32)=33.25$, p=.406) and reliability (PSI) of .89 (Table 12, Nonjudge, Final). The assumption of unidimensionality was also confirmed. Significant DIF by sample was found for Item 35 ($F(1,294)=11.11$, $p < .01$), which was resolved by splitting this item between students and general population. Figure 8 (Nonjudge) shows the item-person thresholds distribution for the final solution of the Nonjudge subscale. There are some minor signs of ceiling and floor effects but 92% of individuals' locations are fairly well targeted by thresholds of the subscale items. Thus, the Nonjudge subscale satisfied the unidimensional Rasch model without any modifications.

Initial analysis of the 8-item Describe subscale revealed acceptable reliability (PSI=.88) but an unsatisfactory overall fit to the model ($\chi^2(32)= 65.19$, $p < .001$). Additionally, unidimensionality could not be confirmed (Table 12). With a fit residual of 4.57 and chi-square of 15.88, Item 32 ('My natural tendency is to put my experiences into words') showed a significant deviation from the Rasch model expectations and was therefore discarded (Table 11). As a result, the model fit became satisfactory ($\chi^2(32)=32.35$, p = .26). The only significant ($F(1,294) = 10.68$, $p < .01$) and consistent DIF observed was by gender for Item 2 (I'm good at finding words to describe my feelings). To resolve the DIF issue, Item 2 was split by gender. This modification resulted in an even better overall model fit ($\chi^2(32)=33.56$, p=.39, PSI=.88) (Table 12). Figure 8 (Describe) illustrates person-item threshold distribution for the Describe subscale after deleting Item 32. As with the Observe subscale, only some slight ceiling and floor effects were noticeable. Despite some gaps that can be seen between item thresholds, the thresholds cover the abilities of 95% of the sample. In this analysis, a good fit of the Describe subscale to the Rasch model was achieved with the minor modification of deleting one item.

The initial analysis of the 7-item Nonreact subscale indicated overall acceptable fit to the model, as indicated by a non-significant chi-square ($\chi^2(28)=29.14$, p=.405) and a PSI of .79. However, unidimensionality of the subscale was not confirmed (Table 12). Local dependency was found between Item 24 ‘When I have distressing thoughts or images, I feel calm soon after’ and Item 33 ‘When I have distressing thoughts or images, I just notice them and let them go’. Due to potential redundancy of these items, the item map was examined and indicated that Item 33 covers a larger range of individual abilities on the latent trait compared to Item 24, which was therefore removed before the analysis continued. After removing Item 24, the overall model improved ($\chi^2(24)=20.14$, p=.69), and unidimensionality of the modified subscale was evident (Table 12). DIF analysis indicated that item functioning was not affected by any of the examined person factors. The person-item threshold distribution for the final analysis of the Nonreact subscale is presented on Figure 9 (top panel) and shows excellent coverage in that 99% of the sample abilities are located within the range of item thresholds.

The initial Rasch analysis of the full FFMQ scale indicated poor overall model fit ($\chi^2(156)=370.52$, p <.001) and multidimensionality (Table 12, FFMQ, Initial). Therefore, the residual correlation matrix was examined and reflected local dependency patterns between items of each individual subscale. Using the subtests approach of Lundgren Nilsson et al. (2013), the FFMQ subscales were treated as subtests in Rasch analysis. In the first analysis all 39 FFMQ items were included in subtests. The overall model fit improved but was still unsatisfactory ($\chi^2(20)=38.91$, $p < .001$) with low PSI (.62) and multidimensionality evident. At the individual item level, the Observe subtest displayed significant misfit with fit residual value of 3.12.

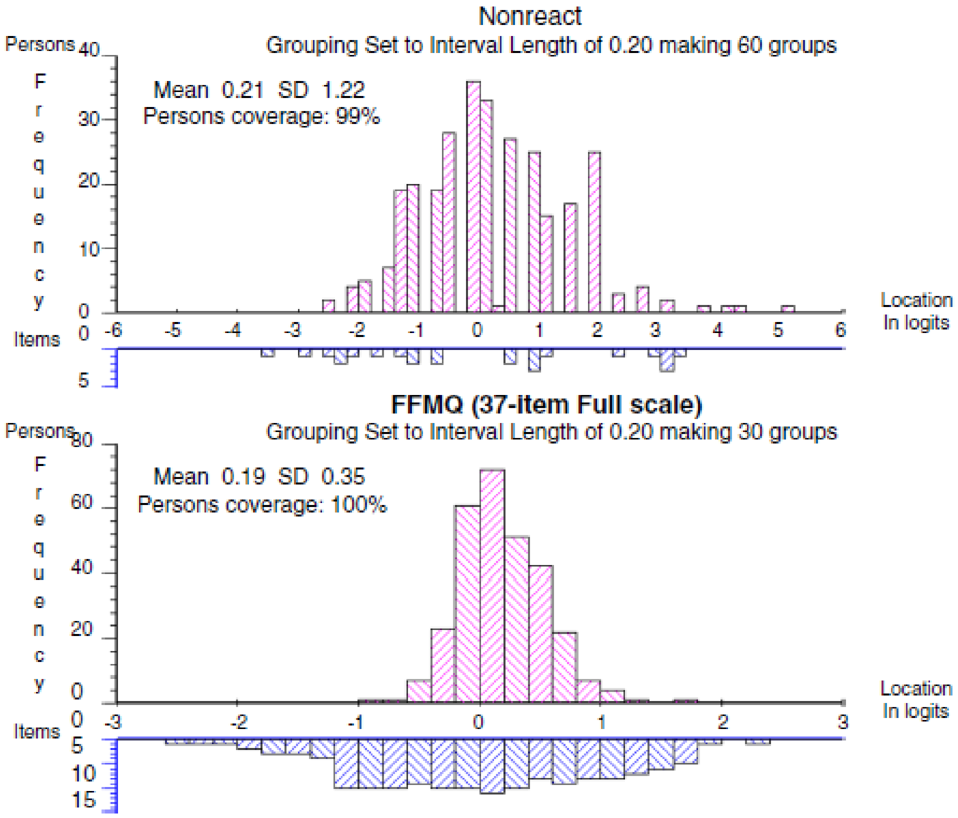


Figure 9. Person-item threshold distributions for the FFMQ subscale Nonreact and the full 37-item scale.

This analysis was replicated by deleting misfitting items 24 and 32 identified from individual subscales analysis prior to creating subtests. This resulted in the best overall model fit ($\chi^2(20)=14.28$, $p=.82$) and unidimensionality was clearly evident. Also, there were no misfitting items and DIF due to personal factors and the PSI improved up to .82. Figure 9 (lower panel) shows that 100% of individual abilities on the latent mindfulness trait were satisfactorily covered by the items' thresholds of the modified FFMQ. However, there are no person locations at the lower end of the scale covered by the item thresholds, which suggests that easy items may be overrepresented in the full scale.

Table 13 and 14 provide conversion scores to transform ordinal raw scores to interval-level data for all five subscales of the FFMQ and the 37-item full scale. For convenience, both ordinal and interval level scores are scaled to the metric of each ordinal subscale. Instructions for use of the conversion table are provided as a table footnote (Tables 13 and 14). This conversion table cannot be used for respondents with missing data.

Table 13. *Converting from ordinal to interval level scores for the subscales of the 37-item FFMQ Observe, Act, Nonjudge, Describe, and Nonreact.*

	Observe		**Act**	**Nonjudge**			**Describe**		**Nonreact**	
Ordinal	Interval Students	Interval General	Interval	Interval Students	Interval General	Ordinal	Interval Male	Interval Female	Ordinal	Interval
8	8.00	8.00	8.00	8.00	8.00	7	7.00	7.00	6	6.00
9	11.17	11.39	10.98	10.71	10.66	8	9.22	9.24	7	8.17
10	13.39	13.60	13.02	12.68	12.60	9	10.88	10.90	8	9.73
11	14.95	15.05	14.41	14.11	14.01	10	12.12	12.13	9	10.84
12	16.17	16.13	15.51	15.27	15.16	11	13.17	13.15	10	11.74
13	17.19	17.01	16.43	16.28	16.17	12	14.11	14.05	11	12.53
14	18.07	17.76	17.27	17.18	17.08	13	14.98	14.87	12	13.25
15	18.85	18.43	18.04	18.01	17.92	14	15.80	15.64	13	13.94
16	19.57	19.04	18.78	18.77	18.70	15	16.57	16.37	14	14.62
17	20.22	19.63	19.49	19.50	19.44	16	17.30	17.08	15	15.30
18	20.83	20.18	20.18	20.20	20.14	17	18.00	17.75	16	16.00
19	21.41	20.72	20.84	20.88	20.82	18	18.68	18.41	17	16.71
20	21.96	21.25	21.47	21.55	21.47	19	19.33	19.06	18	17.45
21	22.50	21.78	22.08	22.20	22.10	20	19.98	19.69	19	18.20
22	23.02	22.31	22.65	22.84	22.73	21	20.63	20.33	20	18.96
23	23.54	22.83	23.21	23.48	23.35	22	21.28	20.97	21	19.74
24	24.05	23.36	23.75	24.12	23.97	23	21.97	21.64	22	20.55
25	24.57	23.90	24.30	24.75	24.59	24	22.68	22.34	23	21.37
26	25.10	24.46	24.85	25.37	25.22	25	23.43	23.08	24	22.23
27	25.64	25.03	25.42	26.00	25.85	26	24.24	23.89	25	23.13
28	26.20	25.62	26.02	26.63	26.48	27	25.10	24.76	26	24.09
29	26.80	26.25	26.66	27.27	27.13	28	26.01	25.71	27	25.13
30	27.42	26.91	27.34	27.92	27.79	29	26.94	26.71	28	26.33
31	28.09	27.61	28.06	28.59	28.48	30	27.90	27.73	29	27.92
32	28.80	28.37	28.83	29.29	29.20	31	28.89	28.77	30	30.00
33	29.56	29.17	29.64	30.03	29.96	32	29.96	29.89		
34	30.39	30.04	30.50	30.83	30.78	33	31.20	31.15		
35	31.30	31.00	31.42	31.71	31.68	34	32.83	32.82		
36	32.33	32.07	32.43	32.70	32.68	35	35.00	35.00		
37	33.52	33.31	33.60	33.86	33.86					
38	35.00	34.84	35.04	35.29	35.30					
39	37.08	36.99	37.09	37.27	37.28					
40	40.00	40.00	40.00	40.00	40.00					

Note: All items are scored from 1 to 5 and negatively worded items 3, 5, 8, 10, 12, 13, 14, 16, 17, 18, 22, 23, 25, 28, 30, 34, 35, 38 and 39 have to be reversed coded prior calculating the total score. For the *Observe* subscale, add items 1, 6, 11, 15, 20, 26, 31 and 36. For the *Describe* subscale, drop item 32 and add items 2, 7, 12, 16, 22, 27 and 37. For the *Act* subscale, add items 5, 8, 13, 18, 23, 28, 34, 38. For the *Nonjudge* subscale, add items 3, 10, 14, 17, 25, 30, 35 and 39. For *Nonreact* subscale, drop item 24 and add items 4, 9, 19, 21, 29, 33. For each subscale sum score, find the equivalent interval-level score in the above conversion table. This table cannot be used for respondents with missing data.

Table 14. *Converting from ordinal to interval level scores for the full 37-item FFMQ.*

Ordinal	Interval	Ordinal	Interval	Ordinal	Interval	Ordinal	Interval
37	**37.00**	**74**	**96.58**	**111**	**111.58**	**148**	**127.39**
38	**49.52**	**75**	**97.02**	**112**	**111.98**	**149**	**127.88**
39	**57.38**	**76**	**97.43**	**113**	**112.37**	**150**	**128.37**
40	**62.33**	**77**	**97.86**	**114**	**112.76**	**151**	**128.87**
41	**65.93**	**78**	**98.28**	**115**	**113.18**	**152**	**129.38**
42	**68.77**	**79**	**98.71**	**116**	**113.57**	**153**	**129.89**
43	**71.10**	**80**	**99.12**	**117**	**113.97**	**154**	**130.40**
44	**73.11**	**81**	**99.54**	**118**	**114.38**	**155**	**130.92**
45	**74.86**	**82**	**99.95**	**119**	**114.77**	**156**	**131.45**
46	**76.42**	**83**	**100.37**	**120**	**115.19**	**157**	**132.00**
47	**77.84**	**84**	**100.76**	**121**	**115.58**	**158**	**132.55**
48	**79.12**	**85**	**101.17**	**122**	**116.00**	**159**	**133.10**
49	**80.30**	**86**	**101.57**	**123**	**116.39**	**160**	**133.68**
50	**81.41**	**87**	**101.98**	**124**	**116.80**	**161**	**134.27**
51	**82.45**	**88**	**102.40**	**125**	**117.22**	**162**	**134.86**
52	**83.40**	**89**	**102.79**	**126**	**117.63**	**163**	**135.47**
53	**84.30**	**90**	**103.20**	**127**	**118.05**	**164**	**136.10**
54	**85.17**	**91**	**103.60**	**128**	**118.46**	**165**	**136.75**
55	**85.98**	**92**	**103.99**	**129**	**118.89**	**166**	**137.44**
56	**86.73**	**93**	**104.41**	**130**	**119.31**	**167**	**138.13**
57	**87.46**	**94**	**104.80**	**131**	**119.74**	**168**	**138.86**
58	**88.17**	**95**	**105.20**	**132**	**120.16**	**169**	**139.63**
59	**88.82**	**96**	**105.61**	**133**	**120.59**	**170**	**140.42**
60	**89.47**	**97**	**106.00**	**134**	**121.02**	**171**	**141.26**
61	**90.08**	**98**	**106.40**	**135**	**121.46**	**172**	**142.15**
62	**90.67**	**99**	**106.81**	**136**	**121.89**	**173**	**143.12**
63	**91.22**	**100**	**107.21**	**137**	**122.32**	**174**	**144.16**
64	**91.77**	**101**	**107.60**	**138**	**122.78**	**175**	**145.28**
65	**92.31**	**102**	**107.99**	**139**	**123.21**	**176**	**146.53**
66	**92.82**	**103**	**108.39**	**140**	**123.66**	**177**	**147.93**
67	**93.33**	**104**	**108.80**	**141**	**124.12**	**178**	**149.50**
68	**93.82**	**105**	**109.20**	**142**	**124.57**	**179**	**151.32**
69	**94.30**	**106**	**109.59**	**143**	**125.04**	**180**	**153.48**
70	**94.77**	**107**	**109.98**	**144**	**125.50**	**181**	**156.15**
71	**95.24**	**108**	**110.38**	**145**	**125.97**	**182**	**159.59**
72	**95.70**	**109**	**110.77**	**146**	**126.44**	**183**	**164.44**
73	**96.13**	**110**	**111.17**	**147**	**126.92**	**184**	**172.29**
						185	**185.00**

Note: All items are scored from 1 to 5 and negatively worded items 3, 5, 8, 10, 12, 13, 14, 16, 17, 18, 22, 23, 25, 28, 30, 34, 35, 38 and 39 have to be reversed coded prior calculating the total score. Drop items 24 and 32, sum all remaining items scores and find the equivalent interval-level score in the above conversion table. This table cannot be used for respondents with missing data.

The ordinal-to-interval scale conversions provided here increase the precision of the FFMQ subscales and the full scale by using scoring algorithms without the need for modifications of the original response format of the instrument.

The ordinal FFMQ responses were transformed into interval level data using conversion Tables 13 and 14, and descriptive statistics were computed including means and standard deviations for meditator and non-meditator groups presented together with Cronbach's alpha coefficients for each subscale of the 37-item FFMQ in Table 15. Independent samples *t*-tests indicated significantly higher means for meditators on the Observe and Nonreact subscales. These results should be interpreted with caution due to disproportionately small number of meditators in the sample.

Table 15. *Means and standard deviations (SD) for meditators and non-meditators based on the interval level data of the 37-item FFMQ, and Cronbach's alpha coefficients.*

Subscale	Meditators (n=34)		Non-Meditators (n=237)		Cronbach's alpha
	Mean	SD	Mean	SD	
Observe	**27.42***	**3.00**	**25.59***	**3.17**	**0.76**
Act	**24.49**	**2.97**	**24.49**	**2.73**	**0.87**
Non-Judge	**25.76**	**4.45**	**24.89**	**5.00**	**0.90**
Describe	**23.71**	**4.42**	**22.73**	**4.30**	**0.88**
Non-react	**18.98***	**3.23**	**18.06***	**2.47**	**0.79**
FFMQ	**118.18***	**8.41**	**115.18***	**6.59**	**0.89**

Note:* Mean difference is significant at the level .05

Discussion

The FFMQ (Baer et al., 2006) is a widely used measure of the five mindfulness facets and more efforts are necessary to increase its precision in discriminating between individual trait levels. The primary focus of this study was to conduct psychometric diagnostics for each individual subscale and the total FFMQ, and to examine internal construct validity of the measure. Also, the current study employed strategies of Rasch analysis to improve the psychometric properties of the FFMQ.

The results of this study have demonstrated a successful application of the Rasch model that allows researchers to improve the precision of the instrument using the ordinal-to-interval conversion algorithms presented here. The two FFMQ subscales Observe and

Nonjudge met expectations of the unidimensional Rasch model without any modifications, while Nonreact, Describe, and Act facets were modified to achieve the best model fit. However, regardless of these modifications, the psychometric properties of all facets can be improved by using the ordinal-to-interval conversion for these subscales because their items have varying degrees of difficulty and hence contribute differently to the total subscale scores. Therefore, the conversion algorithms of Tables 13 and 14 can be used to transform ordinal responses into interval-level data required for parametric statistics, which increases the precision of the instrument. This conversion can be conducted for each subscale as long as there are no missing data and does not require any modification of the current response format because the transformation algorithms already account for modifications. Moreover, an ordinal-to-interval conversion spreadsheet in Excel format will be available online to simplify the conversion, and the author can be contacted if more assistance with data conversion is necessary. Satisfactory overall model fit and relevant subscale model fit could not be achieved without removing items 24 and 32. Therefore, the psychometric properties of the total FFMQ and the Describe and Nonreact subscales will be improved by removing these items.

Overall, the results support internal construct validity, unidimensionality and acceptable reliability of the FFMQ after minor modifications. Unlike the recent MAAS Rasch analysis (Medvedev et al. 2016a), where all items were rescored uniformly to correct disordered thresholds, the FFMQ items showed no disordered thresholds, which supports the utility of the original response options selected by Baer et al. (2006). Only a few minor modifications, such as combining locally dependent items into subtests and removing two misfitting items, were necessary to achieve the best fit to the Rasch model for the remaining three subscales and the full scale. Local dependency found between individual items in 3 out of 5 subscales refers to a high degree of similarity or shared variance between those items, which may result in spurious factors and appear as multidimensionality. However, after successful resolving of local dependency by combining these items into subtests each facet was clearly unidimensional.

Compared to the recent Rasch analysis of the similarly structured KIMS (Medvedev et al., 2016b), where five non-fitting items were removed, all but two FFMQ items displayed acceptable psychometric properties. Therefore, only two items that showed significant issues were excluded to achieve the best model fit in the current analysis. Item 24 ‘When I have distressing thoughts or images, I feel calm soon after’ was removed from the Nonjudge subscale. This item might not work well in assessing a judging attitude in the population of the present sample because it implies that one has distressing thoughts or images at least sometimes. Additionally, setting the time frame as ‘soon after’ can lead to a wide range of interpretations ranging from few a minutes to a few hours. Item 24 also displayed a large bias in the DIF analysis conducted by Van Dam et al. (2009). Re-wording this item and providing a specific time frame (e.g. one hour) instead of ‘soon’ may be necessary to improve its accuracy.

Item 32 (‘My natural tendency is to put my experiences into words’), which is Item 34 in the KIMS, was excluded from the Describe subscale in the current analysis as well as in the earlier Rasch analysis of the KIMS (Medvedev et al. 2016b). This item might be too vague to capture the ability to describe experiences because it asks about ‘tendency’, which is quite different from ‘ability’. For instance, a tendency to jump will not make one win a competition, while the ability might. Consistent misfit of this item found in two studies with different samples suggests that re-wording is necessary, for instance, ‘ability’ can be used instead of ‘tendency’.

The modified FFMQ subscales showed substantial improvement of their psychometric properties. Compared to the recent Rasch analysis of the KIMS (Medvedev et. al. 2016b), the corresponding FFMQ subscales Nonjudge, Describe and Act have demonstrated higher reliability indices and better coverage of individual abilities by subscale item thresholds. In particular, PSI for Act subscale has increased from .60 in the final KIMS analysis up to .80 in the current FFMQ solution suggesting improved psychometric properties of the FFMQ compared to the KIMS. The fact that one item was removed in only two out of five subscales to achieve the best Rasch model fit supports structural validity of all the FFMQ subscales and the total scale. Even though DIF was observed for some items by meditation experience, this did not appear to be a substantial source of DIF, in line with findings of Baer et al. (2010), and could be corrected in the model. To account for differences between students and the general population in responding to items in the Observe and Nonjudge subscales and gender differences in responding to items from the Describe subscale found in this study different conversion tables were produced for use with students and the general population as well as for male and female respondents (Table 13). Different conversion tables by demographic factors are not necessary for the total FFMQ interval score (Table 14), since no DIF was observed at the full scale level.

When facets were treated as subtests and misfitting items 24 and 32 removed, the full FFMQ scale met expectations of a unidimensional Rasch model, confirming the presence of an overarching mindfulness trait. If high precision of measurement is required for assessment of individual aspects of mindfulness then facets interval scores would be more appropriate. In other words, the total interval-level FFMQ score reliably assesses mindfulness as a higher-order construct, and facet interval scores provide more detailed information for each of the five facets.

Limitations and Conclusions

The following limitations should be acknowledged. Although the diversity of New Zealand's ethnic groups is reflected in the sample (Statistics New Zealand, 2013), underrepresented ethnic groups were not purposively sampled. The results might have been affected by disproportional distribution of gender, age, formal meditation experience as well as students compared to general population in the sample. However, gender, age and sample groups were large enough for calibration of the FFMQ items (Linacre, 1994) meaning that conversion algorithms presented here will reliably increase precision of measurement for students and the general population. Moreover, DIF analysis was replicated with matched sample sizes of students and general population to address disproportional distribution in these samples. Therefore, if complete ordinal FFMQ data are collected, they can be transformed from an ordinal to an interval level scale simply by adding individual subscale scores, excluding items 24 and 32, and finding an equivalent interval score in the right column. The advantage of this transformation certainly outweighs the inconvenience, and the authors can be contacted if assistance with data transformation is needed. However, these conversion algorithms may not be suitable for clinical populations (e.g. stroke or trauma) and further studies should replicate these findings with more diverse populations not represented in the current sample.

Mindfulness is an important contributor to both physical and psychological health, which raises the importance of its precise measurement. The current Rasch analysis was conducted to enhance psychometric properties of the widely used measure of five mindfulness facets the FFMQ. The study has demonstrated successful application of the Rasch model that allows researchers to improve precision of the instrument by using the included Rasch transformation algorithms. These findings can be beneficial in many areas where more accurate assessment of mindfulness and its facets is required.

Chapter Six. Assessing the CHIME's Psychometric Properties Using Rasch Analysis.

Introduction

Mindfulness-based interventions are generally found to be beneficial for improving well-being and alleviating symptoms of psychological distress, although accurate measurement of the psychological construct of mindfulness itself remains a challenge. This chapter describes Rasch analysis conducted to investigate the psychometric properties of the CHIME with a sample of 443 participants from the general population aiming at improvement of instrument precision. The author of this work has presented his earlier works that used Rasch analysis to enhance psychometric properties of the MAAS, the KIMS & the FFMQ at the International Mindfulness Conference in Rome 2016 (Medvedev, Siegert & Krägeloh, 2016c). After considering the benefits of Rasch analysis, Dr Claudia Bergomi, who has developed the CHIME with her colleagues (Bergomi et al., 2014), has kindly provided her dataset (n=443) for this study. Bergomi et al. (2013) has also conducted theoretical work that highlighted a wide range of characteristics and aspects of mindfulness, which can be assessed comprehensively by the recently developed eight-factor CHIME. While this 37-item German-language scale has demonstrated acceptable psychometric properties, its ability to discriminate precisely across individual mindfulness levels has not yet been rigorously investigated.

To date, the most widely used multidimensional measure of trait mindfulness is the Five Facet Mindfulness Questionnaire (FFMQ) (Baer et al., 2006) that includes five subscales: observing, describing, act with awareness, non-judging and non-reacting to inner experience. However, recent analysis of validated mindfulness measures identified a wider range of aspects underpinning the construct (Bergomi et al., 2013). As a result, the 37-item Comprehensive Inventory of Mindfulness Experiences (CHIME) (Bergomi et al., 2014) was constructed, which is a multidimensional German-language measure of mindfulness covering eight mindfulness aspects identified across currently available mindfulness measures. The CHIME has eight subscales measuring eight mindfulness aspects including: awareness of internal experiences, awareness of external experiences, acting with awareness, accepting nonjudgmental attitude, nonreactive decentering, openness to experience, awareness of thoughts' relativity, and insightful understanding. The CHIME was validated in a community sample (n=298) and a sample of Mindfulness-Based Stress Reduction (MBSR) course participants (n=161), and overall good internal consistency (α range 0.70-0.90) as well as test-retest reliability (r range 0.70-0.90) were reported. The adequacy of the eight-factor structure of the CHIME was confirmed using another sample (n=202) (Bergomi et al., 2014). Measurement invariance of the single items was tested over groups differing in age, gender, meditation experience, and symptom load, which indicated absence of systematic differences in the semantic understanding of items. Construct validity of the CHIME was supported by strong correlations (r=0.85) with the total score of the FFMQ (Baer et al., 2006) as well as conceptually similar subscale scores (e.g., act with awareness, r=0.63). Moderate correlations were found between the CHIME total score and measures of wellbeing (.40), depression (-.46), and anxiety (-.39) in the predicted directions (Bergomi et al., 2014).

Similar to other mindfulness measures, the CHIME functions at the ordinal level of measurement, which does not satisfy fundamental assumptions of parametric statistical tests such as ANOVA and hence limits its application in research. For instance, an ordinal measurement does not support mathematical operations used to compute means and standard deviations (Merbitz, Morris, & Grip, 1989). Additionally, every item explains a different amount of information relevant to the latent trait meaning that the sum of all item scores might not be an accurate estimate of the latent trait (Stucki et al., 1996; Allen & Yen 1979). As a result, comparisons between CHIME scores and interval measures (e.g. EEG, blood tests) in modern mindfulness research may be limited and even misleading. Therefore, it is necessary to investigate the psychometric properties of the CHIME in order to improve its precision up to an interval level scale. Rasch analysis provides a template for conversion from an ordinal-to-interval measure and represents a suitable psychometric method for this purpose (Rasch, 1961; Tennant & Conaghan, 2007). The aim of the current study was to conduct Rasch analysis to investigate and enhance the psychometric properties of the CHIME and to produce ordinal-to-interval transformation algorithms for use in mindfulness research.

Method

Participants

The sample included 443 German-speaking participants of the general population from the dataset previously reported elsewhere (Bergomi et al., 2014). The sample size was larger than recommended for all purposes of Rasch analysis (Linacre, 1994). Of those, 202 (46%) were male and 241 (54%) were female. Less than half of the sample (n=199) reported some experience with meditation. Ages ranged from 18 to 82 years, with a mean of 36.96 and standard deviation of 13.51. To enable investigation of differential item functioning (DIF) in Rasch analysis across different age groups the following approximately equal-sized age categories were created: 18-27, 28-40, and 41-82. The created age groups can be meaningfully related to life experience because the age 18-27 is mainly associated with acquiring qualification and establishing a life style (unexperienced). The age 28-40 is related overall to a period in life when people tend to build family relationships and establish more consistent life style patterns and mastering skills (intermediate experience). The age 41-82 is normally associated with established lifestyle and substantial life experience (experienced).

Procedure

Participants were recruited through the networks of the researchers based in Switzerland (Bergomi et al., 2014). Participants were contacted directly and invited to complete an online questionnaire. The authors' institutional ethic committee has approved the study.

Measures

The 37-item CHIME (Bergomi et al., 2014) is a self-report questionnaire that includes eight subscales measuring awareness of internal experiences (AwareInt), awareness of external experiences (AwareExt), acting with awareness (ActAware), accepting nonjudgmental attitude (AccNJ), nonreactive decentering (NrDec), openness to experience (Openness), awareness of thoughts' relativity (Relativity), and insightful understanding (Insight). The measure employs a 6-point Likert scale format from 'almost never'=1 to 'almost always'=6, and negatively-worded items (7, 10, 17, 19, 22, 26, 30, 33, 36) need to be reversed coded before calculating subscale scores. Total scores are calculated by adding responses to each individual subscale item with higher scores corresponding to higher levels of mindfulness.

Data Analysis

Rasch analysis was performed using RUMM2030 software (Andrich, Sheridan, & Luo, 2009). The Rasch model requires unidimensionality that will be tested along with other psychometric criteria. Prior to the main analysis, the likelihood-ratio test was computed on the initial analysis output for each CHIME subscale, which supported suitability of the unrestricted Partial Credit version of the model (p<.001). Rasch analyses were performed for each CHIME subscale separately as well as the full scale where all subscales were treated as testlets, consistent with empirically tested methodology (Lundgren Nilsson et al., 2013). Rasch analysis was conducted following several main steps that are described in Chapter Two and elsewhere (Siegert et al., 2010). DIF was investigated by personal factors including gender, age and meditation experience consistent with earlier psychometric studies on mindfulness (Baer et al., 2006; Medvedev et al., 2017).

Results

Fit statistics for each individual item including item locations, fit residuals and chi-square for the initial analysis of the eight CHIME subscales are included in Table 16.

Table 16. *Initial Rasch model fit statistics for the CHIME subscales items.*

Items / Factors	Item Location	Item-fit Residual	[a]Chi-square
AwareInt			
1 notice mood changes	-0.75	0.04	9.29
5 sitting or lying perceive body sensations	0.49	1.02	11.70
14 talk to others and notice my feelings	-0.21	0.51	5.64
29 notice changes in my body such as breathing	0.65	1.56	13.41
34 aware how I am currently feeling	-0.19	-0.80	17.00
AwareExt			
9 calm soon after distressing thoughts/images	-0.23	-2.56	15.69
18 notice distressing thoughts/images without respond	0.23	1.60	7.53
21 in difficult situations can pause without respond	0.13	2.33	9.83
27 it does not take long to notice thoughts/emotions	-0.14	-0.63	10.34
ActAware			
10 break or spill things out of inattention	-0.15	2.10	5.54
12 easy to stay focused on what I am doing	-0.47	0.71	16.95
17 distracted by memories, images or day-dreaming^R	0.45	-0.33	12.39
26 have to reread because thinking of something else^R	0.17	-0.86	9.03
AccNJ			
2 kind to myself in the ups and downs of life	-0.63	-0.04	4.46
7 hard on myself when I make a mistake^R	0.94	1.86	18.51
11 see my mistakes/difficulties without judging myself	-0.44	1.40	9.62
32 treat myself with understanding if making a mistake	0.06	-0.97	12.21
36 resent my own mistakes and weaknesses	0.07	-1.74	10.61
NrDec			
8 calm soon after distressing thoughts/images	0.39	0.41	4.66
13 notice distressing thoughts/images without respond	0.36	-0.83	8.98
16 in difficult situations can pause without respond	-0.10	1.60	6.15
20 it does not take long to notice thoughts/emotions	0.18	-0.31	7.98
25 able to observe thoughts/feelings without distraction	-0.21	-1.03	5.28
28 notice my thoughts/feelings from a distance	-0.61	0.94	4.97
Openness			
19 stay busy to keep specific thoughts/feelings away^R	-0.55	0.40	6.30
22 distract myself when I feel unpleasant emotions^R	0.15	-0.55	9.63
30 try to get rid of angry or fearful feelings^R	0.33	0.33	10.53
33 try to avoid pain sensation as much as possible ^{R}	0.07	1.95	6.48
Relativity			
4 clear that evaluations of situations/people can change	-0.15	1.06	3.39
23 aware that thoughts/interpretations are not facts	0.37	1.58	9.87
31 aware that my view on things is subjective	0.34	-1.07	7.46
35 aware that my own opinions may change	-0.57	-0.20	15.72
Insight			
3 notice difficulties due to negative attitude	0.16	3.24*	22.22
6 notice seeing things more complicated as they are	0.19	0.30	4.81
15 can see self-created hard time with humour	-0.23	-1.12	11.07
24 able to smile about self-created problems	-0.05	-1.03	16.23
37 notice if make life needlessly difficult	-0.07	0.39	6.39

Note: $^{\mathbf{R}}$ reverse-scored item. * Significant misfit (p<0.05); [a]Degree of freedom (df)= 9 for all items.

Significant misfit was only evident for Item 3 ('notice difficulties due to negative attitude') from the insight subscale but no other items. None of the 37 items of the CHIME displayed significantly disordered thresholds. Figure 10 shows response probability curves for Item 1 illustrating ordered thresholds, as is typical for most CHIME items.

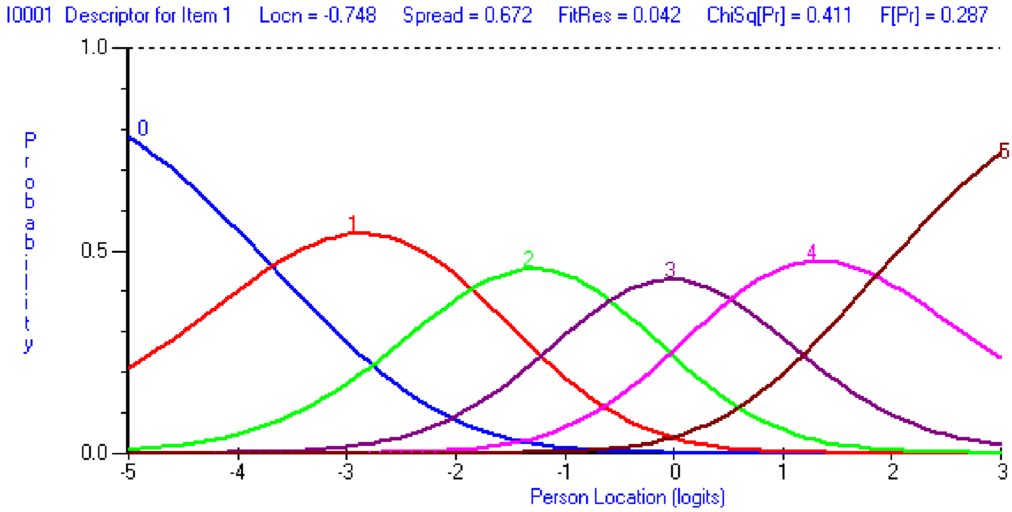


Figure 10. Item category probability curves illustrating ordered thresholds for CHIME item 1, which is typical for most CHIME items.

Table 17 summarises Rasch model fit statistics for the initial and final analysis of each CHIME subscale and the full scale. The initial analysis was also the final one for all CHIME subscales, except for Insight subscale. For these subscales, the best model fit was achieved without modifications (Table 17). In line with model expectations, the overall item-trait interaction for these subscales was not significant with chi square ranging from 32.94 to 57.04 (*p*>.05). There were no signs of local dependency between any of the items in these subscales, no DIFs due to personal factors, and evidence of unidimensionality was obtained.

Figures 11 to 13 show item-person thresholds distributions for the final solutions of the CHIME subscales and the full scale, which provides visual illustration of how well the trait abilities of both meditators and non-meditators of the current sample are covered by thresholds of individual items. The graphs show distributions of meditators and nonmeditators separately because non-parametric Mann-Whitney tests indicated significant group differences across all CHIME subscales and the total score (*p*<.05). Both the AwareExt (Figure 11, top panel) and the AwareInt subscales (Figure 11, centre panel) showed some signs of ceiling effects, which is linked to meditation experience for the AwareInt subscale. Nevertheless, items of both subscales still covered approximately 90% of individuals in the present sample. For the following subscales there were no significant ceiling or floor effects, and over 95% of the sample's abilities were well covered by items: ActAware (Figure 11, bottom panel), AccNJ, NrDec, and Openness

(Figure 12). Even though the ceiling effect was not significant for the openness to experience subscale (*p*>.05), a few meditators with higher levels of the trait were outside of the scale coverage (Figure 12, bottom panel). Figure 13 (top panel) shows that a majority of the sample distribution was above the mean of the relativity subscale items thresholds suggesting that easy items are overrepresented in the Relativity subscale. Even though items thresholds covered 95% of the sample, there were item thresholds outside of the person distribution at the lower end of the scale.

Table 17 *Summary of fit statistics for the initial and the final Rasch analyses of the eight CHIME sub-scales: AwareInt, AwareExt, ActAware, AccNJ, NrDec, Openness, Relativity, Insight.*

	Item fit residual		Person fit residual		Goodness of fit		PSI	Independent *t*-test	
Analyses	Value	SD	Value	SD	χ^2 (df)	*p*		%	%LB[a]
AwareInt									
Final	0.47	0.91	-0.39	1.18	57.04 (45)	.11	.71	3.61	2.26
AwareExt									
Final	0.18	2.22	-0.48	1.19	43.38 (36)	.19	.75	4.06	2.03
ActAware									
Final	0.40	1.30	-0.39	1.08	43.90 (36)	.17	.66	3.84	1.81
AccNJ									
Final	0.10	1.53	-0.50	1.21	55.40 (45)	.14	.84	4.97	2.94
NrDec									
Final	0.13	1.03	-0.53	1.38	38.01 (54)	.41	.84	7.00	4.97
Openness									
Final	0.53	1.04	-0.49	1.20	32.94 (36)	.61	.74	4.29	2.26
Relativity									
Final	0.34	1.20	-0.57	1.40	36.44 (32)	.27	.74	5.87	3.84
Insight									
Initial	0.36	1.76	-0.61	1.54	60.71 (45)	.06	.72	7.22	5.19
Final	0.28	2.19	-0.59	1.27	28.69 (36)	.80	.67	4.29	2.26
CHIME									
Initial	0.41	1.84	-0.41	2.13	642.49 (333)	.00	.92	23.93	21.90
Testlets	0.08	1.82	-0.42	1.27	95.35 (72)	.03	.82	9.48	7.45
Final	0.20	1.25	-0.40	1.21	70.76 (63)	.23	.82	5.42	3.39

Note: [a]LB = lower bound of the 95-% confidence interval.

Initial analysis of the insight subscale indicated the overall satisfactory model fit (χ^2 (45)=60.71, *p*=.06), but there was evidence of multidimensionality (Table 17, Insight) and significant misfit of the Item 3 (‘notice difficulties due to negative attitude’) (Table 16). Therefore, the residual correlation matrix was examined, which revealed local dependency between items 15 (‘can see self-created hard time with humour’) and 24 (‘able to smile about self-created problems’). Given the similarity of these items’ content, this may be expected.

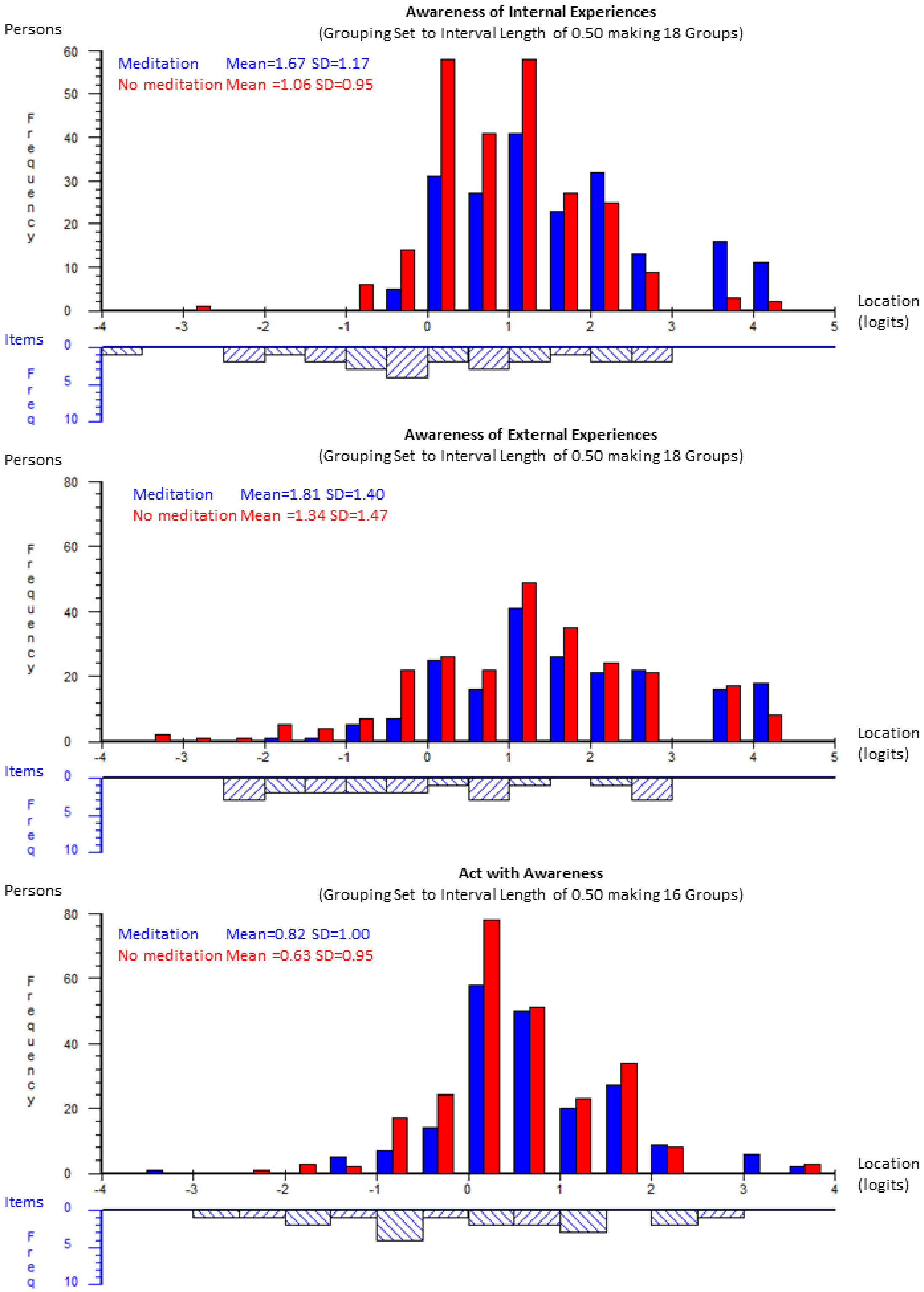


Figure 11. Person-item threshold distribution for the CHIME subscales awareness of internal experiences (AwareInt), awareness of external experiences (AwareExt), acting with awareness (ActAware).

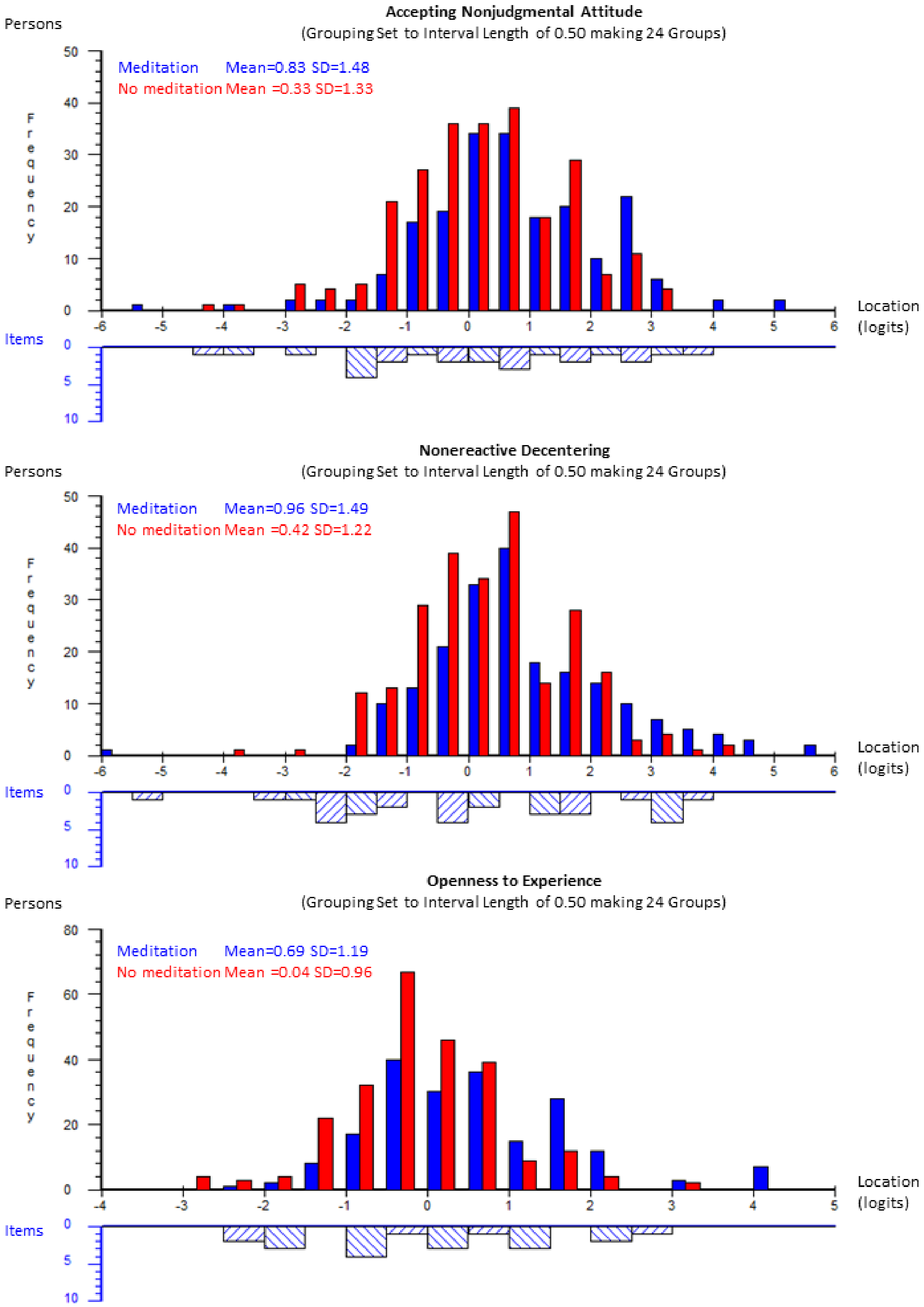


Figure 12. Person-item threshold distribution for the CHIME subscales accepting nonjudgmental attitude (AccNJ), nonreactive decentering (NrDec), openness to experience (Openness).

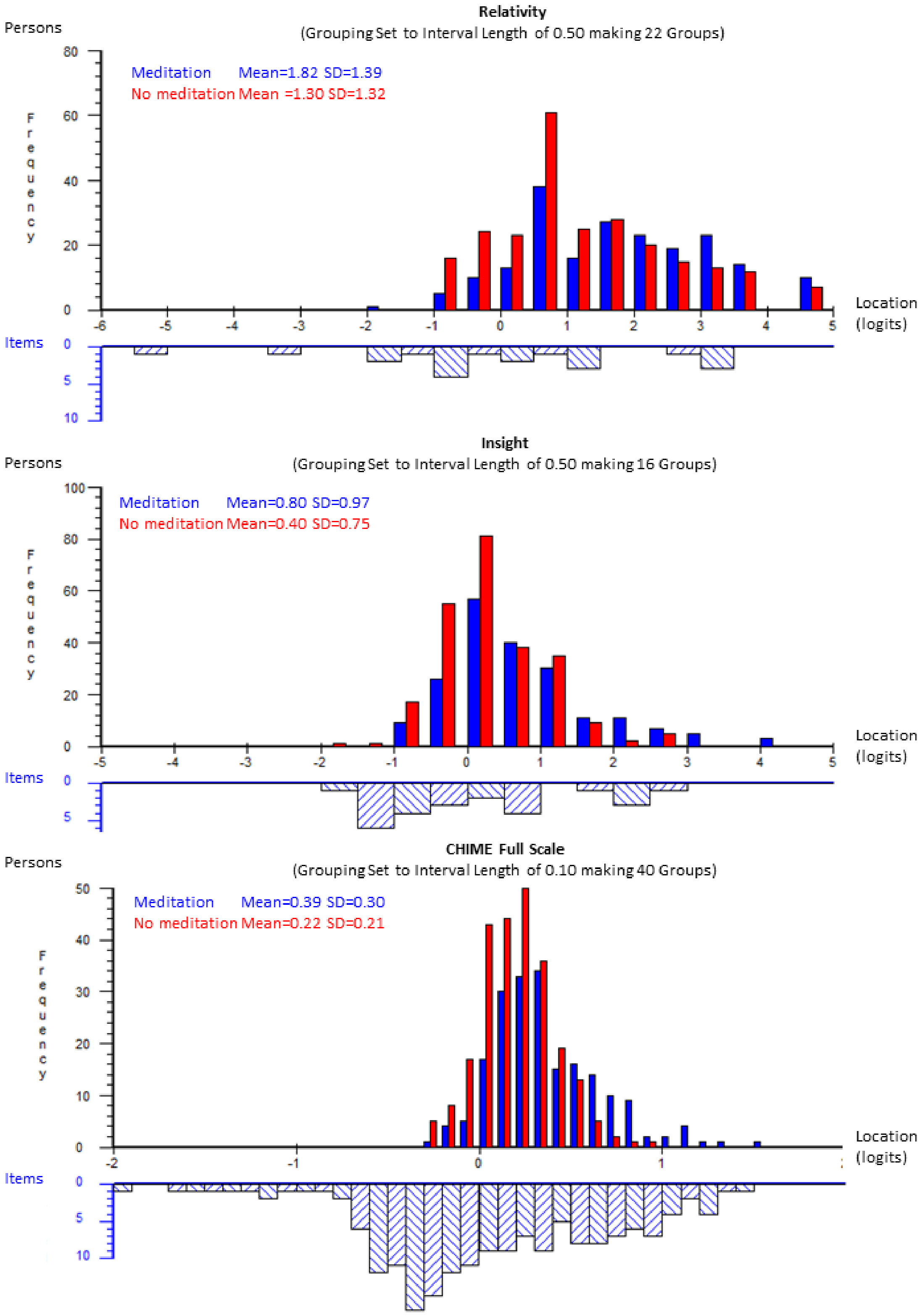


Figure 13. Person-item threshold distribution for the CHIME subscales Relativity, Insight and the full scale.

Local dependency found between items 15 and 24 was resolved by combining these items into a testlet resulting in strict unidimensionality of the subscale and improvement of the chi square (χ^2 (36)=28.69, *p*=.80) (Table 17, Insight). At this stage, item thresholds adequately covered over 95% of the sample population meaning that there were no significant floor or ceiling effects (Figure 13).

The subscales AccNJ and NrDec both exhibit good reliability as indexed by PSI of .84 (Table 17, AccNJ, NrDec), and reliability of other subscales (AwareInt, AwareExt, Openness, and Relativity) was above .70 and thus acceptable. However, the discriminating ability of the ActAware and Insight subscales were less satisfactory as indexed by a PSI of .66, meaning that the precision of these subscales may not be adequate if high precision is required such as when making clinical judgements about individuals.

Finally, the full CHIME scale was fitted to the Rasch model and initially exhibited a poor model fit (χ^2 (333)=642.49, p<.01) as well as evidence for multidimensionality (Table 17, CHIME, Initial). Therefore, items of each subscale were combined into testlets following the recommended approach by Lundgren Nilsson et al. (2013). After creating the testlets, the overall model fit improved substantially, although chi square was still significant (χ^2 (72)=95.35, *p*=.03) and unidimensionality was still not confirmed. At this stage, the correlation matrix between the eight subscales was examined and revealed that the acting with awareness subscale had a moderate correlation with the nonreactive decentering subscale and low correlations with all other subscales. As this indicates that ActAware and NrDec testlets share common variance, these subscales were subsequently combined into one testlet. This resulted in a satisfactory model fit (χ^2 (63)=70.76, *p*=.23), and unidimensionality of the full scale was clearly evident (Table 17, CHIME, Final). At this final stage, there was no DIF or significant misfit to the Rasch model, and reliability of the instrument was satisfactory (PSI=.82).

Figure 13 (bottom panel) shows the item-person threshold distribution of the full CHIME scale derived from the final analysis. It can be seen that items thresholds provide nearly full coverage (99%) of the sample's ability on the overarching latent trait of mindfulness. However, there are many item thresholds at the lower side of the scale, which are located outside of the sample distribution and indicate that easy items are overrepresented in the measure.

Conversion from Ordinal to Interval Scales

Table 18 and 19 include conversion algorithms to transform ordinal-scale scores to interval-level data for the full scale and all eight subscales of the CHIME using the original scale metric. Conversion tables are user friendly and include instructions provided as a table footnote (Tables 18 and 19). These conversion tables should only be used for respondents with no missing data. Conversions from the ordinal-to-interval level scale included here increase the precision of the CHIME and its subscales by applying the scoring algorithm without the need to alter the original response format of the measure.

Table 18. *Converting from ordinal (Raw) to Interval-level scores for the CHIME scale.*

Raw	Interval	Raw	Interval	Raw	Interval	Raw	Interval	Raw	Interval
37	37.00	68	112.76	99	121.55	130	129.42	207	161.58
38	54.47	69	113.14	100	121.79	131	129.69	208	162.55
39	65.58	70	113.51	101	122.03	132	129.96	209	163.58
40	72.67	71	113.86	102	122.27	133	130.23	210	164.71
41	77.84	72	114.19	103	122.52	134	130.50	211	165.92
42	81.92	73	114.54	104	122.79	135	130.79	212	167.30
43	85.29	74	114.86	105	123.03	136	131.06	213	168.81
44	88.14	75	115.16	106	123.27	137	131.36	214	170.56
45	90.62	76	115.48	107	123.51	138	131.63	215	172.56
46	92.81	77	115.75	108	123.76	139	131.93	216	174.93
47	94.75	78	116.05	109	124.03	140	132.22	217	177.79
48	96.50	79	116.34	110	124.27	141	132.49	218	181.37
49	98.09	80	116.64	111	124.51	142	132.79	219	186.12
50	99.52	81	116.91	112	124.76	143	133.09	220	192.91
51	100.87	82	117.18	113	125.02	144	133.38	221	203.96
52	102.08	83	117.45	114	125.27	145	133.68	222	222.00
53	103.19	84	117.72	115	125.51	146	134.00		
54	104.21	85	117.99	116	125.78	147	134.30		
55	105.16	86	118.26	117	126.02	148	134.60		
56	106.02	87	118.50	118	126.26	149	134.92		
57	106.83	88	118.77	119	126.53	150	135.22		
58	107.58	89	119.01	120	126.78	151	135.54		
59	108.26	90	119.28	121	127.05	152	135.86		
60	108.90	91	119.52	122	127.29	153	136.19		
61	109.50	92	119.77	123	127.56	154	136.51		
62	110.03	93	120.04	124	127.83	155	136.83		
63	110.55	94	120.28	125	128.07	156	137.16		
64	111.03	95	120.55	126	128.34	157	137.48		
65	111.52	96	120.79	127	128.61	158	137.83		
66	111.95	97	121.03	128	128.88	159	138.15		
67	112.35	98	121.28	129	129.15	160	138.50		

Note: Reverse code items 7, 10, 17, 19, 22, 26, 30, 33, 36, add all items together and find your ordinal score and a corresponding interval level score on the right-hand side.

Table 19. *Converging from ordinal (Raw) to Interval-level scores for the eight CHIME subscales.*

	AwareExt	***ActAware***	***AccNJ***		***NrDec***			***Openness***	***Relativity***	***Insight***	
Raw	**Interval**	**Interval**	**Raw**	**Interval**	**Raw**	**Interval**	**Raw**	**Interval**	**Interval**	**Raw**	**Interval**
4	4.00	4.00	5	5.00	6	6.00	4	4.00	4.00	5	5.00
5	5.89	6.23	6	7.55	7	9.53	5	6.12	8.35	6	7.06
6	7.23	7.84	7	9.44	8	11.75	6	7.61	10.71	7	8.47
7	8.17	9.01	8	10.85	9	13.14	7	8.66	11.99	8	9.44
8	8.94	9.97	9	11.96	10	14.17	8	9.51	12.80	9	10.21
9	9.61	10.79	10	12.90	11	15.01	9	10.25	13.39	10	10.86
10	10.24	11.52	11	13.72	12	15.73	10	10.93	13.87	11	11.44
11	10.84	12.20	12	14.46	13	16.39	11	11.56	14.30	12	11.98
12	11.45	12.84	13	15.16	14	17.00	12	12.17	14.72	13	12.50
13	12.07	13.46	14	15.82	15	17.59	13	12.77	15.15	14	13.00
14	12.72	14.06	15	16.47	16	18.17	14	13.36	15.61	15	13.51
15	13.41	14.66	16	17.11	17	18.76	15	13.96	16.10	16	14.05
16	14.15	15.27	17	17.73	18	19.35	16	14.59	16.62	17	14.61
17	14.94	15.89	18	18.36	19	19.95	17	15.24	17.19	18	15.20
18	15.79	16.55	19	18.98	20	20.58	18	15.95	17.82	19	15.84
19	16.71	17.27	20	19.60	21	21.22	19	16.73	18.51	20	16.53
20	17.72	18.07	21	20.23	22	21.89	20	17.61	19.30	21	17.26
21	18.84	19.01	22	20.88	23	22.57	21	18.65	20.18	22	18.05
22	20.14	20.17	23	21.57	24	23.28	22	19.92	21.18	23	18.88
23	21.82	21.77	24	22.29	25	24.01	23	21.65	22.44	24	19.77
24	24.00	24.00	25	23.08	26	24.76	24	24.00	24.00	25	20.73
			26	23.95	27	25.53				26	21.81
			27	24.95	28	26.32				27	23.07
			28	26.17	29	27.13				28	24.64
			29	27.80	30	27.97				29	26.87
			30	30.00	31	28.85				30	30.00
					32	29.79					
					33	30.83					
					34	32.07					
					35	33.75					
					36	36.00					

Note: Reverse code items 7, 10, 17, 19, 22, 26, 30, 33, 36, add items for each CHIME subscale together and find a corresponding interval level score on the right-hand side.

Discussion

The current study conducted Rasch analysis to examine the psychometric properties of the CHIME (Bergomi et al., 2014) and to produce ordinal-to-interval transformation tables that increase precision of the instrument. Overall, the findings of this study provide support for internal structural validity, unidimensionality, and acceptable reliability of the CHIME. Seven of the eight CHIME subscales demonstrated fit to the unidimensional Rasch model without the need for modifications. For the full scale and the remaining insightful understanding subscale, adequate model fit was achieved by minor modifications that involved combining items into testlets.

Despite the good psychometric properties of the CHIME, ordinal-to-interval conversion is still necessary to improve the precision of the instrument because individual CHIME items have varying degrees of difficulty (location) and hence contribute differently to the total subscale scores as can be seen in Table 16. Not considering item difficulty when calculating total scores is a common limitation of ordinal measures, as it increases measurement error and thus negatively affects accuracy of assessment (Bond & Fox, 2007; Norquist et al. 2004). Therefore, using the conversion algorithms presented in Tables 18 and 19, ordinal responses to CHIME can be transformed into interval-level data suitable for parametric statistics. This transformation can be computed for each subscale and the full scale as long as there are no missing data meaning that the effects of mindfulness and its specific aspects in context of MBIs can be explored with greater precision.

Compared to the recent Rasch analysis of the MAAS where disordered thresholds were corrected by uniform rescoring (Medvedev et al. 2016a), the CHIME items showed no significantly disordered thresholds, which supports utility of the current response format of the scale. Unlike the recent Rasch analysis of the multidimensional KIMS, which required removing five non-fitting items (Medvedev et al. 2016b), excellent fit in the current analysis of the CHIME was achieved without removing any items. Only one item displayed significant misfit to the Rasch model, which was resolved by addressing local dependency. Furthermore, no differences in item functioning were found across the person factors gender, age, and meditation experience, making the CHIME a very promising measure with a potentially wide range of applicability across different populations. Even though there were expected significant mean differences between meditators and non-meditators across all subscales and the full CHIME scale, the item-person plots (Figures 11-13) demonstrated that both groups can be adequately measured by the instrument. The only exception is the awareness of internal experiences that showed ceiling effects associated with meditation practice.

Limitations and Conclusions

Two of the eight subscales of the CHIME (acting with awareness and insightful understanding) exhibited reliability indexed by PSI slightly below .70, which limits their applicability at subscale level. However, after Rasch transformation the full CHIME scale has good reliability and met expectations of the unidimensional Rasch model. Item-person threshold distributions for the full CHIME scale and individual subscales suggest that easy items may be overrepresented in the measure. However, lack of individuals endorsing lower response options in the current sample does not affect precision of the CHIME if Rasch-transformed interval scores are used. Availability of easy items in the scale might be an advantage if the scale is applied to clinical populations where lower mindfulness scores may be expected (Park et al., 2013). Given that the CHIME covers a wide range of aspects of mindfulness and has demonstrated good psychometric properties, further psychometric work should be conducted to translate and validate the CHIME in languages other than German. So far, only the adolescent version of the CHIME has been adapted into English (Johnson, Burke, Brinkman, & Wade, 2016). Additionally, the scale will need to be validated with clinical populations, and the potential advantage of easy items for these sample groups can then be examined.

Mindfulness is an important predictor of psychological health and well-being, which requires accurate measurement of the construct. This study applied Rasch analysis to examine and enhance the psychometric properties of the CHIME, a multidimensional measure of mindfulness. The study findings support the internal structural validity and reliability of this measure and allows researchers to improve the precision of the instrument by using provided transformation algorithms. The CHIME is a new emerging self-report instrument with enhanced psychometric properties to measure eight aspects of mindfulness, which represents both a reliable and a valid alternative to existing mindfulness measures.

Chapter Seven. Rasch Analysis of the Perceived Stress Scale

Introduction

Stress is described as heightened emotional states associated with physiological changes (McEwen & Stellar, 1993; Helton & Näswall, 2015). Research has shown that extended exposure to stress can lead to negative health effects (Cohen et al., 1998; Hillhouse et al., 1991). Both, effective stress management and stress reduction are critical in reducing the negative effects of stress on an individual's health. Development of effective methods to manage and reduce stress requires accurate assessment of perceived stress levels to evaluate and compare the unique contributions of various predictors, situations and behaviours that may trigger and maintain stress. One of the first mindfulness-based treatment programs, MBSR (Kabat-Zinn, 1982, 1990) was specifically developed for stress reduction. However, to evaluate effectiveness of such and similar programs or its components reliably accurate assessment of both mindfulness and stress levels pre/post and during a MBI is required. In particular, precise assessment of perceived stress is critical because it reflects the subjective evaluation of environmental events (Bloch et al., 2004), which directly influences physiological responses responsible for adverse health effects (LeDoux, 2000; Medvedev et al., 2015).

The Perceived Stress Scale (PSS) (Cohen & Williamson, 1988) was constructed as a subjective measure of perceived stress to assess the extent to which a person's life is perceived as "unpredictable, uncontrollable, overloading," (p.387) relative to the individual's coping abilities. The scale is very widely used, approaching 12,000 citations by the beginning of 2017, according to Google Scholar. The original PSS version contains 14 items and has good internal consistency (Cronbach's alpha > .80) and satisfactory construct validity (Cohen & Williamson, 1988). The authors subjected the 14-item PSS to principal component analysis (PCA) and found four items that displayed poor loadings on the first principal component in the range of .11 - .39, which were removed resulting in the popular 10-item version of the PSS (called PSS-10). A four-item PSS version was also introduced as a quick assessment tool to be used in time-constrained situations. However, it has been established that the PSS-10 has better internal consistency (alpha=.78) compared to the 4-item PSS version (alpha=.60) and was recommended by the authors for use in future research (Cohen & Williamson, 1988). A recent factor analysis study using a community sample also reported that the factor structure of the four-item version of the PSS is problematic (Ingram, Clarke, & Lichtenberg, 2016). The

PSS-10 has been used with different populations for both clinical assessments and empirical investigations including validation studies, all confirming its satisfactory psychometric properties (Mitchell, Crane, & Kim, 2008; Roberti, Harrington, & Storch, 2006; Taylor, 2015). Currently, there is no agreement on the factor structure of the PSS-10, with some studies including the original validation report confirming unidimensionality (Cohen & Williamson, 1988; Cole, 1999), while others argue that a two-factor solution provides a better fit (Barbosa-Leiker et al., 2013; Taylor, 2015; Teh, Archer, Chang, & Chen, 2015). Thus, alternative method to test dimensionality of the PSS-10 can be useful to answer this question and is applied in the current investigation.

The psychometric properties of the PSS-10 have been investigated mainly using traditional psychometric methods with just a few exceptions that used item response theory (IRT) to investigate functioning of individual items (Cole, 1999; Sharp, Kimmel, Kee, Saltoun, & Chang, 2007; Taylor, 2015). IRT extends classical test theory because it provides more comprehensive analysis of a scale and individual item functioning, leading to improvement of measurement precision which is particularly important for health-related assessments (Allen & Yen, 1979; Thomas, 2011). IRT analysis is also useful as it controls for item bias or DIF. Cole (1999) used IRT to investigate potential item bias of the PSS-10 with a large US sample (*n*=2,264). Significant, but very small differences in performance of several items were reported by gender, ethnicity and education level categories, with the overall conclusion of satisfactory item performance. The study also provided evidence for unidimensionality of the PSS-10, which is a prerequisite for the unidimensional IRT model used in their analysis. Sharp et al. (2007) tested the performance of the PSS-10 items in a clinical sample of asthma patients and reported that few items function differently across ethnic and literacy factors, with an overall conclusion that there was no DIF. Recently, Taylor (2015) used the graded response IRT model and reported satisfactory functioning of individual items in a two-factor model of the PSS-10. Even though, these findings provide useful diagnostics of the individual PSS-10 items' function, they do not provide feasible solutions to improve the measurement precision of the overarching latent variable of perceived stress.

Although the PSS is an ordinal scale, researchers have used the PSS with parametric statistics (Gitchel, Roessler & Turner, 2011; Chavez-Korell & Torres, 2014), which may violate their fundamental assumptions given that the scale suitability for such techniques has not been thoroughly examined. It should be noted that an ordinal scale will not become an interval scale simply because of its popularity or by adding individual items scores together (Stucki et al., 1996; Allen & Yen, 1979). For instance, usage of the ordinal PSS-10 in research may affect comparisons with neurophysiological and biological data (e.g. heart rate, cortisol), an especially important consideration in stress and mindfulness research. Therefore, further research is necessary to improve the precision of the PSS-10 up to an interval level scale and to address structural validity issues, which can be achieved using Rasch analysis (Rasch 1960; Tennant & Conaghan, 2007).

In sum, it is assumed that the PSS-10 is a measure of perceived stress and has generally accepted psychometric properties that can be applied to a wide range of clinical and non-clinical populations. However, its ability to discriminate precisely between perceived stress levels has not been investigated in sufficient detail. Rasch analysis is a suitable method to investigate the ability of the scale and individual items to discriminate on their overarching latent factor. However, to the best of our knowledge, this advanced technique has not yet been applied to scrutinize and to improve the psychometric properties of the PSS. The present study aims to apply Rasch analysis to explore strategies to improve the psychometric properties of the PSS-10 up to an interval level of measurement. This psychometric investigation explicitly focuses on the measurement of the overarching latent factor of perceived stress, rather than its underlying facets already explored in the literature (Roberti et al., 2006; Taylor, 2015). To ensure that results are generalizable to diverse populations, a combined sample of respondents from New Zealand and the United States is used, consisting of university students as well as respondents from the general population. Additionally, the dataset is sufficiently large to allow splitting of the combined sample into two sets, thus allowing replication of the Rasch analysis of one half of the sample with the other half. To enable users of the scale to use parametric statistics with the PSS, an ordinal-to-interval conversion table was generated.

Method

Figure 14 shows a flow diagram of the steps used to arrive at ordinal-to-interval transformation algorithms for the PSS-10 scale. This includes accessing datasets of respondents from the New Zealand general population as well as datasets of New Zealand and US university students. From these three datasets, randomly selected respondents were extracted to create two samples of n=450 in which respondents from each dataset were equally represented. Each stage of the diagram is explained in greater detail in the subsequent sections of this report.

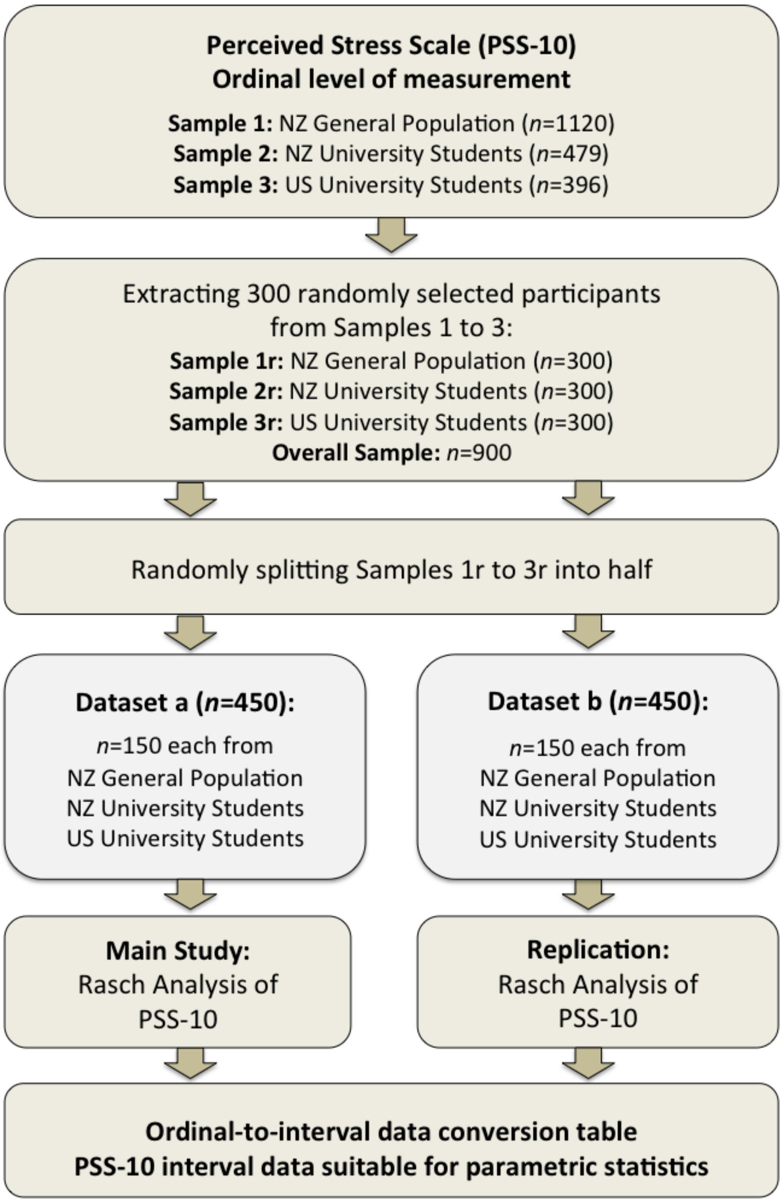

Perceived Stress Scale (PSS-10)
Ordinal level of measurement
Sample 1: NZ General Population (n=1120)
Sample 2: NZ University Students (n=479)
Sample 3: US University Students (n=396)
Extracting 300 randomly selected participants from Samples 1 to 3:
Sample 1r: NZ General Population (n=300)
Sample 2r: NZ University Students (n=300)
Sample 3r: US University Students (n=300)
Overall Sample: n=900
Randomly splitting Samples 1r to 3r into half
Dataset a (n=450):
n=150 each from
NZ General Population
NZ University Students
US University Students
Dataset b (n=450):
n=150 each from
NZ General Population
NZ University Students
US University Students
Main Study:
Rasch Analysis of PSS-10
Replication:
Rasch Analysis of PSS-10
Ordinal-to-interval data conversion table
PSS-10 interval data suitable for parametric statistics

Figure 14. Flow chart outlining the steps involved in the present Rasch analysis to create ordinal-to-interval transformation algorithms for the PSS-10.

Participants

The present study combined three independently collected samples (Figure 14). Sample 1 (n=1,102) consisted of Auckland (New Zealand) residents who participated in a postal survey on noise sensitivity and health (Hill et al., 2014). The mean age was 51 (SD=16.42), and 65% were female. Sample 2 (n=479) contained university students enrolled in various health science courses at the University of Auckland and Auckland University of Technology. The majority (76%) were female, and the overall mean age was 19.96 (SD=4.47). Sample 3 (n=396) consisted of students enrolled in first-year psychology classes at West Chester University in the United States. The mean age was

19.18 (SD=2.21), and 45% were female. From each of these three samples, 300 participants were randomly selected to create an overall sample of 900 participants from the New Zealand general population, New Zealand university students, and US university students. Each subset was randomly divided in half to create two samples of 450 participants, where 150 participants were included from each sample population (Figure 14). As the ethnic profile for the New Zealand and US samples was very different, common categories were created so that DIF by ethnicity could be compared. In the overall sample of 900, 65% were classified as Caucasian, 5% as Polynesian, 9% as Asian, and 19% as other.

Procedure

The procedure for collecting the data for the New Zealand general population (Sample 1) is reported in detail elsewhere (Hill et al., 2014). Auckland residents living in roads with various levels of traffic flow received a questionnaire on noise sensitivity, perceived stress, and health, which they subsequently posted back to the researchers using a self-addressed pre-paid envelope. Participants in Sample 2 were university students enrolled in health science courses at two major universities in New Zealand completing a survey on motivation to learn, quality of life, and perceived stress. Students at the University of Auckland received an invitation to complete an online survey and thus responded at a time of their convenience. Students at Auckland University of Technology were approached in lecture theatres (with permission from lecturers) and completed the questionnaire during the lecture break or after the lecture. For Sample 3, students taking an introductory psychology class at West Chester University in Pennsylvania, United States, completed a research study on motivation to learn, quality of life and perceived stress as an option in fulfilling their class research credit.

While the questionnaires in the three above-mentioned studies each contained various scales, the present study focused on evaluating the PSS-10 only. The PSS-10 is a 10-item self-report questionnaire of perceived stress operationalized as subjective evaluation of lack of control, unpredictability and overload in participants' daily life (Cohen & Williamson, 1988). The instrument uses a five-point Likert-scale response format (1="Never" to 5="Very often"), and a total score is calculated after reverse-coding items 4, 5, 7 and 8 and then adding scores of all ten items together.

Data Analysis

Descriptive statistics, reliability, and exploratory PCA of the PSS-10 were computed using IBM SPSS v.22 for subsequent comparisons with outcomes of Rasch analysis. Then, data were formatted and saved as an ASCII file to satisfy the requirements of the RUMM2030 software for Rasch analysis (Andrich et al., 2009). The initial output analysis was subjected to a likelihood-ratio test, which indicated the appropriateness of the unrestricted (Partial-Credit) version of the Rasch model for the current dataset. The Rasch analysis was conducted following the sequential stages described in Chapter Two.

As outlined in Figure 14, two equal sized datasets (a and b) were created from an overall dataset (n=900). The first sample (a) of 450 was used for the main Rasch analysis, which was then replicated using the second half of the sample (b). Both samples (Datasets a and b) contained a large enough number to satisfy the recommended sample size estimates for the Rasch analysis (Linacre, 1994) and to allow investigation of DIF.

Results

Cronbach's alpha for the PSS-10 with the current data set (N=900) was .88 indicating good internal consistency with all 10 items having item-to-total correlations in the range from .49 to .74 (Table 20). PCA (principal axis factoring) extracted two factors with an eigenvalue > 1.00, but most of the variance (48.84%) was explained by the first principal component. Also, all PSS-10 items showed high loadings on the first principal component ranging from .58 to .81, which would be expected for a unidimensional measure.

Table 21 provides a summary of the Rasch model fit statistics for the initial and the final analysis including both the overall and the individual item fit indices for the sample (a) and the overall fit indices for replication of the analysis with sample (b). Overall, the person location mean was within the acceptable range in both samples suggesting good targeting of the sample by the scale items (Table 21). Also, the scale showed good person separation reliability of .88. However, in both samples the initial overall fit to the Rasch model was affected by significant item-trait interaction indexed by chi-square test: $\chi^2(90)=129.26$ for sample (a) and $\chi^2(90)=150.62$ for sample (b), $p <.001$ (Table 21). This means that the scale cannot adequately discriminate between respondents at different levels of the latent trait (perceived stress). Although, no items with disordered thresholds were identified, Item 10 displayed significant misfit to the Rasch model with fit residual below the acceptable cut-off point of -2.50 (Table 20), in in both samples. Also, Item 4 showed deviation from the model expectations with a positive fit residual above 2.50 in sample (b) but not in sample (a).

Table 20. *Corrected item-to-total correlation and loadings on the first principal component (PC) for PSS-10 items (n=900).*

	Item	Item-to-total correlation	Item Loadings on the 1st PC
1	Been upset	.62	.70
2	Felt unable to control	.67	.75
3	Felt nervous and stressed	.66	.74
4	Felt confident	.52	.61
5	Things are going your way	.63	.71
6	Could not cope	.59	.68
7	Able to control irritations	.49	.58
8	Felt on top of things	.63	.71
9	Been angered	.57	.66
10	Could not overcome	.74	.81

Table 21. *Rasch model fit statistics for the initial (1a) and final (2a) analysis of the PSS-10 (n=450), and its subsequent replication 1b and 2b, respectively (n=450).*

	Item residual		Person residual		Goodness of fit		Person	Independent *t*-test	
Tests	**Value**	**SD**	**Value**	**SD**	**χ2 (df)**	***p***	**separation**	**%**	**^{a}LB**
1a	**0.18**	**1.92**	**-0.45**	**1.44**	**129 (90)**	**.004**	**0.88**	**10.67**	**8.65**
1b	**0.11**	**1.89**	**-0.48**	**1.47**	**150 (90)**	**<.001**	**0.88**	**12.67**	**10.65**
2a	**0.22**	**1.97**	**-0.54**	**1.05**	**29 (27)**	**.365**	**0.80**	**5.19**	**3.16**
2b	**0.18**	**1.64**	**-0.57**	**1.11**	**36 (27)**	**.11**	**0.81**	**4.74**	**2.71**

Note: ^{a}LB = lower bound of the 95-% confidence interval.

Given the inconsistent fits of Item 4, the residual correlation matrix was examined because local dependency between items affects both discrimination parameters and test information associated with the Rasch model fit (Lundgren Nilsson et al., 2013).

Local dependency

The residual correlation matrix was inspected for residual correlations higher or at the level of the cut-off point of .20 above the mean of all residual correlations. Correlations above this value were found for items 1, 2 and 9; items 4, 5, 7, and 8; and items 6, 10 and 3; indicating local dependency between those items. To verify this observation, a Spearman’s correlation matrix between all PSS-10 items was generated and confirmed higher correlations between these items in the range of .50 to .60. These observations together confirmed three groups of locally dependent items, which were combined into three subtests to address local dependency (Lundgren Nilsson et al., 2013). These minor modifications provided an alternative reliable solution (PSI=.80-.81) to satisfy the expectations of the Rasch model in both samples (χ^2(2(27)=29.92, *p*=.36 (a), χ^2(2(27)=36.30, *p* =.11 (b), Table 21). At this stage, good model fit was also achieved at the individual item level without the need to remove any of the PSS-10 items.

Differential item functioning (DIF)

ANOVA indicated significant DIF effects by sample on the subtest one (*F*(2)=6.27, *p*=.002, Bonferoni adjusted *p*=.006), but not for the subtests two (*F*(2)=3.46, *p*>.006) and three (*F*(2)=0.36, *p*>.006). However, post-hoc comparisons between sample groups showed no significant difference with *p* range from .065 to .217. There were no significant DIFs in functioning of subtest items due to other personal factors including gender, age, ethnic groups, and education levels.

Test for unidimensionality

To test the unidimensionality of the final model solution, the person estimates from the subtest with the highest positive loadings on the first principal component were compared with the estimates from the subtest with the highest negative loadings. Unidimensionality was confirmed for both samples with 5.19% significant t-tests overlapping 5% cut-off point on the lower bound of the confidence interval (3.16%) for sample (a) and 4.74% of significant *t*-tests and cut-off overlap of 2.71% for sample (b) (Table 21).

Item-person threshold distribution

Figure 15 shows the person-item threshold distribution of the modified PSS-10 after combining locally dependent items into three subtests (Table 21, analysis 2a). Here, subtest item thresholds and person ability levels on the latent factor measured by the PSS-10 are plotted using the same metric in logit units. Distribution of persons is close to normal and the modified PSS-10 item thresholds satisfactorily cover 98% of the participants’ abilities on the latent factor (perceived stress).

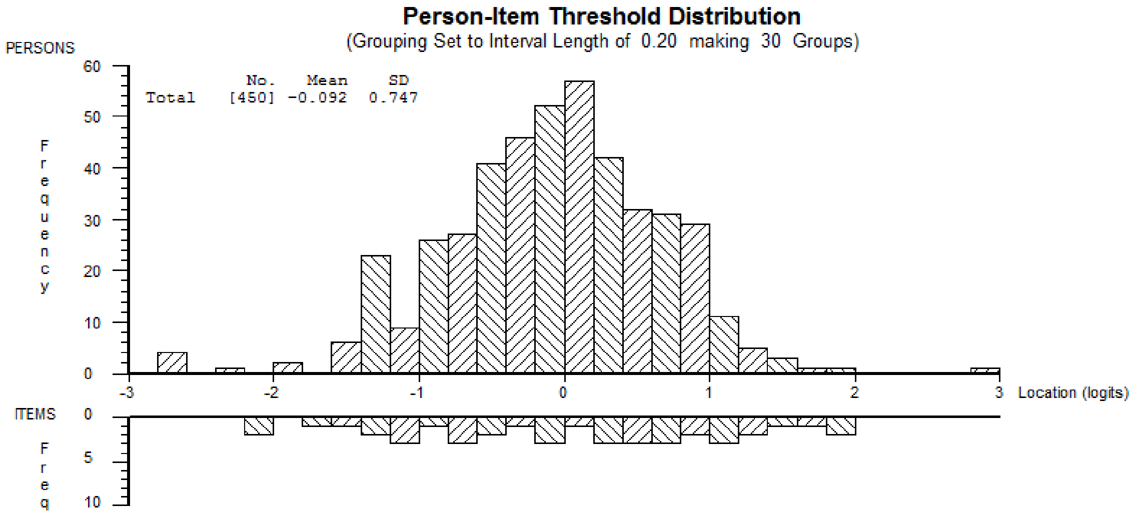


Figure 15. Person-item threshold distribution for the PSS-10 (Table 21, 2a).

Equating t-test

The means of person estimates from the original PSS-10 and the modified version were compared by a paired-samples *t*-test. The difference between the person estimates of the two versions was significant (*t*(449)=7.22, *p*<.01), indicating successful alteration of the ability of the final model to discriminate between individual levels of perceived stress compared to the original version. This confirms that the implemented modifications (subtests) resulted in an improved solution for the PSS.

Ordinal-to-interval conversion table

Table 22 provides interval scores in both logit units and the original PSS-10 scale format that allows researchers to convert ordinal raw scores to interval-level scores. Researchers who have already used the PSS-10 to collect data or are planning to use the scale can apply the results of this study as follows: Calculate the raw score by reverse-scoring questions 4, 5, 7 and 8 and then adding scores of all 10 items together. Next, use Table 3 to convert these scores to the corresponding interval-scale scores ranging from 10 to 50, identical to the range of scores of the original PSS scoring system. By using the conversion table provided here, users are able to increase the precision of the PSS-10. It should be noted that conversion from the ordinal-to-interval scale proposed here does not require altering the original response format of the PSS-10 scale. This conversion table was independently generated with the second sample (b) (*n* = 450), showing almost identical results.

Table 22. *Converting from a raw PSS-10 score (10 to 50) to an interval scale in logit units and the original scale metrics.*

Ordinal measure	Interval measure		Ordinal measure	Interval measure	
Raw score	**Logit**	**Scale**	**Raw score**	**Logit**	**Scale**
10	-3.34	10.00	31	0.19	32.71
11	-2.66	14.37	32	0.28	33.28
12	-2.21	17.23	33	0.37	33.85
13	-1.92	19.11	34	0.45	34.40
14	-1.70	20.53	35	0.54	34.96
15	-1.52	21.68	36	0.63	35.52
16	-1.37	22.66	37	0.71	36.08
17	-1.24	23.53	38	0.80	36.66
18	-1.11	24.32	39	0.90	37.25
19	-1.00	25.06	40	0.99	37.87
20	-0.89	25.77	41	1.09	38.52
21	-0.78	26.45	42	1.20	39.21
22	-0.68	27.12	43	1.31	39.93
23	-0.58	27.77	44	1.43	40.69
24	-0.48	28.43	45	1.56	41.52
25	-0.38	29.06	46	1.70	42.45
26	-0.28	29.70	47	1.87	43.51
27	-0.18	30.33	48	2.08	44.88
28	-0.09	30.94	49	2.40	46.93
29	0.01	31.54	50	2.88	50.00
30	0.10	32.13			

Note: The raw score is calculated by reverse-scoring questions 4, 5, 7 and 8 and then adding scores of all 10 items together. This table cannot be used for respondents with missing data.

Discussion

The PSS is a widely used instrument to measure perceived stress on an ordinal scale; however its precision has not been yet fully optimized. The current study used strategies of modern Rasch analysis to improve the psychometric properties and precision of the 10-item PSS up to an interval level scale. This Rasch analysis contributed to the limited number of IRT-based studies (Cole, 1999; Sharp et al., 2007; Taylor, 2015) that focused on the functioning of individual PSS-10 items by increasing the precision of the PSS-10 and addressing both local dependency and DIF. Using a subtests approach similar to Lundgren Nilsson et al. (2013), good model fit was achieved after combining locally dependent items into three subtests, and no systematic DIF by personal factor such as gender, ethnicity, education and sample population was evident. After these minor modifications, the psychometric properties of the PSS-10 are robust, and transformation from an ordinal to an interval level scale can be conducted using the conversion algorithm provided in Table 22.

Local dependency found between items 4, 5, 7, and 8 is consistent with earlier research (Roberti et al., 2006; Taylor, 2015), where the same items were proposed as a second factor in a two-factor PSS-10 solution. All these four items are negatively worded and thus measure coping abilities as opposed to perceived stress, which could explain previous findings of these items loading together as a factor (Roberti et al., 2006; Taylor, 2015). However, after combining locally dependent items into subtests, unidimensionality of the PSS-10 was clearly evident in both the main (a) and the replication (b) samples, suggesting that after Rasch modifications the PSS-10 is clearly tapping into one overarching latent factor namely perceived stress. Local dependency found between items 1, 2 and 9 is not surprising because these items all contain themes of control over external events. Another subtest included items 3, 6, and 10, which are explicitly related to perceived helplessness (e.g. could not overcome or cope) and hence explain local dependency. These clusters of locally dependent items may have influenced variability of the earlier factor analysis (Roberti et al., 2006; Taylor, 2015) and may have even generated spurious factors (Lundgren Nilsson et al., 2013). However, the clear evidence of unidimensionality of the PSS-10 based on three subtests replicated by two random samples together with the large amount of shared variance suggest that a total PSS-10 interval score reflects perceived stress levels of the majority of people.

The following limitations are acknowledged. The samples may not reflect the full diversity of New Zealand’s or the US’s ethnic groups and no efforts were made to purposively sample under-represented groups. The response rate of 15% for the New Zealand general population sample (Hill et al., 2014) was low, which could reflect a self-selection bias. However, such response rates are not uncommon for research of this nature in New Zealand (Krägeloh et al., 2013).

The main contribution of this study is that the PSS-10 raw score can now be converted from an ordinal scale to an interval scale, which means that parametric statistics can be conducted without violating their fundamental assumptions. The interval-level estimates of the latent factor offer researchers the opportunity to examine the effects of mediators and moderators of perceived stress in various contexts. Rasch interval transformed PSS-10 scores can reliably be used in such models given that potential item biases (DIF) and local dependency issues are ultimately resolved by Rasch analysis if data fits the model expectations. The improved precision of the instrument is also highly desirable in clinical assessment, where it informs accuracy of diagnosis and treatment of stress-related conditions. It is important to note that these improvements were possible without the need to alter the original PSS-10 response format meaning that existing datasets can easily be re-analysed to provide interval-level measurement. The modified PSS-10 satisfactory covers 98% of the sample abilities, however, there are still a few individuals uncovered by item thresholds at both the upper and lower level of the scale. Future studies may consider exploring whether using more response options with extreme categories will provide better coverage for individuals with lower and higher stress levels.

Conclusion

Stress may affect both physical and mental health and its accurate assessment represents an ongoing challenge. The current study has demonstrated that after minor modifications the widely used perceived stress measure PSS-10 satisfies the expectations of the unidimensional Rasch measurement model. The precision of the PSS-10 can be optimized up to an interval level scale by using the ordinal-to-interval conversion tables published here. The current study focused on the measurement of the overarching latent factor of perceived stress, rather than investigating its underlying facets and, therefore, is best assessed from the perspective of the modern IRT and specifically the Rasch model.

Chapter Eight. The Oxford Happiness Questionnaire: Rasch Analysis

Introduction

Mindfulness practice was found beneficial to psychological well-being (Josefsson et al., 2014; Bennet & Dorje, 2015) and trait mindfulness was identified as a major predictor in various models of psychological well-being (Brown and Kasser 2005; Pearson et al. 2015). Therefore, scientific studies investigating the relationship between mindfulness and subjective well-being, which is widely used as synonymous with happiness, require reliable and valid measurement tools to assess both. This is also important for accurate assessment of psychological and cognitive changes in individuals undergoing MBIs. Reliable comparisons between mindfulness and well-being measures require that both measures should be at least interval level of measurement, which can be achieved using Rasch analysis (Rasch, 1961; Tennant & Conaghan, 2007). While earlier Chapters focused on enhancement of widely used mindfulness measures this Chapter focuses on investigating and improving psychometric properties of a widely used well-being measure, the OHQ, by applying Rasch analysis.

Happiness has been the most important goal for humans throughout history (Compton, 2005). Aristotle considered his *eudaemonia*, usually translated as ‘happiness’, to be the ultimate goal of humans, and superior to all other goals (Diener, Sapyta, & Suh, 1999). Eudaemonic happiness was suggested to be related to psychosocial functioning and distinct, albeit correlated, with pleasure-driven hedonic happiness (Joshanloo, 2015). Cross-cultural research conducted in 47 countries indicated that happiness is rated higher than all other personal values such as health, love or wealth (Kim-Prieto, Diener, Tamir, Scollon, & Diener, 2005). Happiness has also been found to be a highly valued component of quality of life, superior to other values such as money, health or sex (Skevington, MacArthur, & Somerset, 1997). Therefore, happiness can be considered the most desirable condition among humans, and other goals may only be valued as potential determinants of happiness (Csikszentmihaliy, 1992). From an evolutionary perspective, happiness can be seen as a psychological reward for adaptive functioning associated with evolutionary fitness (Nesse, 1990).

The early tendency of psychological research was to focus mainly on mental illness and abnormalities associated with social or occupational dysfunction (Argyle, 2001; Carr, 2004) and related cognitive distortions (Beck, 1991; Ellis, 2002). Towards the end of the 20th century, psychologists started to display an increased interest in the positive dimensions of human life (e.g., well-being, happiness and quality of life), reflected by the growing body of research focused on these constructs (Argyle, 2001; Diener, 1984). Happiness and subjective well-being are often used as synonyms (Diener, 1984, 2006) and became mainstream in economic research contexts (Kristoffersen, 2010). Specifically, subjective well-being data are now widely used and studied along with economic indicators (Kahneman & Krueger, 2006; Spruk & Kešeljević, 2015). For instance, in the past few years, the Organisation for Economic Co-operation and Development (OECD) has been measuring and reporting the Better Life Index (OECD, 2015), indicating that subjective well-being or happiness plays a crucial role in the life of individuals living in developed nations. However, scientific study of happiness requires accurate measurement of the construct that satisfies assumptions of parametric statistics and thus allows both researchers and clinicians to make reliable and valid comparisons with the relevant data sources - such as comparing the mean values from different sample groups.

Theoretically, happiness can be explained by bottom-up processes representing the affective component of happiness, and top-down processes, relating to cognitive components (Andrews & McKennell, 1980). According to the bottom-up approach, happiness is evaluated as a total sum of aggregated positive and negative feelings (Diener, 1984). It should be noted that accurate measurement should include both positive and negative affect because a number of studies have suggested that positive affect is not the opposite of negative affect (Andrews & McKennell, 1980; Argyle, 2001; Brandburn, 1969), and the correlation between them is only moderate (Tellegen et al., 1988). On the other hand, top-down approaches suggest that happiness is largely a product of an individual's cognitions and refers to subjective evaluations of one's experiences and expressions of life satisfaction (Andrews & McKennell, 1980; Diener, 1984). Even though the top-down approach has been well argued with some supporting evidence (Andrews & McKennell, 1980; Beck, 1991; Diener, 1984), it has been shown that the subjective evaluation of life events influences emotional responses (LeDoux, 2000; Medvedev et al., 2015) meaning that the approaches are likely to complement each other. Accordingly, there is research evidence indicating a single dimension of happiness that includes positive affective, negative affective and cognitive components (Argyle, 2001; Hills & Argyle, 1998, 2002; Joseph & Lewis, 1998). Cognitive components may also include personality traits like internal locus of control, extraversion and optimism (Carr, 2004; Fordyce, 1988; Mayers, 1992).

Different measures have been developed to assess happiness, however, operational definitions of these instruments appear inconsistent (Andrews & McKennell, 1980; Argyle, 2001; Brandburn, 1969). Inconsistency in defining subjective happiness and well-being constructs, and considering their relevant components, has resulted in various limitations of existing happiness measures (Eid & Larsen, 2008). The most common limitation is not considering both cognitive and affective facets as suggested by theory and research (Andrews & McKennell, 1980; Argyle, 2001; Diener, 1984; Fordyce, 1988).

The 29-item Oxford Happiness Questionnaire (OHQ) (Hills & Argyle, 2002) is a widely-used scale to assess personal happiness. It is a new version of the original Oxford Happiness Inventory (OHI) (Argyle, Martin, & Crossland, 1989), which contains 29 items, each accompanied by four statements representing response options similar to the Beck Depression Inventory (Argyle et al., 1989; Beck, Steer, & Brown, 1996). The OHI has been scrutinized using Mokken scaling analysis probing whether questionnaire items can correspond reliably to the range of individual abilities on a latent trait (Stewart, Watson, Clark, Ebmeier, & Deary, 2010). The authors concluded that out of 29 items only 12 items could reliably reflect individual happiness levels.

Hills and Argyle (2002) used the terms well-being, subjective well-being, and psychological well-being as synonymous to happiness in describing the OHI and OHQ. Both measures are based on theoretical considerations supported by research findings indicating a single dimension of happiness that covers positive and negative affect, and cognitive evaluations such as life satisfaction and happy traits (Andrews & McKennell, 1980; Argyle, 2001; Diener, 1984). Furthermore, there are items reflecting specific cognitive components and traits found within the single happiness factor labeled as sociability, sense of control, physical fitness, positive cognition, mental alertness, self-esteem, cheerfulness, optimism and empathy (Hills & Argyle, 1998, 2002). From a theoretical perspective, the item content of the OHQ was criticized for being too broad – spreading over many areas of human experience rather than specifically focusing on happiness (Kashdan, 2004).

Hills and Argyle (2002) validated the 29-item OHQ using a 6-point Likert-scale response format worded as *strongly disagree*, *moderately disagree*, *slightly disagree*, *slightly agree*, *moderately agree*, and *strongly agree* and reported a Cronbach's alpha just above 0.90. A five-point response format worded as *agree strongly, agree, uncertain, disagree, and disagree strongly* has also been used for the OHQ, and comparable Cronbach's alpha of 0.90 was reported (Robbins, Francis, & Edwards, 2010). Due to the large number of factors extracted using the Kaiser criterion in principal component factor analysis (eigenvalue > 1.00), Hills and Argyle (2002) used the Direct Oblimin rotation method, which extracted only one second-order component that was interpreted as evidence of unidimensionality.

The OHQ is an ordinal scale and as such it should not be used with parametric statistics such as ANOVA without violating fundamental assumptions of these tests. Also, usage of the ordinal OHQ in research may compromise the validity of comparisons with interval level data. Therefore, further research is necessary to improve psychometric properties of the OHQ up to an interval-level measure, and in particular Rasch analysis, a method that can be used for this purpose (Rasch, 1960; Tennant & Conaghan, 2007).

Taken together, the OHQ (Hills & Argyle, 2002) has generally accepted psychometric properties, but its item functioning and ability to discriminate precisely between levels of subjective wellbeing has not been sufficiently investigated. Rasch analysis is a suitable technique to test structural validity and to improve the ability of an instrument and individual items to discriminate on their overarching latent trait. However, to the best of our knowledge, Rasch analysis has not yet been used to investigate and to advance the psychometric properties of the OHQ. The present study aimed to apply modern Rasch strategies to improve the psychometric properties of the OHQ up to an interval-level scale. Thus, to enable researchers to analyse the OHQ data using parametric statistics and to make valid comparisons with neurophysiological data ordinal-to-interval conversion tables were produced.

Method

Participants

The sample was based on the recommended optimal sample size estimates for most purposes of Rasch analysis (Linacre, 1994) and included 281 university students from two countries: 180 from Auckland University of Technology, New Zealand, and 101 from the University of Nottingham, Malaysia Campus. The New Zealand sample includes 35 males (20%), 141 females (80%) and 4 participants who did not provide gender information, and the Malaysian sample had 48 (48%) males and 53 (52%) females. The mean age for the pooled sample was 23.70 years, with a standard deviation of 7.32. Ethnicities of the combined sample include 86 (31%) Caucasian, 36 (13%) Polynesian, 105 (37%) Asian, and 45 (16%) of other ethnic groups. Personal factors such as sample category, gender, ethnic group and age were used to investigate DIF in Rasch analysis. Three age categories, each representing approximately a third of the age distribution in the sample, were created as follows: 16 to 18 years, 21 to 23 years and 24 to 64 years. Less than 1 percent of the data was missing.

Instruments

The 29-item OHQ (Hills & Argyle, 2002) is a self-report questionnaire that employs a 6-point Likert scale response format from *strongly disagree* = 1 to *strongly agree* = 6, with the higher scores corresponding to higher levels of happiness (Appendix C4). There are 12 negatively worded items that require reverse coding before calculating the total happiness score, which is a sum of individual item scores. Examples of positively worded items include 'I am very happy' and negatively worded items 'I rarely wake up feeling rested'. According to Hills and Argyle (2002), the OHQ is supposed to measure personal happiness as a broad unidimensional construct and has high internal consistency, with Cronbach's alpha at the level of 0.90 and above.

Procedure

The study questionnaires were completed by the New Zealand participants in lecture theatres of Auckland University of Technology before the start of lectures. The data from the Malaysian sample were obtained before the start of a lecture at the University of Nottingham's Malaysia campus and as part of a mindfulness and subjective wellbeing neuroscience research study. The Malaysian sample had a good standard of English and all scored above 6 points on the International English Language Testing System or 79 points on the Test of English as a Foreign Language (IBT) as part of their entry requirement for their studies. Both studies complied with local and international ethical guidelines and were approved by the ethics committees of the Department of Psychology, Faculty of Health and Environmental Sciences, Auckland University of Technology and the School of Psychology, Faculty of Science, University of Nottingham.

Data Analysis

Descriptive statistics, internal reliability, and exploratory PCA of the 29-item OHQ were computed using IBM SPSS v.22. Data were then formatted and saved as an ASCII file and imported into the RUMM2030 software for Rasch analysis (Andrich et al., 2009). A likelihood-ratio test was performed on the initial output analysis and indicated the appropriateness of the unrestricted (Partial-Credit) version of the Rasch model for the current dataset. Rasch analysis followed the sequential stages explained in Chapter Two and elsewhere (Siegert et al., 2010):

Results

Internal consistency of the 29-item OHQ with the current dataset was high (Cronbach's alpha .89), consistent with the original study (Hills & Argyle, 2002). However, five items (2, 5, 7, 14 and 23) displayed low item-to-total correlations (<.30), and removing any of these items did not result in any noticeable improvement of Cronbach's alpha.

Though not a principal aim of this study, a significantly higher mean happiness level was found for Malaysian students (128.78) compared to New Zealand students (121.20), Mann Whitney U test, U = 6371.50, p = .01.

Initial Test of the Overall Rasch Model Fit

Initial analysis indicated good reliability of the original scale (PSI=.90), but the overall fit to the Rasch model was unsatisfactory (χ^2(87)= 314.37, p < .001, Table 23, Analysis 1). In Rasch analysis, chi square for item-trait interaction should be non-significant meaning that measurement ability is independent from the level of the latent trait possessed by a person. Furthermore, a substantial number of the OHQ items displayed disordered thresholds (items: 1, 3, 6, 7, 9, 10, 11, 16, 23, 24, 27, 28, 29), and threshold ordering of other items was only marginally acceptable. Therefore, items were rescored to correct the thresholds before the analysis continued.

Table 23. *Rasch model fit statistics for the original OHQ version (1), after uniform item rescoring (2), after removing items 2, 5, 14 and 23 (3), and the final solution (4).*

	Item residual		**Person residual**		**Goodness of fit**		**PSI**	**Independent *t*-test**	
Tests	**Value**	**SD**	**Value**	**SD**	**χ2 (df)**	***p***		**%**	**LB[a] 95% CI**
1	**0.95**	**2.18**	**-0.13**	**1.75**	**314.37 (87)**	**<.001**	**.90**	**28.57**	**26.02**
2	**0.31**	**1.62**	**-0.30**	**1.73**	**222.18 (87)**	**<.001**	**.89**	**24.64**	**22.09**
3	**0.26**	**1.59**	**-0.33**	**1.69**	**152.12 (75)**	**<.001**	**.89**	**21.79**	**19.23**
4	**0.16**	**0.81**	**-0.45**	**0.99**	**13.03 (15)**	**.60**	**.82**	**4.56**	**1.73**

Note: [a]LB = lower bound of the 95-% confidence interval.

Rescoring of the OHQ Items

The OHQ items were rescored in an iterative way with subsequent goodness of fit testing. The best fit and optimal ordering of thresholds was achieved when using uniform rescoring of all items by collapsing response categories 'moderately disagree' with 'slightly disagree', and 'slightly agree' with 'moderately agree'. Figure 16 (top panel) shows disordered thresholds on the response probability curves for Item 1 typical for the original OHQ items, and bottom panel illustrates the effect of rescoring for the same item, which is representative for all other items.

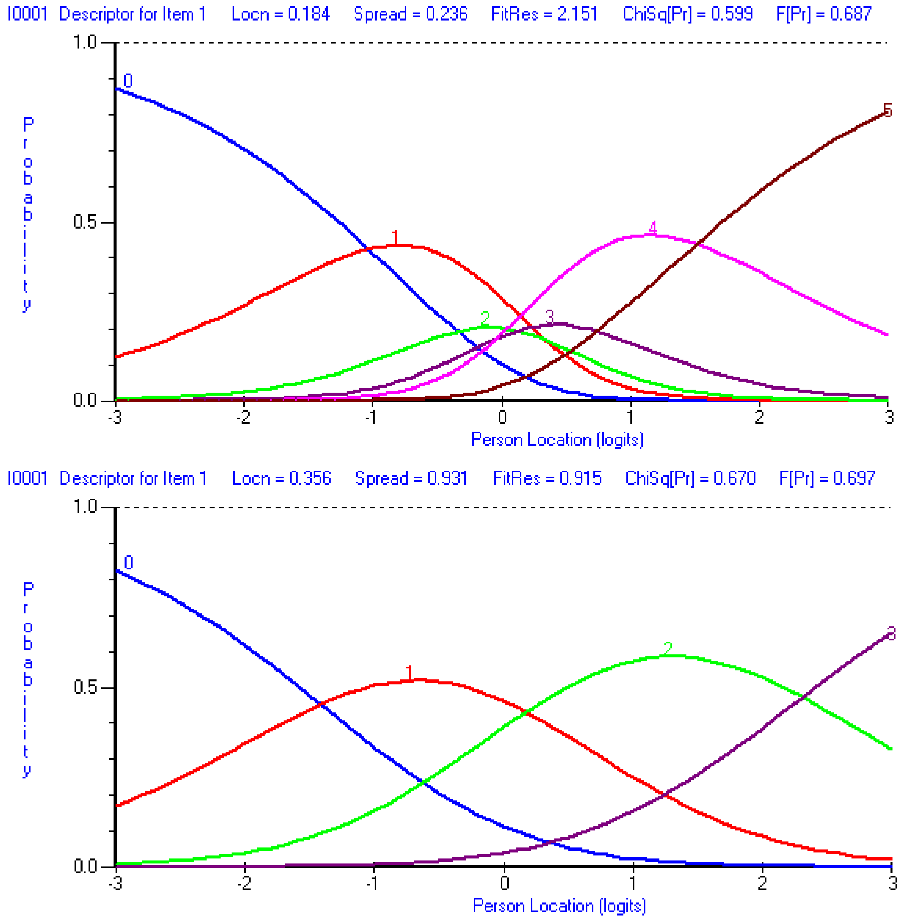


Figure 16. Response probability curves for item 1 of the OHQ before rescoring (top panel) and after rescoring (bottom panel).

All disordered threshholds were corrected after uniform rescoring, and the overall model fit was improved, although it remained below acceptable levels (χ^2(87)= 222.19, p < .001, Table 23, Analysis 2).

Removing Items Not Fitting to the Model

After uniform rescoring, fit residuals of all individual items were analyzed. Table 24 shows item fit residuals and chi-square values after uniform rescoring together with item-to-total correlations and loadings on the first principal component for the original OHQ version. Item 5 ‘I rarely wake up feeling rested’ and Item 23 ‘I do not find it easy to make decisions’ showed highest fit residuals above 3.00 indicating a poor fit to the Rasch model. Item-to-total correlations and loadings on the first principal component for these items were below .30 suggesting poor relationships with the latent trait (Table 24). Both items were removed, with subsequent testing of the overall model fit. Noticeable improvement was evident but expectations of the model were still not reached (χ^2(81)= 192.79, *p*<.001, PSI=.89).

Table 24. *Corrected item-to-total correlation and loadings on the first principal component (PC) for the original OHQ, and item-fit residuals and chi-square values after uniform rescoring (1) and after removing items 5 and 23 (2).*

		Item-to-	**Item**	**Item-fit**		**Item Chi-**	
		total	**Loadings**	**Residuals**		**Square**	
	Item	**correlation**	**on the 1st PC**	**(1)**	**(2)**	**(1)**	**(2)**
1	pleased with self [R]	0.45	0.44	0.92	1.14	1.55	3.88
2	**interested in others**	**0.20**	**0.21**	**2.52**	**2.94**	**15.46**	**17.96**
3	life is rewarding	0.45	0.49	-0.15	-0.12	0.93	1.00
4	warmth for others	0.48	0.53	-0.17	-0.20	2.23	2.49
5	**wake up rested [R]**	**0.30**	**0.28**	**3.80**	**-**	**22.61**	**-**
6	optimistic [R]	0.47	0.46	0.19	0.60	3.77	6.00
7	find things amusing	0.28	0.31	1.26	1.36	5.09	6.47
8	committed and involved	0.41	0.44	0.64	0.89	1.42	4.07
9	life is good	0.59	0.65	-1.40	-1.47	6.85	7.32
10	world is good [R]	0.35	0.34	2.98	3.40	6.45	10.24
11	laugh a lot	0.37	0.45	1.27	1.03	0.29	0.09
12	satisfied with life	0.61	0.69	-1.18	-1.28	11.20	12.96
13	look attractive [R]	0.50	0.51	-0.35	-0.12	1.03	0.97
14	**done things wanted [R]**	**0.28**	**0.28**	**2.36**	**3.21**	**9.92**	**17.56**
15	very happy	0.69	0.76	-2.29	-2.44	19.11	21.53
16	find beauty in things	0.37	0.42	-0.04	-0.02	2.06	3.80
17	cheerful effect on others	0.39	0.46	0.12	0.09	1.58	3.15
18	can organise time	0.47	0.52	0.37	0.61	5.40	6.42
19	feel in control [R]	0.48	0.46	0.54	1.24	0.18	2.17
20	feel able do most things	0.57	0.63	-0.97	-1.11	11.00	6.34
21	mentally alert	0.60	0.64	-1.38	-1.17	4.64	2.48
22	joy and elation	0.65	0.71	-2.02	-2.13	20.04	16.46
23	**make decisions easily [R]**	**0.23**	**0.21**	**3.15**	**-**	**27.01**	**-**
24	life has meaning and purpose [R]	0.48	0.46	0.82	1.42	3.65	1.29
25	feel energetic	0.59	0.67	-1.46	-1.36	8.42	5.12
26	good influence	0.58	0.65	-0.99	-1.12	7.80	7.31
27	have fun with others [R]	0.50	0.48	-1.33	-1.12	11.80	8.10
28	feel healthy [R]	0.52	0.51	-0.62	-0.21	1.93	1.73
29	happy memories [R]	0.42	0.41	2.13	2.89	8.78	15.85

Note: [R] Negatively worded items. Removed misfitting items 2, 5, 14 and 23 are presented in bold.

Further examination of individual items fit residuals indicated that items 2 ‘I am intensely interested in other people’ and 14 ‘There is a gap between what I would like to do and what I have done’ had fit residuals above the 2.50 cutoff point and the highest chi-square values (above 17) for item-trait interaction, which indicates a poor fit to the Rasch model. Both items also showed low item-to-total correlations and low loadings (< 0.30) on the first principal component (Table 24). Removing items 2 and 14 improved the overall model fit, but chi square was still at a significant level (χ^2(75)= 152.12, *p* < .001, PSI=.89, Table 23, Analysis 3).

Local Dependency

At this stage the residual correlation matrix was examined, which revealed three groups of locally dependent items displaying residual correlations above the acceptable level of .20 above the mean of all residual correlations (Marais & Andrich, 2008; Christensen et al., 2016). This observation was confirmed by a principal component factor analysis with principal axis factoring using Varimax rotation. Fixing the number of factors to three clearly highlighted the locally dependent items as three distinct groups (Table 25).

Table 25 *Principal component factor analysis with principal axis factoring using Varimax rotation and number of factors fixed to three after misfitting items 2, 5, 14 and 23 were removed (n=281).*

Factor **Items**	**Negative worded**	**Cognitive**	**Affective**
27 have fun with others R	0.75		
24 life has meaning and purpose R	0.72		
29 happy memories R	0.70		
1 pleased with self R	0.65		
28 feel healthy R	0.65		
6 optimistic R	0.63		
19 feel in control R	0.59		
10 world is good R	0.51		
13 look attractive R	0.46		
8 committed and involved		0.67	
21 mentally alert		0.66	
9 life is good		0.64	
16 find beauty in things		0.62	
3 life is rewarding		0.60	
12 satisfied with life		0.58	
18 can organise time		0.51	
4 warmth for others		0.45	
11 laugh a lot			0.79
17 cheerful effect on others			0.78
7 find things amusing			0.56
25 feel energetic		0.49	0.54
15 very happy		0.48	0.53
20 feel able to do most things		0.44	0.48
22 joy and elation		0.41	0.48
26 good influence		0.44	0.46

Note: R Negatively worded items. Coefficients below 0.40 are suppressed for clarity

One larger group included all negatively worded items (1, 6, 10, 13, 19, 24, 27, 28 and 29); the second group items were more cognitive and thus focusing on attitude (3, 4, 8, 9, 12, 16, 18, 21 and 3); and the third group included mainly items focused on affect (7, 11, 15, 17, 20, 22, 25 and 26). However, items 15, 20, 22, 25 and 26 cross-loaded on both attitude and affect, with consistently higher coefficients representing the affective factor.

These three factors correlate at the level of .50 and above, providing evidence of an overarching single factor but also confirm local dependency between these three item groups. Therefore, locally dependent items were combined into three subtests providing a desirable alternative to improve fit to the Rasch model without excluding further OHQ items (χ^2(9)= 18.00, *p* =.35).

Differential Item Functioning (DIF)

Testing for DIF examined the influence of personal factors such as age, gender, ethnicity, and sample (i.e., New Zealand *versus* Malaysian students) on functioning of individual items. DIF analysis involves comparing distributions of individual scores aggregated by class intervals (CI) mean scores between groups of each personal factor and per each individual item using ANOVA. If the effect of a personal factor (Bonferroni adjusted) is significant for an item(s), it is followed by visual examination of the item characteristic curve (ICC) with CI means for all groups plotted on the ICC. A significant DIF effect by ethnic group was found for subtests 1 (*F*(3,280) = 5.93, *p* <.001) and 2 (*F*(3,280) = 5.77, *p*<.001), Bonferroni adjusted. However, a visual examination of plots revealed that these differences were not consistent across observed CIs with at least one shared CI point at each level , which is considered as non-uniform DIF (Andrich & Hagquist, 2013). Also, there was a significant DIF effect by sample for subtest 1 (*F*(1, 280) = 59.23, *p*<.001), subtest 2 (*F*(1,280) = 59.68, *p*<.001) and subtest 3 (*F*(1,280) = 11.73, *p*<.001), Bonferroni adjusted, but the means were only systematically different for subtests 1 and 2. Therefore, subtests 1 and 2 were split for DIF by sample resulting in the same subtests measuring the New Zealand and the Malaysian sample groups independently. After the subtests were split, DIF analysis were repeated for all personal factors and no significant DIFs were evident. This solution permitted transformation from ordinal to interval level data for each sample without any further modifications and resulted in the best overall model fit (χ^2(15)=13.03, *p* = .60, PSI=.82, Table 23, Analysis 4).

Unidimensionality Test

Unidimensionality of the final model solution was tested by comparing the person estimates from the subset with the highest negative loadings on the first principal component with the estimates from the subset with the highest positive loadings. Strict unidimensionality was evident with the percentage of significant *t*-tests below 5% (Table 23).

Item-person Thresholds Distribution

Figure 17 illustrates the person-item threshold distribution of the original OHQ after uniform rescoring (Table 23, Analysis 2) and Figure 18after removing misfitting items 2, 5, 14 and 23, creating three subtests and splitting subtests by sample DIF (Table 23, Analysis 4). Person ability on the latent factor (happiness) and item difficulty are plotted here using the same metric in logit units. Figure 17 shows extreme misfitting items thresholds on the left-hand side that are outside of the sample abilities and, therefore, do not discriminate between individual happiness levels.

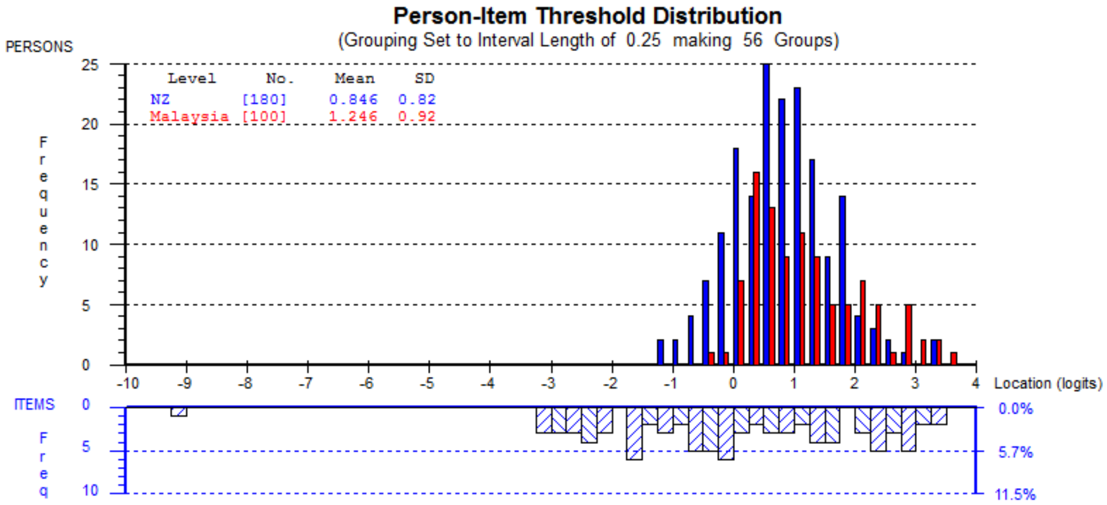


Figure 17. Item-person threshold distribution after uniform rescoring of the OHQ items (Table 23, Analysis 2)

However, after the Rasch modifications that involved removing misfitting items, creating subtests and DIF splitting, the person-threshold distribution closely resembles a normal distribution. The modified OHQ item thresholds perfectly cover the abilities of the sample on the latent factor (happiness), with the Malaysian sample showing higher 'abilities' on the latent factor compared to the New Zealand sample (Figure 18).

Equating t-test

The main purpose of the equating *t*-test is to verify that Rasch modifications produced a significant difference in measuring individual happiness levels by comparing person estimates of the original OHQ and the Rasch modified version using a paired-samples *t*-test. A significant difference was found between the person estimates of the two versions (*t*(279)=21.36, *p*<.01), which confirmed that the implemented Rasch modifications were successful and resulted in an improved solution for the 25-item OHQ version.

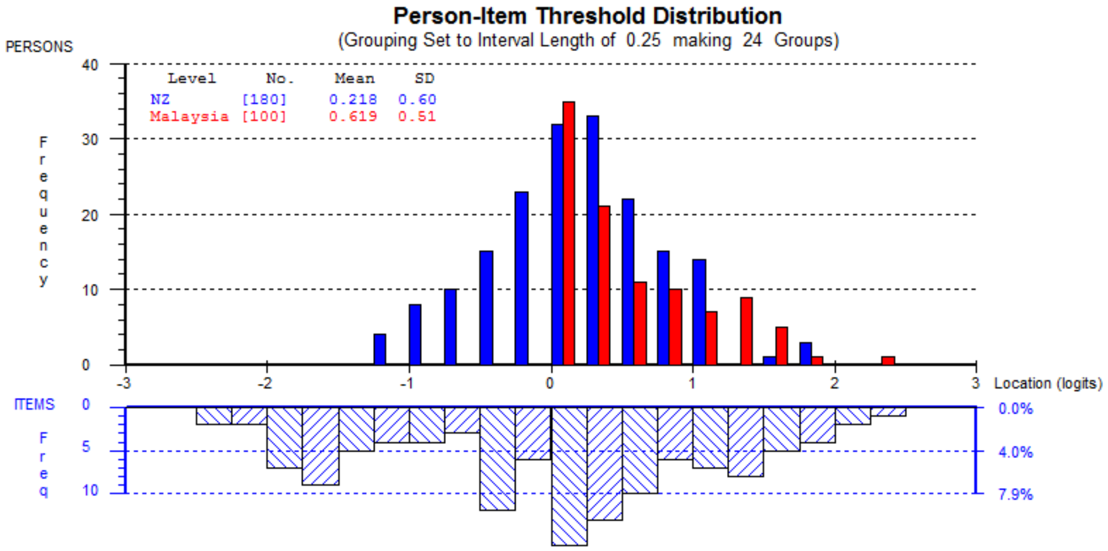


Figure 18. Item-person threshold distribution for the final solution of the OHQ after removing misfitting items 2, 5, 14 and 23, creating subtests and splitting DIF items per sample (Table 23, Analysis 4).

Ordinal-to-interval conversion tables

Table 26 provides a simple algorithm to convert ordinal OHQ scores to interval-level data. The raw scores presented here should be calculated after items 2, 5, 14 and 23 have been excluded and after response options *strongly disagree* have been recoded as 0, *moderately disagree* and *slightly disagree* as 1, *slightly agree* and *moderately agree* as 2, and *strongly agree* as 3. Then, the re-coded 25-item responses (score range 0 to 75) should be added for each person. Corresponding interval-level scores are presented on the right-hand side in both logit units and the raw score scale metric for convenience. Researchers who are currently using the OHQ or have already collected their data can use this algorithm to increase the precision of the OHQ. Considering the DIF by sample reported above, two different conversion tables are presented for use with samples comparable to the New Zealand and to the Malaysian student populations. It should be noted that the proposed ordinal-to-interval conversion algorithm can be used without the need to modify the original OHQ response format. However, these conversion tables can only be used for the respondents without missing data. The author may be contacted to assist with score conversion (Medvedev et al., 2016c).

Table 26. *Converting from a uniformly rescored 25-item OHQ[a] raw score (0 to 75) to an interval scale in logit units and in the same scale metric (0-75) for the New Zealand and Malaysian student populations.*

Raw score[b]	**Interval measure**				**Raw score[b]**	**Interval measure**			
	New Zealand		**Malaysian**			**New Zealand**		**Malaysian**	
	Logit	**Scale**	**Logit**	**Scale**		**Logit**	**Scale**	**Logit**	**Scale**
0	-3.37	0.00	-2.27	0.00	38	-0.42	33.99	0.16	31.55
1	-2.90	5.44	-2.03	3.19	39	-0.35	34.76	0.17	31.76
2	-2.62	8.72	-1.84	5.54	40	-0.28	35.54	0.19	31.94
3	-2.44	10.70	-1.71	7.29	41	-0.21	36.32	0.20	32.14
4	-2.32	12.12	-1.60	8.74	42	-0.15	37.11	0.22	32.37
5	-2.22	13.23	-1.49	10.11	43	-0.08	37.90	0.23	32.54
6	-2.14	14.14	-1.39	11.45	44	-0.01	38.68	0.25	32.76
7	-2.07	14.94	-1.28	12.92	45	0.06	39.46	0.27	33.00
8	-2.01	15.64	-1.15	14.56	46	0.13	40.24	0.29	33.26
9	-1.96	16.26	-0.93	17.40	47	0.20	41.02	0.31	33.54
10	-1.91	16.84	-0.41	24.15	48	0.26	41.79	0.33	33.88
11	-1.86	17.40	-0.44	23.86	49	0.33	42.56	0.37	34.29
12	-1.82	17.91	-0.23	26.58	50	0.40	43.32	0.41	34.83
13	-1.77	18.42	-0.19	27.09	51	0.46	44.07	0.46	35.50
14	-1.73	18.91	-0.17	27.36	52	0.53	44.82	0.52	36.24
15	-1.69	19.41	-0.15	27.59	53	0.59	45.55	0.58	37.02
16	-1.64	19.90	-0.14	27.78	54	0.65	46.28	0.64	37.83
17	-1.60	20.40	-0.12	28.00	55	0.71	46.99	0.70	38.63
18	-1.56	20.89	-0.11	28.16	56	0.78	47.70	0.76	39.45
19	-1.51	21.41	-0.10	28.24	57	0.84	48.42	0.83	40.27
20	-1.47	21.93	-0.09	28.39	58	0.90	49.12	0.89	41.12
21	-1.42	22.46	-0.07	28.59	59	0.96	49.82	0.96	41.98
22	-1.37	23.00	-0.06	28.72	60	1.02	50.54	1.02	42.85
23	-1.32	23.56	-0.06	28.80	61	1.09	51.26	1.09	43.75
24	-1.27	24.15	-0.04	28.98	62	1.15	52.00	1.16	44.66
25	-1.22	24.74	-0.03	29.11	63	1.22	52.76	1.24	45.61
26	-1.17	25.36	-0.03	29.20	64	1.29	53.56	1.31	46.58
27	-1.11	26.00	-0.01	29.39	65	1.36	54.38	1.39	47.62
28	-1.05	26.66	0.00	29.50	66	1.43	55.26	1.47	48.72
29	-1.00	27.33	0.01	29.68	67	1.52	56.20	1.56	49.87
30	-0.94	28.03	0.02	29.85	68	1.60	57.21	1.66	51.15
31	-0.87	28.73	0.04	30.03	69	1.70	58.33	1.77	52.57
32	-0.81	29.46	0.05	30.23	70	1.81	59.58	1.90	54.21
33	-0.75	30.19	0.07	30.43	71	1.94	61.04	2.05	56.15
34	-0.68	30.93	0.09	30.67	72	2.09	62.85	2.23	58.57
35	-0.62	31.69	0.10	30.89	73	2.30	65.26	2.48	61.83
36	-0.55	32.45	0.12	31.12	74	2.63	69.04	2.88	66.96
37	-0.48	33.22	0.14	31.35	75	3.15	75.00	3.49	75.00

Note: [a] OHQ=Oxford Happiness Questionnaire. [b] To calculate the 25-item OHQ raw score exclude items 2, 5, 14 and 23; uniformly rescore other 25 OHQ items as following: 1 to 0; 2 to 1; 3 to 1; 4 to 2; 5 to 2; 6 to 3; and add rescored values together for each person. This table cannot be used for respondents with missing data.

Discussion

The OHQ (Hills & Argyle, 2002) is a widely-used ordinal measure of subjective happiness, but so far the precision of the instrument has not been investigated in sufficient detail. The current study used modern strategies of Rasch analysis to advance the psychometric properties and precision of the OHQ up to an interval-level measure. The results show that good fit to the Rasch model was achieved by minor modifications including uniform item rescoring, removing four misfitting items and combining locally dependent items into subtests. The psychometric properties of the modified 25-item OHQ are therefore robust, and ordinal responses to the original OHQ items can be transformed into an interval-level scale using the conversion algorithms provided in Table 4. This transformation can be conducted without the need to modify the original 29-item OHQ response format meaning that existing datasets can easily be re-analyzed to provide interval-level measurement.

Similar to Medvedev et al. (2016a), uniform rescoring has proved to be the best strategy to improve disordered thresholds of all OHQ items, as illustrated in Figure 16. These findings provide support for the four options response format used in the OHI (Argyle et al., 1989) as the most appropriate to use with the OHQ items. However, the OHI uses response options in a form of statements, which increases both questionnaire length and completion time. Alternatively, four Likert-scale response options defined as *strongly disagree*=1, *slightly disagree*=2, *slightly agree*=3, *strongly agree*=4, can be used because distinctions between *agree* and *slightly agree*, and between *slightly disagree* and *disagree* were shown to be unreliable (Figure 16).

Items 2, 5, 14 and 23 critically affected individual estimates and appeared to be poorly related to the latent trait of happiness based on both Rasch and conventional psychometric analysis results. Item 2 ‘I am intensely interested in other people’ may have a face validity issue because it implies that happy people should be extremely preoccupied with other people, which is not supported by our results. It seems more likely that happy people naturally attract others and become an object of interest. Item 5 ‘I rarely wake up feeling rested’ seems to focus on perceived level of physical energy at the time point of waking up and does not appear to be a reliable estimate of the latent trait according to our data. Perceived energy levels may vary substantially during the day, and assessing levels after waking up only is unlikely to reflect overall happiness. The results show that Item 14 ‘There is a gap between what I would like to do and what I have done’ is also not a reliable estimate of individual happiness levels. This item seems not to be fully consistent with the theory emphasizing a cognitive component of happiness (Diener, 1984; Diener, Lucas, & Oishi, 2005) because an individual might be totally happy with the state of affairs regardless of a gap. This item might need to be reworded and ask about satisfaction with actual achievements rather than about the difference between one's plans and facts. Finally, negatively worded Item 23 'I don't find it easy to make decisions' implicitly suggests that lack of difficulty when making decisions contributes to greater happiness, which is not supported by our results and available theories on well-being and happiness. Modern life is challenging, and making a decision that ensures future well-being may involve thorough investigation of a subject matter, which is not necessarily an easy task. These four misfitting items were removed, which improved precision of the proposed 25-item OHQ version.

The disparity between findings from the Mokken analysis (Steward et al., 2010) and the current Rasch analysis are due to differences between these two methods. For instance, if an item doesn't meet expectations of the Mokken analysis (e.g. item is very similar to another item), the only available option for researcher is to exclude that item. However, if an item doesn't fit expectations of a Rasch model for the same reason, which refers to local dependency in Rasch terminology, locally dependent items can be combined into subtests without a need to remove any of them. The subtests approach based on Lundgren-Nilsson et al. (2013) was also effective in solving local dependency issues found between items of three distinct groups. This is another distinct advantage of Rasch modelling that permits fine tuning of a scale without a need to remove a large amount items. Interestingly, one group of locally dependent items included all negatively worded items of the OHQ loaded together on one factor, and other items were split into more cognitive and more affective clusters (Table 21) confirming theoretical expectations (Diener, 1984; Diener et al., 2005). However, after locally dependent items were combined into three subtests, unidimensionality of the 25-item OHQ was clearly evident, confirming structural validity of the modified scale. Our results are consistent with earlier research (Argyle, 2001; Hills & Argyle, 1998, 2002; Joseph & Lewis, 1998) suggesting a single dimension of happiness, which contains positive and negative affective components together with cognitive facets such as life satisfaction and happy traits.

One group of items works differently for Malaysian students compared to New Zealand as evidenced by DIF found between the New Zealand and Malaysian samples, and as a result separate ordinal-to-interval conversion tables were generated for each sample

(Table 26). Figure 17 shows the item-person threshold distribution after uniform rescoring but before other Rasch modifications such as removing misfitting items, creating subtests and DIF split. There are item thresholds located outside the sample range of happiness trait and signs of a ceiling effect, which increase measurement error. Figure 18 illustrates the item-person threshold distribution of the 25-item OHQ after Rasch modifications including removal of misfitting items 2, 5, 14 and 23, creating subtests and splitting by sample DIF. It can be seen that the range of happiness levels possessed by the sample are perfectly covered by the item thresholds of the modified OHQ version, which provides clear evidence for successful completion of this Rasch analysis.

The limitations of this study have to be acknowledged. The student samples may not reflect the full diversity of New Zealand's or the Malaysian's ethnicities, and no attempts were made to sample under-represented groups. Also, these findings were not replicated with a sample derived from the general population, which could increase generalizability.

The important contribution of this study is to allow researchers and clinicians to convert the OHQ raw score from an ordinal- to an interval-level scale and to use transformed data with parametric statistics without violating their fundamental assumptions. Also, the interval-level scores of the latent happiness factor provide researchers with the opportunity to study the effects of moderators and mediators of happiness in various contexts (e.g. mindfulness). OHQ scores that are transformed to an interval scale can be analysed reliably in such models because local dependency issues and potential item biases (DIF) are ultimately addressed by Rasch analysis if the data fits the model. The improved precision of the instrument may also be useful to evaluate effects of mindfulness-based treatment or in clinical assessment, where the subjective wellbeing score might reflect recovery from psychological conditions such as anxiety, stress and depression.

Conclusion

Striving to increase subjective wellbeing or happiness in greater society with modern psychological approaches requires accurate assessment of the construct to determine its reliable predictors and obscuring factors. The current study conducted Rasch analysis of the OHQ and has demonstrated that after minor modifications the 25-item OHQ version satisfies the Rasch model expectations. Therefore, the OHQ's precision can be increased up to an interval-level scale by using the ordinal-to-interval conversion tables presented here. These findings permit researchers to use the OHQ data with parametric statistics and to make valid comparisons with interval- and ratio-level data such as EEG, heart rate or blood sugar levels.

Chapter Nine. Rasch analysis of the UK Functional Assessment Measure

Introduction

Systematic reviews of the studies applying mindfulness-based interventions in rehabilitation provide increasing evidence for their multiple benefits but fail to adequately address measurement of both functional independence and mindfulness levels (Siegert et al., 2015). A recent systematic review focused on applying MBIs to the stroke population and included promising results based ordinal scales data (Lawrence, Booth, Mercer & Crawford, 2013). Given limited precision of ordinal measures it is necessary to enhance psychometric properties of both functional independence and mindfulness measures by applying modern psychometric methods such as Rasch analysis. This main focus of this study is on reliable measurement of functional independence in stroke populations.

The Functional Independence Measure is a global measure of disability, extensively used by rehabilitation clinicians to assess in a reliable and valid manner change in severity of disability. It is an 18-item scale comprising 13 'motor' and 5 'cognitive' items (Keith et al.,1987; Hamilton et al.,1987). The Functional Assessment Measure was originally developed in the United States as an extension of the FIM in the mid-1990s (Hall et al.,1993), adding a further 12 items to extend its coverage of cognitive and psychosocial function, for use in patients with more complex disabilities following acquired brain injury. Adapted for use in the UK, the UKFIM+FAM was published in 1999 (Turner-Stokes et al., 1999). It consists of a 30-item scale encompassing physical, cognitive, communicative and psychosocial function. An optional add-on module addresses extended activities of daily living (Law, Fielding, Jackson, & Turner-Stokes, 2009), designed primarily for use in the community.

In the UK, following a stroke, the majority of patients will progress down the pathway to recovery with the help of their local non-specialist rehabilitation services. However, a smaller group of adults, principally of working age, have more complex disabilities (i.e. physical, cognitive, communicative, emotional, and/or behavioural problems) requiring treatment in either tertiary specialised (Level 1) or local specialist (Level 2) in-patient rehabilitation services (Specialised Neurorehabilitation Service Standards, 2015). The UK Rehabilitation Outcomes Collaborative (UKROC) provides the national clinical database collating outcomes for these Level 1 and 2 services and the UKFIM+FAM is now the principal outcome measure within the dataset (Turner-Stokes, Williams, Bill, Bassett, & Sephton, 2016).

It is pertinent to understand the psychometric properties of outcome measures for the population in which they are being used. In a previous paper we examined the psychometric properties of the 30-item UKFIM+FAM in a general neuro-rehabilitation cohort using both Classical Test Theory (CTT) and non-parametric Item Response Theory (IRT) methods (Turner-Stokes & Siegert, 2013). This analysis demonstrated two distinct domains - motor (16 items) and cognitive (14 items) - the latter dividing into a 5-item communicative and 9-item psychosocial component. This yielded an overall factor structure of three subscales (physical, communication and psychosocial), each with a Cronbach’s alpha >0.90 and Cohen’s d effect sizes ranging from 0.86-1.29 between admission and discharge.

A subsequent analysis in stroke patients (Nayar, Vanderstay, Siegert, & Turner-Stokes, 2016) demonstrated the same 3-factor structure (which accounted for 69% of the total variance) and also the anticipated score differences related to hemispheric location of the stroke. Left hemispheric stroke patients had lower scores in the communication subscale (in keeping with the predominance of aphasia in this group), whilst right hemispheric strokes had lower scores for physical function (most probably reflecting the presence of dyspraxia and motor planning difficulties). The scale was considered to be valid, reliable and responsive to changes occurring in this study population, as well as sensitive to differences that resonate with clinical experience. In this paper, we use parametric Item Response Theory (specifically the Rasch model), to evaluate further the psychometric properties of the UKFIM+FAM within a cohort of patients with complex disability admitted for specialist rehabilitation following stroke (Rasch, 1960).

A recent review of the literature found more than 50 published studies that explore how well FIM data conform to the Rasch model. The authors highlight the development and refinement of the model over time and the variety of solutions obtained for the FIM scale, which were tested with and without re-ordering of disordered response categories (Lundgren Nilsson & Tennant, 2011). By contrast, the FIM+FAM has received little exposure to Rasch modelling. Two previous studies have explored the benefits of Rasch transformation of the original US version in patients following stroke (Linn et al., 1999) and traumatic brain injury (Hawley, Taylor, Hellawell, & Pentland, 1999) - but as yet there have been no published Rasch analyses of the UKFIM+FAM in any population.

The aim of this study was to determine whether data from the UKFIM+FAM satisfies the Rasch model expectations in a population of patients with complex disability after stroke.

We also wished to explore whether Rasch-transformed scores were sensitive to the expected differences between left and right hemispheric strokes, and to draw up a Rasch transformation table for converting ordinal scores into interval level data using the simplest possible scale structure.

Method

An extensive literature provides guidance on methodology for Rasch analysis (11,12, 19-25). A recent article by Lundgren-Nilsson and Tennant 2011 (16) examined specifically the literature applying the Rasch model to the FIM™. They described the methodological evolution in approach that has occurred over the 21 years or so since its first application in this context and made the following recommendations to improve the rigor of future analyses:

1. Sample size should be a minimum 20 cases per item in the largest subscale or 243 participants, whichever value is larger (Lundgren Nilsson & Tennant, 2011; Linacre, 1994).
2. Use of the Rating Scale versus the Partial Credit Model chosen according to the Likelihood-Ratio test.
3. Use analytical pathways with and without re-ordering disordered thresholds.
4. Creation of 'testlets' (a combination of two or more items) to deal with local dependency.
5. Unidimensionality tested using Rasch principal components analysis of the residuals and the equating test with paired *t*-tests across all participants.
6. Where present Differential Item Function might require splitting the sample according to the relevant person factor (e.g. age, sex, localisation of injury, etc.).
7. Item removal only as a last resort (in order to maintain the clinical integrity of the instrument).
8. Where possible, production of a transformation table to convert raw sores to Rasch-transformed scores, thus encouraging clinicians to use interval scores.

The present analysis followed all the above steps to deal with each of these issues, when they arose.

Data source, sampling and measure

The data source was the UKROC database, which was set up in September 2008 at Northwick Park Hospital funded by a National Institute for Health Research Programme Grant (Specialised Neurorehabilitation Service Standards, 2015; Turner-Stokes et al., 2016). The dataset comprises socio-demographic and clinical data as well as information on rehabilitation needs, inputs and outcomes on admission and discharge from in-patient rehabilitation. Since April 2013, reporting of the full UKROC dataset is a mandated requirement for commissioning of all Level 1 and 2 specialist rehabilitation services. However, reporting was voluntary until that date, so not all services routinely reported UK FIM+FAM data. Within these Level 1/2 services, which have a mean length of stay of approximately 80 days (sd = 60), the UK FIM+FAM is usually completed for each patient within 10 days of admission and during the last week before discharge to evaluate the functional gains made during the episode of care.

The sample was extracted from the cohort of all 1318 stroke patients consecutively admitted to the 58 Level 1/2 specialist rehabilitation centres in England that submitted data to the UKROC database between January 1, 2010 and May 30, 2013, for whom a complete UKFIM+FAM score was available at both admission and at discharge from the unit. FIM+FAM scores are expected to be lower on admission and higher at discharge from rehabilitation. Mallinson (2011) approach was used to investigate item parameter drift over time, where only one set of responses is included for each patient randomly selected from either admission or discharge data. To ensure that the data represented the full range of the scale, admission and discharge scores were pooled from the complete sample of N=1318, into one dataset. In order not to violate the Rasch assumption of local independence between observations (i.e. to prevent the same patient contributing two entries in the data) only one time point was included, i.e. admission or discharge, for each patient. Taking into account the largest sub-division of the UKFIM+FAM identified from previous factor analyses (i.e. the 16 motor items) we used a randomly selected sample of 320 cases (representing 20 cases per item for this domain) to fulfil the sampling criteria (Turner-Stokes & Siegert, 2013). Figure 19 summarises the process of sample extraction and analysis.

Measure

UK FIMFAM: UK Functional Independence Measure and Functional Assessment Measure. Within the UK FIM+FAM, each of the 30 items is scored on the same seven-point ordinal scale as follows: 1 (Total assistance); 2 (Maximal assistance); 3 (Moderate assistance); 4 (Minimal assistance); 5 (Supervision/set-up); 6 (Independent with device) and 7 (Fully independent). A category of 6 or 7 implies no help from another person while for categories 1 to 4 the assessment is based on the amount of help required, e.g. the percentage of task performed by patient. The UKROC software automatically produces a 'FAM-Splat' or radar chart, presenting a visual impression of change at item level. This may be used to describe change in individual scores or median scores for a population in a format that is clinically interpretable by rehabilitation professionals. By way of example, Figure 20 shows a composite FAM-Splat for median admission and discharge scores within this dataset.

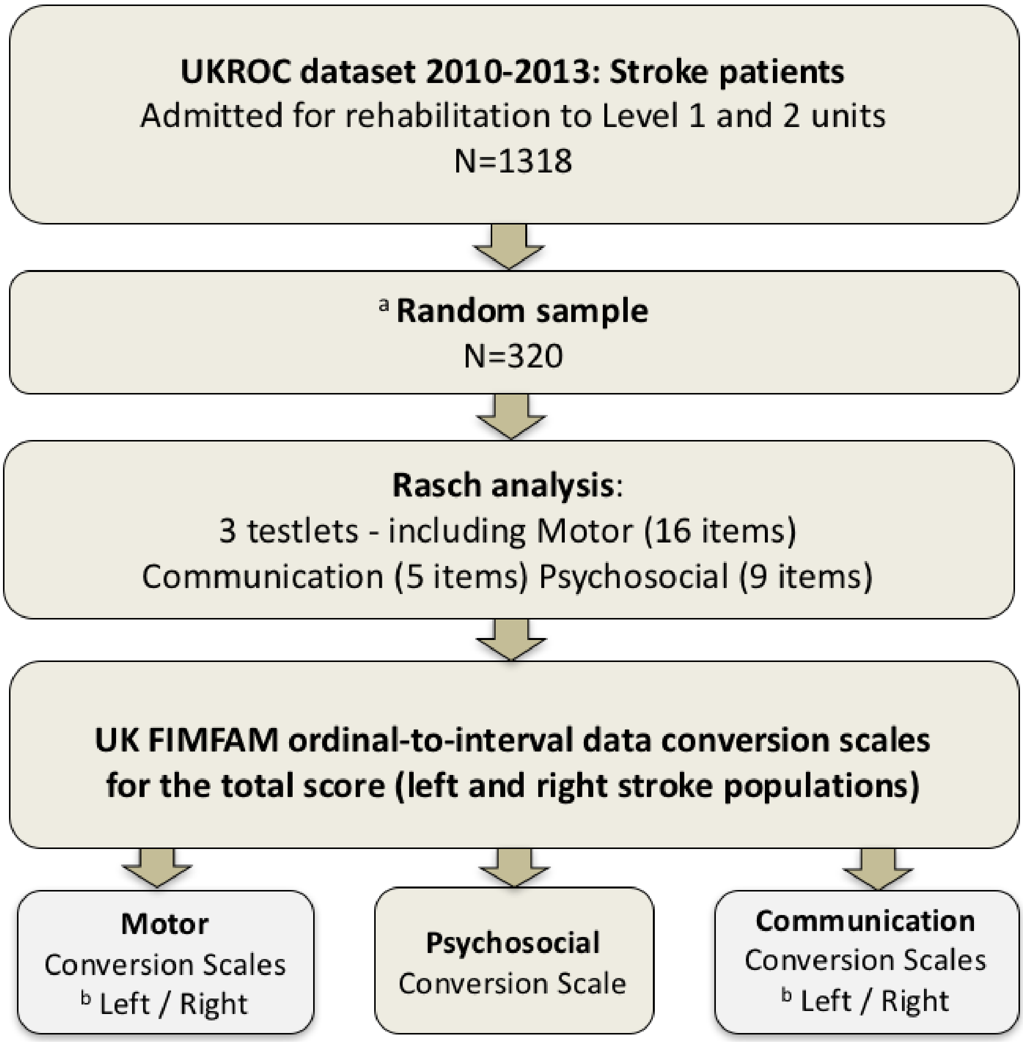

UKROC dataset 2010-2013: Stroke patients
Admitted for rehabilitation to Level 1 and 2 units
N=1318
a Random sample
N=320
Rasch analysis:
3 testlets - including Motor (16 items)
Communication (5 items) Psychosocial (9 items)
UK FIMFAM ordinal-to-interval data conversion scales for the total score (left and right stroke populations)
Motor
Conversion Scales
b Left / Right
Psychosocial
Conversion Scale
Communication
Conversion Scales
b Left / Right

Figure 19. Flow chart of the study sample extraction and analysis.

Legend: [a] Random sample extracted from the dataset derived across admission and discharge values so that each patient is only in the dataset once but both time points are equally represented;

[b] Left/Right stroke (Differential Item Functioning by stroke location led to different conversion scales for left and right stroke)

UKROC: UK Rehabilitation Outcome Collaborative database; UK FIMFAM: UK Functional Independence Measure and Functional Assessment Measure.

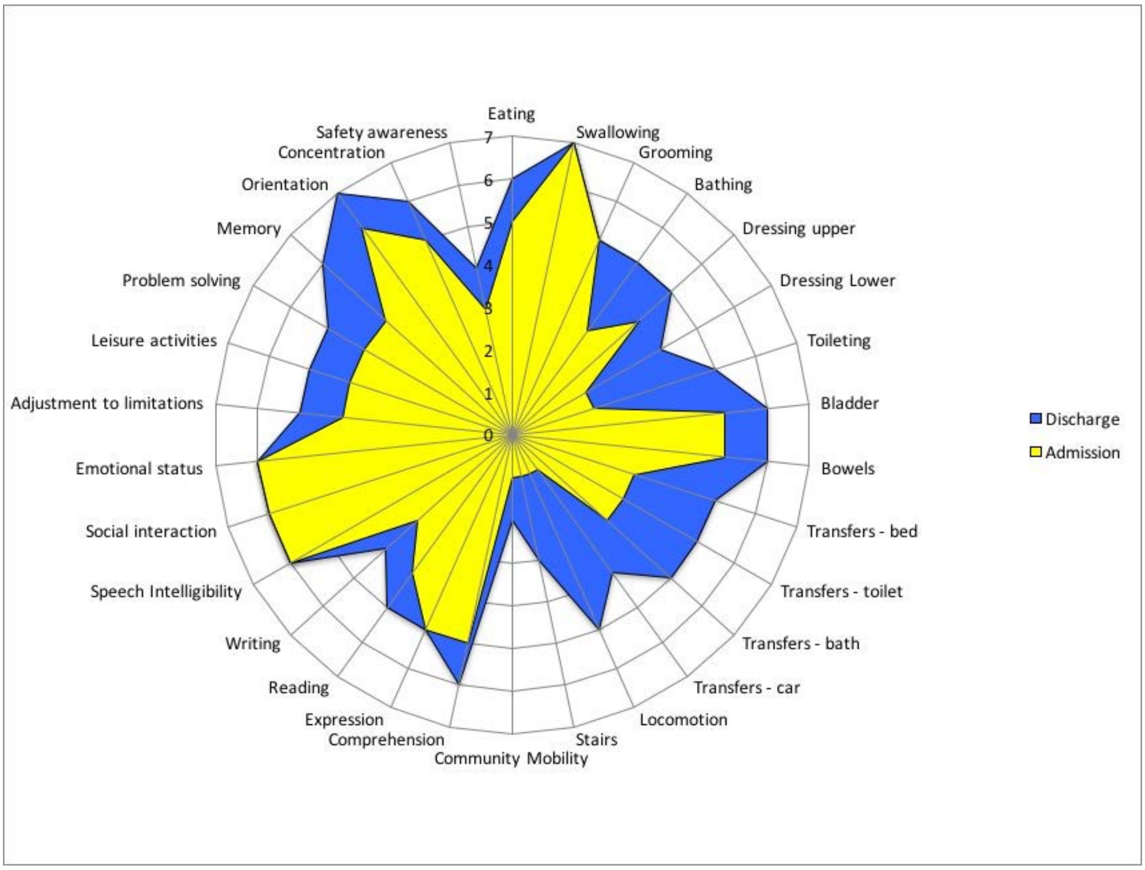


Figure 20. Composite FIM+FAM-splats of the median admission and discharge scores for each item within this dataset (n=320).

Legend: The radar chart (or 'FAM splat') provides a graphic representation of the disability profile from the FIM+FAM data. The 30-scale items are arranged as spokes of a wheel. Scoring levels from 1 (total dependence) to 7 (total independence) run from the centre outwards. Thus, a perfect score would be demonstrated as a large circle. These composite radar charts illustrate the median admission and discharge scores within this dataset. The yellow-shaded portion represents the median admission scores and the blue-shaded area represents the difference between median scores on admission and discharge.

Summing the item scores gives a total range from 30 to 210 where a maximum score of 210 indicates total independence. The seven-category structure implies, in Rasch terms, that each item has 6 possible thresholds or points between two response categories where either response is equally probable (i.e.1 to 2, 2 to 3, etc).

Psychometric analysis of the UK FIM+FAM

Descriptive analysis was carried out using the IBM SPSS 22 software. Rasch analysis was performed using RUMM2030 software (Andrich et al., 2009). A significance value of .05 was used throughout. The Likelihood-Ratio test, to determine whether the Rating scale or Partial Credit Model for Rasch analysis was most appropriate. The summary statistics of the Rasch model were assessed based on the mean item and person location, individual item fit residual, the overall item-trait interaction chi-squared test/ p value and the Person Separation Index (PSI), interpreted according to fit criteria specified in Chapter Two.

In accordance with the recommendations of Lundgren Nilsson & Tennant (2011), several analytical pathways were explored. In the first, all 30 items were fitted to the Rasch model without adjustment of any kind. The second two pathways used a ‘testlet’ approach based on the method used by Lundgren Nilsson et al., (2013) to solve local dependency issues between items of distinct domains. Locally dependent items were combined to produce a single testlet. The 3-factor structure demonstrated by factor analysis (Turner-Stokes & Siegert, 2013; Nayar et al., 2016) formed the basis for these testlets but, in addition, local dependency between items was examined using a residual inter-item correlation matrix.

The second analytical pathway used testlets without re-scoring and the third involved re-scoring of any significantly disordered thresholds at individual item level prior to further analysis. A disordered threshold occurs when people higher in the ability or construct being measured (in this case *independence*) do not consistently obtain correspondingly higher response options (i.e. 1, 2, 3…7) for an item. In Rasch analysis disordered thresholds are corrected by collapsing adjacent response categories. Items with significantly disordered thresholds were rescored by collapsing adjacent categories in a meaningful way (e.g. ‘total’ and ‘maximal assistance’; ‘supervision/set-up’ and ‘modified independence’, leaving a separate category for ‘complete independence’).

In both the second and third pathways, item bias (DIF) was examined across important person factors such as age group (0-44, 45-54, 55-64, 65-74, 75 plus), gender, ethnicity, type of stroke (haemorrhagic, infarct, sub-arachnoid and other), stroke location (left or right hemisphere) and time point (admission or discharge) (Holland & Wainer, 1993). Items displaying differential item functioning (DIF) were split to allow variation by the corresponding factor.

As it was desirable to keep the original structure of the UKFIM+FAM scale, item removal was considered only as a last resort to improve the fit. The items at risk of deletion were those exhibiting significant misfit, i.e., excessive residual values (> ± 2.5) and a *p* value significant at the 0.05 level, with a Bonferroni adjustment for multiple tests (Bland & Altman, 1995).

Results

Within our random sample of 320 cases, the mean age was 58.70 (SD=15.27) years, range 16 to 89. To confirm that this group was representative of the cohort from which it was drawn, we compared the socio-demographic and clinical characteristics of our Rasch study sample with the full cohort. No significant differences were seen (Table 27). The likelihood-ratio test (*p*<.0001) indicated the suitability of the Partial Credit Model.χ^2 (120) = 540.87, *p*<.001). Table 28 presents the Rasch fit statistics for all three analytical pathways described above.

Table 27. *The UKROC: stroke population and the Rasch random sample characteristics.*

	UKROC Study sample *n*=1318	**Random sample (Rasch analysis)[a] *n*=320**
Age, years	***n* (%)**	***n* (%)**
< 44	220 (16.7)	50 (15.6)
45-54	293 (22.2)	74 (23.1)
55-64	298 (22.6)	66 (20.6)
65-74	250 (19.0)	54 (16.9)
74+	231 (17.5)	68 (21.3)
Unknown	26 (2.0)	8 (2.5)
Male, n (%)	752 (57.1)	189 (59.1)
Ethnicity		
White	951 (72.2)	227 (70.9)
Asian/Asian British	98 (7.4)	21 (6.6)
Black/Black British	110 (8.3)	29 (9.1)
Other	41 (3.1)	10 (3.2)
Unknown	118 (8.9)	33 (10.3)
Length of stay, days,		
Mean (SD)	77.7 (57.3)	78.9 (52.6)
Diagnosis localisation	***n* (%)**	***n* (%)**
Right hemisphere	638 (48.4)	159 (49.7)
Left Hemisphere	680 (51.6)	161 (50.3)
Diagnosis subcategory		
Haemorrhagic	386 (29.3)	93 (29.1)
Infarct	707 (53.6)	174 (54.4)
Sub-Arachnoid	136 (10.3)	32 (10.0)
Other	89 (6.8)	21 (6.6)

[a] Random sample extracted from the dataset (n=1318) derived across admission and discharge values so that each patient is only in the dataset once but both time points are equally represented

Table 28. *The UK FIM FAM: Rasch model summary statistics (overall fit of the scale).*

UK FIM FAM Rasch model	Item Location Mean (SD)	Item Fit residual Mean (SD)	Person Location Mean (SD)	Person Fit residual Mean (SD)	Item –Trait Interaction χ square /DF	*p* value	PSI	Unidimensional *t*-test (%)
<u>Pathway 1</u> (All items)	0.00 (0.42)	0.01 (2.99)	0.16 (0.98)	-0.21 (1.52)	540.87/120	.000	.95	No (41.88)
<u>Pathway 2: Three testlets – no re-scoring</u>								
Analysis 2A	0.00(0.04)	0.22(1.72)	0.05(0.25)	-0.33(0.89)	20.81/12	.053	.82	Yes (2.19)
Analysis 2B with DIF	0.00(0.07)	0.24(1.20)	0.05(0.28)	-0.35(0.94)	14.48(10)	.152	.83	Yes (2.19)
<u>Pathway 3: Three testlets - with re-scoring</u>								
Analysis 3A	0.00 (0.05)	0.26 (1.29)	0.06 (0.27)	-0.34 (0.90)	18.65/12	.097	.81	Yes (2.50)
Analysis 3B with DIF	0.00(0.08)	0.33(0.82)	0.05(0.29)	-0.36(0.95)	19.91(20)	.463	.83	Yes (2.50)

UK FIM FAM = UK Functional Assessment Measure; DF = degrees of freedom; PSI = Person Separation Index.

Analytical pathway 1: Initial analysis of the full 30-item scale

The initial analysis including all 30 items showed good reliability (PSI=.95) but misfit at both individual item and overall level with significant item-trait interaction. Table 29 presents Rasch model fit statistics for each individual item, along with the frequency distribution of responses for each of the 7 scoring categories within the 30 items. There were 5 items, which had a single category endorsed by less than 10 persons, which represents less than 5% of available response options. Also, 13 out of 30 items show significant misfit to the Rasch model. At this stage the residual correlation matrix was examined and it displayed local dependencies between three groups of items that mirrored previously reported results of factor analysis (9, 10) - i.e. Motor (16 items), Communication (5 items) and Psychosocial (9 items) function). For the next stages of the analysis, the 30 items were combined into three testlets representing motor, communication and psychosocial function.

Analytical pathway 2: Testlet analysis without re-scoring

Pathway 2a: Testlet analysis without rescoring produced satisfactory overall model fit with (χ^2 (12) = 20.81, p =.053; PSI=.82) and confirmed unidimensionality with only 2.19% of *t* tests significant (see Table 28). DIF analysis indicated significant uniform DIF for the Motor (*F*(1,319)=39.69, p<0.05) and Communication subtests (*F*(1,319)=97.13, p<.05) by stroke localisation, but no other DIF was identified.

Table 29: *Frequency distribution of responses and Rasch model fit statistics for the UKFIM+FAM items (initial analysis), and domain subtests split by localisation (without rescoring).*

Item	Description	Location	Fit Residual	Chi-Square	Frequency distribution across scoring categories						
					Cat 1	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 5**	**Cat 6**	**Cat 7**
1	[a] Eating	-0.54	-0.13	2.84	21	**7**	10	18	114	43	104
2	[a] Swallowing	-0.96	1.70	7.49	11	**8**	11	10	36	22	219
3	Grooming*	-0.25	**-4.32**	**37.10**	25	31	21	41	84	38	77
4	Bathing*	0.19	**-4.43**	**37.25**	46	39	52	54	49	33	44
5	Dressing - upper*	-0.03	**-4.41**	**33.17**	42	36	40	48	44	36	71
6	Dressing - lower*	0.33	**-5.05**	**29.48**	77	57	33	32	33	27	58
7	Toileting*	0.16	**-4.34**	**22.86**	78	52	28	22	20	36	81
8	[a] Bladder	-0.14	2.37	10.25	62	27	24	20	21	36	127
9	[a] Bowel	-0.15	-2.31	11.58	64	19	23	16	29	37	129
10	[a] Bed transfers*	0.05	**-2.93**	**16.46**	71	24	35	26	42	31	88
11	[a] Toilet transfers*	0.11	**-3.91**	**19.68**	78	21	32	29	37	42	78
12	Bath transfers	0.35	-0.29	2.89	114	17	31	22	36	37	60
13	[a] Car Transfers	0.48	-0.73	1.73	159	**8**	13	25	28	21	63
14	[a] Locomotion	0.40	0.44	6.03	132	17	**8**	10	41	54	55
15	[a] Stairs	0.68	-0.68	1.18	184	2	13	12	28	36	42
16	[a] Community Mobility	1.05	0.88	1.65	183	38	38	**8**	22	11	17
17	Comprehension	-0.37	1.92	8.86	16	27	34	28	65	71	76
18	Expression*	-0.11	**3.87**	**33.13**	38	37	29	19	47	68	79
19	[a] Reading	0.12	1.92	10.95	69	19	22	35	60	47	65
20	[a] Writing*	0.33	**3.16**	**32.50**	98	36	24	24	45	31	59
21	Speech intelligibility*	-0.40	**4.04**	**46.20**	26	18	23	26	36	53	135
22	Social Interaction*	-0.67	**3.75**	**22.20**	13	18	21	16	44	78	127
23	Emotional Status*	-0.32	**6.55**	**76.91**	26	26	21	15	40	84	105
24	Adjustment to limitations	0.04	1.54	12.94	29	47	56	35	52	60	38
25	[a] Use of leisure time	0.27	0.03	7.34	49	46	43	27	36	96	20
26	Problem Solving*	0.31	-1.57	**17.57**	45	55	39	36	70	44	28
27	Memory	-0.13	1.99	9.47	32	35	41	40	44	51	74
28	[a] Orientation	-0.47	1.48	8.69	28	13	27	26	32	37	154
29	Concentration	-0.34	0.21	7.29	20	27	37	35	66	48	84
30	[a] Safety Awareness	0.02	-0.39	5.21	16	96	50	37	26	51	41
	Subtests										
1	Motor Left	0.01	-0.85	0.73							
2	Motor Right	0.07	-1.22	1.73							
3	Communication left	0.05	0.85	6.84							
4	Communication right	-0.10	0.88	3.18							
5	Psychosocial	-0.03	1.54	2.00							

*Significant misfit to the Rasch model (*p*<.05, Bonferoni adjusted);

[a] Denotes items with significantly disordered thresholds ($p < 0.05$)

Pathway 2b: When the motor and communication subtests were split by localisation (left/right) to control for DIF, this produced the best model fit with (χ^2 (10) = 14.48, *p* =.152) and an improved PSI of .83 (Table 28).

Analytical pathway 3: Testlet analysis with re-scoring.

Pathway 3a: Applying the third analytical pathway (with rescoring), significantly disordered thresholds were identified in 15 out of 30 items. Table 29 indicates the items with significantly disordered thresholds. Notably, of the 15 items with disordered thresholds only 3 items (No. 8 (Bed transfers), No. 9 (Toilet transfers), and No. 20 (Writing)) are mis-fitting. All 15 items with disordered thresholds were re-scored before the analysis continued. After rescoring, the items showed similar patterns of local dependency and were combined into motor, communication and psychosocial subtests. The resultant fit indices were comparable to those achieved without rescoring (χ^2 (12) = 18.65, *p* =.097) and a PSI of .81.

Pathway 3b: When the motor and communication subtests were split by localisation (left/right stroke) after rescoring of the 15 items, this produced very good model fit (χ^2 (12) = 19.91, *p*=.463) with a PSI of .83, equal to the result of the second analytical pathway (without rescoring). Table 29 also includes fit statistics for each individual testlet of the final solution, which all have comparable level of difficulty and satisfy Rasch model expectations.

Figure 21 presents the item-person threshold distributions of the best solution with and without re-scoring (bottom and top panel respectively). Both show that abilities of the sample are fairly well targeted by item thresholds with minor signs of a ceiling effect for the right stroke population and floor effect for the left stroke population. However, person distribution without rescoring (top panel) is closer to a normal distribution than the rescored analysis (bottom panel). Figure 22 shows scatter plots for Motor, Communication and Psychosocial domain interval level scores as a function of ordinal raw scores including DIF by localisation. The Communication and Psychosocial scales show a reasonable gradient, but the distribution of the motor scale is notably ‘flat’ in the middle part of the scale, which may potentially affect the sensitivity to change of the interval scores, requiring further evaluation in clinical practice.

Both analytic pathways (without rescoring/with rescoring) resulted in very good fit to the Rasch model, but there is a major advantage in using the simpler conversion algorithm. Therefore, Tables 30A and 30B contain ordinal-to-interval conversion scores estimated from the analysis without rescoring disordered thresholds.

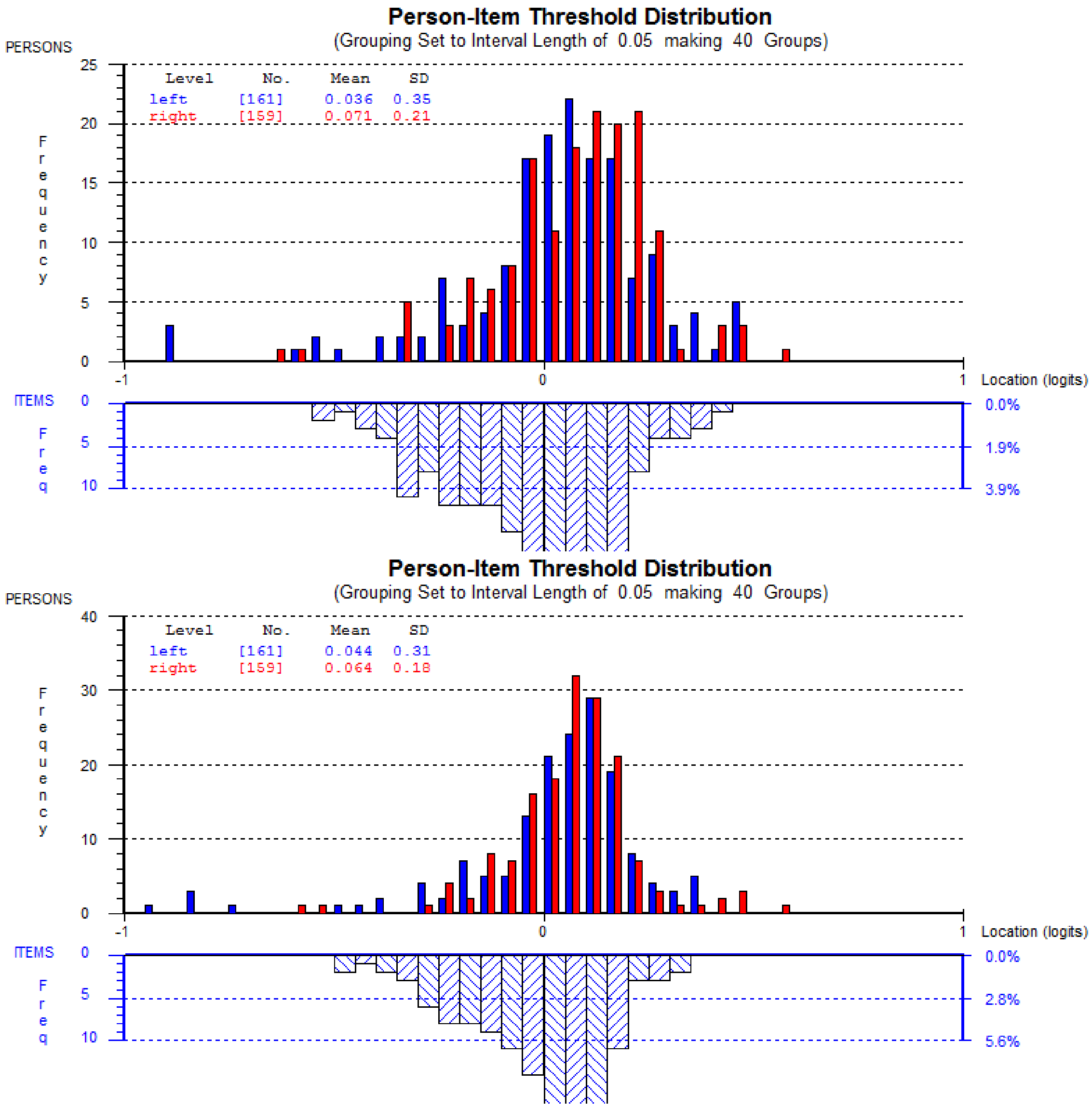


Figure 21. Person-item threshold distributions for the final solution without re-scoring (top panel) and with re-scoring (bottom panel) for the left and right stroke populations.

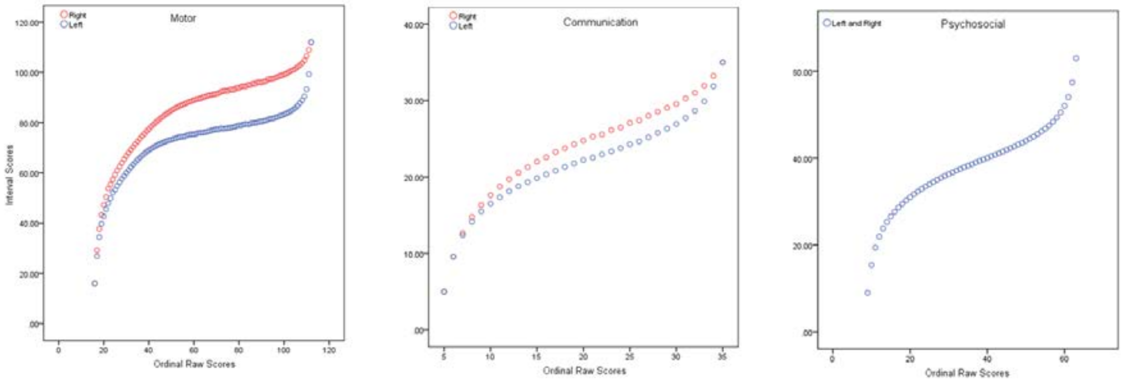


Figure 22. Scatter plots for the UKFIM+FAM Motor, Communication and Psychosocial domains interval level scores as a function of ordinal raw scores including DIF by localisation.

Table 30A. *The UKFIM+FAM total score ordinal-to-interval conversion scale for left and right strokes.*

Raw	Interval		Raw	Interval		Raw	Interval		Raw	Interval		Raw	Interval		Raw	Interval	
Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Left	Right
30	30.00	30.00	61	111.70	115.36	92	120.59	124.64	123	124.94	131.11	154	127.78	135.31	185	133.88	139.24
31	55.96	57.20	62	112.06	115.80	93	120.80	124.86	124	124.99	131.28	155	127.89	135.37	186	134.25	139.35
32	70.18	72.01	63	112.48	116.19	94	121.01	125.09	125	125.04	131.44	156	128.04	135.53	187	134.71	139.63
33	77.99	80.09	64	112.84	116.57	95	121.16	125.31	126	125.20	131.61	157	128.04	135.53	188	135.07	139.85
34	83.62	85.45	65	113.20	116.96	96	121.32	125.53	127	125.35	131.78	158	128.20	135.70	189	135.59	140.12
35	86.88	89.37	66	113.56	117.29	97	121.53	125.75	128	125.30	131.94	159	128.35	135.81	190	136.06	140.40
36	89.78	92.36	67	113.92	117.68	98	121.73	125.97	129	125.51	132.11	160	128.45	135.92	191	136.62	140.68
37	92.15	94.85	68	114.23	118.01	99	121.84	126.19	130	125.51	132.27	161	128.56	136.09	192	137.19	141.06
38	94.12	96.89	69	114.60	118.34	100	122.04	126.41	131	125.61	132.44	162	128.71	136.14	193	137.76	141.39
39	95.77	98.66	70	114.91	118.67	101	122.20	126.63	132	125.66	132.55	163	128.87	136.25	194	138.38	141.78
40	97.22	100.21	71	115.22	119.00	102	122.35	126.86	133	125.82	132.72	164	128.97	136.42	195	139.05	142.28
41	98.51	101.54	72	115.53	119.28	103	122.51	127.08	134	125.92	132.88	165	129.13	136.47	196	139.83	142.78
42	99.70	102.75	73	115.84	119.61	104	122.66	127.30	135	125.92	132.99	166	129.28	136.64	197	140.66	143.55
43	100.79	103.91	74	116.10	119.89	105	122.82	127.52	136	126.13	133.16	167	129.44	136.70	198	141.54	144.21
44	101.72	104.91	75	116.41	120.22	106	122.97	127.74	137	126.08	133.27	168	129.59	136.70	199	142.57	145.04
45	102.60	105.85	76	116.66	120.50	107	123.08	127.96	138	126.23	133.38	169	129.75	136.81	200	143.71	146.04
46	103.43	106.73	77	116.98	120.77	108	123.23	128.18	139	126.39	133.49	170	129.90	136.92	201	144.95	147.25
47	104.20	107.56	78	117.23	121.05	109	123.34	128.35	140	126.44	133.65	171	130.11	137.03	202	146.40	148.58
48	104.93	108.28	79	117.49	121.33	110	123.49	128.57	141	126.44	133.77	172	130.26	137.19	203	148.16	150.29
49	105.60	109.00	80	117.75	121.60	111	123.59	128.79	142	126.59	133.88	173	130.47	137.36	204	150.33	152.34
50	106.22	109.66	81	118.01	121.88	112	123.75	128.96	143	126.64	134.04	174	130.68	137.41	205	152.96	154.83
51	106.84	110.33	82	118.27	122.16	113	123.85	129.18	144	126.85	134.15	175	130.88	137.58	206	156.38	158.26
52	107.41	110.93	83	118.53	122.43	114	123.96	129.40	145	126.90	134.26	176	131.14	137.64	207	161.19	162.79
53	107.98	111.49	84	118.78	122.65	115	124.06	129.62	146	127.06	134.32	177	131.40	137.86	208	168.79	170.20
54	108.49	112.04	85	118.99	122.93	116	124.16	129.79	147	127.16	134.54	178	131.66	137.97	209	183.11	184.02
55	109.01	112.59	86	119.25	123.21	117	124.32	130.01	148	127.11	134.59	179	131.92	138.13	210	210.00	210.00
56	109.48	113.09	87	119.46	123.43	118	124.42	130.17	149	127.32	134.76	180	132.18	138.35			
57	109.94	113.59	88	119.72	123.65	119	124.52	130.39	150	127.42	134.76	181	132.49	138.46			
58	110.41	114.03	89	119.92	123.93	120	124.63	130.56	151	127.42	134.98	182	132.80	138.69			
59	110.87	114.53	90	120.13	124.15	121	124.68	130.72	152	127.52	134.98	183	133.16	138.80			
60	111.29	114.97	91	120.39	124.37	122	124.83	130.89	153	127.68	135.09	184	133.47	138.96			

Table 30B. *The UKFIM+FAM motor, communication and psychosocial domains ordinal-to-interval conversion scale for left and right strokes.*

Raw	Motor Interval		Raw	Motor Interval		Raw	Motor Interval		Raw	Motor Interval		Raw	Comm Interval		Raw	Interval	Raw	Interval
Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Left	Right	Score	Psychs	Score	Psychs
16	16.00	16.00	47	72.35	83.01	78	78.28	93.25	109	90.46	104.73	5	5.00	5.00	9	9.00	40	40.05
17	26.85	29.22	48	72.66	83.62	79	78.67	93.97	110	93.27	106.57	6	9.56	9.58	10	15.39	41	40.46
18	34.42	37.72	49	72.98	84.23	80	78.91	93.76	111	99.28	108.93	7	12.40	12.68	11	19.38	42	40.77
19	39.73	43.25	50	73.21	84.85	81	78.83	94.38	112	112.00	112.00	8	14.16	14.75	12	21.92	43	41.12
20	42.69	47.25	51	73.52	85.36	82	79.30	94.58				9	15.49	16.33	13	23.79	44	41.50
21	45.50	50.42	52	73.76	85.87	83	79.45	94.38				10	16.51	17.63	14	25.29	45	41.92
22	47.84	53.81	53	74.15	86.28	84	79.30	95.09				11	17.37	18.75	15	26.57	46	42.27
23	49.87	55.45	54	74.30	86.80	85	79.53	94.89				12	18.14	19.71	16	27.65	47	42.65
24	52.14	57.39	55	74.38	87.10	86	79.92	95.50				13	18.79	20.57	17	28.62	48	43.14
25	53.31	59.24	56	74.54	87.51	87	79.84	95.40				14	19.33	21.28	18	29.49	49	43.52
26	54.87	60.98	57	75.00	87.92	88	80.00	96.02				15	19.86	22.00	19	30.29	50	43.97
27	56.27	62.51	58	75.24	88.33	89	80.47	96.12				16	20.35	22.59	20	31.02	51	44.46
28	57.60	63.95	59	75.32	88.54	90	80.39	96.02				17	20.81	23.27	21	31.71	52	44.98
29	58.77	65.38	60	75.32	89.05	91	80.70	96.32				18	21.33	23.76	22	32.34	53	45.53
30	60.02	66.72	61	75.55	89.26	92	80.86	96.53				19	21.77	24.29	23	32.93	54	46.16
31	61.11	67.94	62	75.94	89.46	93	81.33	97.35				20	22.21	24.75	24	33.48	55	46.78
32	62.20	69.17	63	75.94	89.66	94	81.56	97.45				21	22.53	25.28	25	34.00	56	47.51
33	63.30	70.30	64	76.02	90.18	95	81.64	97.45				22	22.95	25.56	26	34.52	57	48.38
34	64.23	71.43	65	76.18	90.28	96	81.95	97.81				23	23.35	26.15	27	35.01	58	49.35
35	65.17	72.45	66	76.41	90.79	97	82.19	98.17				24	23.77	26.49	28	35.46	59	50.53
36	66.03	73.58	67	76.57	90.89	98	82.65	98.58				25	24.28	27.07	29	35.91	60	52.03
37	66.89	74.60	68	76.96	91.10	99	82.89	98.78				26	24.65	27.41	30	36.33	61	54.08
38	67.67	75.53	69	77.19	91.30	100	83.28	99.19				27	25.19	28.00	31	36.75	62	57.44
39	68.45	76.55	70	77.35	91.51	101	83.67	99.50				28	25.77	28.53	32	37.16	63	63.00
40	69.07	77.37	71	77.27	91.71	102	84.06	100.12				29	26.33	29.06	33	37.51		
41	69.62	78.39	72	77.74	92.33	103	84.60	100.63				30	26.93	29.55	34	37.93		
42	70.17	79.21	73	77.58	92.53	104	85.15	100.93				31	27.72	30.29	35	38.27		
43	70.71	80.03	74	77.66	92.74	105	85.85	101.45				32	28.65	31.01	36	38.62		
44	71.18	80.85	75	77.81	92.64	106	86.63	102.16				33	29.91	31.93	37	39.00		
45	71.57	81.57	76	77.97	93.25	107	87.80	102.88				34	31.88	33.24	38	39.39		
46	71.96	82.29	77	78.13	93.05	108	88.90	103.70				35	35.00	35.00	39	39.70		

Tables 30A and B provide algorithms to convert ordinal scores into interval scores in the original UK FIM+FAM scale format. Ordinal-to-interval conversion can be conducted by calculating ordinal domain and full scale scores and finding corresponding interval scores on the right hand side without altering the original response format of the UK FIM+FAM. It should be noted that these tables cannot be used for patients with missing data.

Discussion

The study presented here represents the first Rasch analysis of the UK FIM+FAM a measure which is the primary outcome measure within the UKROC national clinical dataset for all specialist rehabilitation services in the UK treating patients with complex disabilities.

The best fit to the Rasch measurement model was achieved when three groups of locally-dependent items were treated as testlets consistent with earlier results of factor analysis (Turner-Stokes & Siegert, 2013; Nayar et al., 2016). This solution was tested by applying different analytical pathways, with and without rescoring items with disordered thresholds, and produced similar fit indices that both satisfy the expectations of the unidimensional Rasch model. Together these findings suggest that the UKFIM+FAM satisfies the unidimensional Rasch model without the need to rescore disordered thresholds in a random sample of stroke patients. These results have practical utility including retaining all seven original response options for all 30 items and allowing for very simple conversion from raw scores to an interval metric.

Two previous studies have explored Rasch analysis of the original US version (USFIM+FAM), using the WINSTEPS software (Linn et al., 1999; Hawley et al., 1999). Linn et al. (1999) also reported a number of misfitting items, but they were principally interested in whether the FAM solved the problem of ceiling effects in the FIM. This has limited relevance to the present study as the UKFIM+FAM has dealt with ceiling effects in a different way - by providing a separate module addressing extended activities of daily living (Law et al., 2009) as well as a related scale of workability (Turner-Stokes et al., 2014).

Hawley et al. 1999 examined the USFIM+FAM in a cohort of 652 patients with traumatic brain injury (TBI). They used a principal component analysis to identify two separate dimensions (Motor-16 items and Cognitive 14 items), which conformed only partially to the Rasch model. As they point out, the imperfect fit is hardly surprising given the heterogeneity of a typical brain injury sample and the diverse nature of the items captured within the FIM+FAM. It does not necessarily indicate that the scale is fundamentally flawed in a clinical sense. The question that arises, however, is what further division of subscales is necessary to improve the fit - and these considerations may apply equally to a complex stroke population.

These early studies reported the goodness of overall and individual item fit to the Rasch model, but typically went little further. They frequently relied on deleting items to attain satisfactory fit and rarely provided a table to permit the conversion of raw scores to interval level scores in routine clinical practice. A major methodological strength of our study is that we were able to draw upon 21 years of experience in Rasch studies on the FIM, following the methodology described by Lundgren Nilsson and Tennant (2011) and Lundgren Nilsson et al., (2013) to explore how well the UKFIM+FAM fits the Rasch model according to more current analytical techniques. A range of steps were used including the formation of testlets to eliminate local dependency among items to achieve reasonably good fit for the three dimensions underpinning the UK FIM+FAM. Importantly, this was achieved without deleting any items and it was also possible to produce a conversion table for left and right hemisphere strokes, to account for differential item functioning between these two groups.

The chief advantage of measures that conform to the Rasch model is that their data can be analysed with parametric statistics rather than relying on non-parametric statistics implying greater statistical power and precision. Whilst the use of interval level scales has some clear advantages for the generation of robust metrics for the purpose of research, further work is necessary to explore the impact and benefits of transformed scores in the clinical setting. We recognise that, despite the many conversion tables that have been produced for FIM in different contexts (Lundgren Nilsson & Tennant, 2011), the uptake of these by clinicians has been limited because the ordinal scores within each item are interpretable at a clinical level and are widely used as an aid to clinical reporting and decision-making. Usefulness of conversion table for clinical practice depends on scale sensitivity in the midrange. In many cases Rasch transformation results in low sensitivity in the midrange and high sensitivity on the upper and lower end. Even though a Rasch transformed interval scale accurately reflect change of person ability on the latent trait it may be problematic to differentiate between patients in the middle range of the scale, which might be necessary due to clinically important distinctions.

The FAM splat is particularly valued by UK clinicians in this context, and for this reason we would not necessarily recommend using transformed scores at individual item level, but they may nevertheless prove valuable when presenting summed items in subscale and total scores, particularly if the transformed data prove to be more sensitive (Hobart et al., 2010). However, this requires further evaluation, especially in view of the relatively flat distribution in the middle part of the motor subscale, as noted above.

A number of methodological limitations to this study is also recognised. All the participants were stroke patients drawn randomly from the larger UKROC dataset, which collates a selected population of patients (mainly of working age) with complex neurological disabilities. These findings cannot necessarily be extrapolated to the more general population of stroke patients, who are mainly older with shorter lengths of stay in rehabilitation. Moreover, the present study focused solely on inpatients and it is possible that ceiling effects might be observed with a community sample post-discharge.

We used a sample of 320 patients selected at random from the 1318 stroke patients recorded on the UKROC database in order to satisfy the practical requirements of the Rasch model. As the model is tested by a series of Chi-squared tests, for both overall model fit and for individual item and person fit, large samples make it difficult to achieve 'goodness of fit' which is judged acceptable only when p >.05 (notwithstanding Bonferroni corrections for multiple tests). Consequently, while the random sample was indistinguishable from the full cohort on any demographic or clinical variables (Table 27), there is a small possibility that they are not completely representative of all the UKROC stroke patients. Thus further research on the UK FIM+FAM and the Rasch model with more diverse samples is indicated, as well as exploration in other patient groups (e.g. traumatic brain injury).

Given promising results of applying MBIs to stroke populations (Lawrence et al., 2013), future research should replicate Rasch analysis of widely used mindfulness measures such as the FFMQ with clinical populations (i.e. stroke) to provide researcher and clinicians with tools for reliable assessment of mindfulness in clinical populations.

In conclusion, this analysis suggests that the UKFIM+FAM meets the Rasch model requirements with good reliability, acceptable targeting of each of the three domains, and with no item deletion in a population of complex stroke patients. A conversion table that accommodates DIF by stroke location is now ready for further evaluation in clinical practice and in research.

Chapter Ten. Generalizability Theory and the State - Trait Distinction

Introduction

Mindfulness can be conceptualized as either a state or a trait, but currently there is no reliable psychometric method to distinguish clearly between the two in psychological measures. Notwithstanding the clinical effectiveness of mindfulness, any specific element of mindfulness treatment can only be evaluated by comparing state and trait changes using techniques that allow such changes to be measured. Generalizability Theory (GT) is a suitable method to differentiate between state and trait variance components, and its application is illustrated here with an empirical example using the Toronto Mindfulness Scale (TMS). Person x occasion interaction is a marker of individual state changes and should explain the largest amount of variance in a valid state measure. To assess state variability, data were collected on three separate occasions: (i) after a holiday, (ii) immediately after a mindfulness exercise and before a stressful event (i.e. exam). Generalizability analysis was applied to examine sources of true and error variances. The TMS captured a larger amount of variance attributed to a state and only a small amount associated with trait mindfulness, which is consistent with the purpose of the measure. The study described in this Chapter has demonstrated that GT can be usefully applied to distinguish between state and trait components in a measure, and it is recommended as the appropriate psychometric method to validate state and trait measurement tools.

Mindfulness practice has become popular as a safe, non-invasive method for the management of stress and emotional problems and for the improvement of psychological and physical wellbeing (Chiesa & Serretti, 2010; Ivanovski & Malhi, 2007). With the increased application of mindfulness-based interventions, the accurate measurement of both a general tendency to be mindful (a *trait*) and the actual mindfulness level at any particular point in time (a *state*) has become an important clinical and research issue (Park et al., 2013). A trait is generally conceptualized as a relatively enduring characteristic of a person, while a state refers to a pattern displayed in a present moment situation, or condition (Hamaker et al., 2007; Spielberger et al., 1970). Therefore, a state is defined as interaction between person and occasion and describes a unique adaptation of a person to their immediate environment (Buss, 1989; Epstein, 1984). Lack of an appropriate methodology to distinguish clearly between state and trait measures affect both, reliability and validity of psychological measures such as mindfulness and health-related outcomes. Reliable measurement of state and trait mindfulness and related outcomes is necessary during both therapeutic interventions and neurophysiological studies (e.g. EEG) on mindfulness (Cahn & Polich 2006; Chiesa & Serretti 2010). Therefore, development of an appropriate methodology is important for distinguishing reliably between the two, otherwise effectiveness of therapeutic interventions cannot be evaluated over time.

The current method to evaluate state and trait scales has been merely to examine test-retest reliability coefficients. Generally, test-retest scores above .70 are considered as a characteristic of a trait measure and below .60 as an indicator that a scale is measuring state (Ramanaiah et al., 1983; Spielberger et al., 1970; Spielberger, 1999). This method is entirely based on a single correlation coefficient between total scores at two different occasions, which fails to account for variability due to interaction between person and occasion, which is an essential determinant of state changes in an individual (Buss, 1989; Epstein, 1984; Chaplin et al., 1988). Also, a correlation coefficient does not account for different contributions made by item effects, scale effects, person effects and occasion effects to changes in trait and state.

Essentially, trait scores are not expected to vary a great deal across situations. Instead, an interaction between the person and the occasion is naturally expected, which is a state by definition (Epstein, 1984; Chaplin et al., 1988). To date, the exploration of state and trait variability is limited to structural equation modelling (SEM) approaches (Geiser et al., 2015; Hamaker et al., 2007; Kenny & Zautra, 1995; Steyer et al., 1992) that are generally useful to study state-trait relationships. However, none of the proposed SEM methods account for various sources of variance (e.g. an item) contributing to the measurement error associated with state and trait variability, which limits their applicability for validation of state and trait measures. Such differences in variability require a more detailed study of how factors or components that can affect state and trait, including person and situation, can be quantified. That way, changes in state and trait can be predicted by knowing of changes in person and situation, which is a true generalisability, in other words.

GT is a statistical method used to analyse data collected by means of psychometric measures. It provides techniques to estimate the generalisability of the influence due to any specific factor (e.g. occasion) based on limited amount of data collected from a specific testing situation to all possible situations and contexts (Cronbach et al., 1963). Unlike CTT, GT accounts for numerous sources of variance contributing to the measurement error associated with the main variable of interest (e.g. a mindfulness score). Thus, GT offers an accurate method to evaluate various factors and their interactions contributing to measurement error leading to the improvement of methodology and precision of an assessment instrument (Allal & Cardinet, 1976). GT and its applicability to distinguish between state and trait variance components in a measure are described in greater detail in Chapter One.

This study applies GT to extract and evaluate the amount of variance uniquely explained by the person, the item and the occasion plus their respective interactions (Brennan, 2001; Bloch & Norman, 2012). The 'stateness' of a measure is directly reflected by person-occasion interaction and the 'traitness' by variance due to a person (Buss, 1989; Epstein, 1984; Chaplin et al., 1988). In the current study GT analysis is used to examine both total scales and individual items, which is a unique feature of GT method. GT analysis of individual items can reliably distinguish between true 'state items' and items that are not truly sensitive to occasion. A G-study was conducted to estimate variance of persons, which is a common object of measurement in psychometrics, and influencing facets such as occasions, items, and related interactions. Estimation of variance components was based on observed values acquired from the universe of all possible observations. State measures/items should reflect a large amount of variance due to person-occasion interaction and low generalisability across occasions (e.g. G <.70). In contrast, reliable trait measures/items should be stable over time reflected by greater generalisability of scores (G ≥ .80) (Arterberry et al. 2014; Gardinet et al. 2009).

Currently, there are no commonly accepted criteria for the relative proportions between state and trait variance components in a valid state or trait measure. Therefore, we propose the state component index (SCI) to estimate this relationship as follows:

$$\mathrm{SCI} = \frac{\sigma_s^2}{\sigma_s^2+\sigma_t^2} \qquad (9)$$

In the above formula, the variance component of a state (σ_s^2=σ_{po}^2) is essentially the noise or error variance due to person-occasion interaction that affects trait scores. This reformulation of the original ratio equation is essentially identifying the ratio of state to trait including noise in both which we can assume be equal because the trait (persons) component (σ_t^2=σ_p^2) is the basic component of the state variance. To ensure accuracy of measurement, the SCI calculation should use an absolute value of variance due to person-occasion interaction derived from G-analysis that accounts for all sources of error variance identifiable in the data. SCI is developed in line with GT logic and is easy to interpret. For instance, SCI=1.00 would mean that there is no trait component and only individual state is measured, which appears unlikely because a trait is a basic predictor of a state (Buss 1989; Epstein 1984). SCI=.50 would mean that state and trait components are the same and a scale cannot be classified as either state or trait measure. However, SCI>.60 can be considered as characteristic of a state measure with higher scores corresponding to better ability of an instrument to capture state changes. Similarly, trait component index (TCI) can be used to validate a trait measure using the same metric:

$$\mathrm{TCI} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_s^2} \qquad (10)$$

Therefore, more precise distinction between scales measuring states and traits can be made based on G-study results. The D-study (decision study) is based on G-study results and involves experimenting with designs (e.g. fixed or random) in an attempt to reduce measurement error (Brennan 2001; Shavelson et al. 1989). It can be used to identify those items that are not consistent with the purpose of the measure (e.g. items measuring trait in a state measure) and thus to improve an instrument by removing them.

While GT was applied to assess reliability of trait measures (e.g. Arterberry et al. 2014), we are not aware of any studies to date that have used GT methods to distinguish between state and trait components in a state measure. The aim of this study is to demonstrate application of GT to investigate state- and trait-related variance components in the Toronto Mindfulness Scale (TMS) (Lau et al. 2006), the first and the most frequently cited instrument designed exclusively to assess state mindfulness. Rasch analysis (Rasch, 1960, 1961) was used to assess the general psychometric properties of the TMS subscales from the perspective of Item Response Theory (Allen & Yen, 1979) and their suitability for parametric GT. In particular, the appropriate ordering of item thresholds and item fit to the Rasch model were investigated. However, no ordinal-to-interval conversion was implemented because that would limit GT investigation to total scores properties and prevent individual item analysis. GT analysis was based on the procedure described elsewhere (Gardinet et al., 2009; Bloch & Norman, 2012). Two-way repeated measures ANOVA was used in the G-study design to assess the variance due to object of measurement (persons) and sources of error variance due to occasion, item, person-occasion, person-item and person-occasion-item interactions of the TMS subscales. It was expected that individual scores would have low generalizability across occasions (G<.70) (Arterberry et al. 2014) and high amount of variance due to person-occasion interactions reflected by the proposed SCI above .60 as characteristics of a valid state measure. The D-study was conducted to demonstrate how the functioning of the TMS subscales and individual items can be investigated and optimized by varying facets designs.

Method

Participants

The sample size (*n*=55) satisfied criteria for a reliability study in medical research (Shoukri, Asyali & Donner, 2004) and is adequate for generalizability analysis because G-coefficients are essentially similar to reliability coefficients (Bloch and Norman 2012). Given the experimental nature of this study, where the focus is on an initial measurement of the sample followed by an intervention that is subsequently measured, no attempt was made to set up a control group. Also, any biases introduced by the convenience sampling method involved (all participants were locally available and indicated willingness to participate) are assumed to be distributed evenly throughout the sample. All 55 participants, who provided data at three different occasions, were New Zealand university students, (78.2% females, 21.8% males) with a mean age of 23.44 (SD=6.32) and range of 18 to 44. Ethnic groups include Caucasians (49.1%), Polynesians (16.4%), Asians (14.5%) and other ethnicities (20 %).

Procedure

Potential participants were approached during lectures and invited to complete the survey on three different occasions and to hand the survey directly back to the researchers or submit it to a locked collection box at their respective faculty. Three occasions were chosen to increase variability of state mindfulness and data were collected ‘after a holiday’, ‘after a mindfulness exercise’ and ‘before a stressful event’. On the first occasion, the first lecture after the summer holiday served as the baseline. Here, students completed the questionnaire in class before the lecture or during a short lecture break. The second occasion occurred after a one-week interval, where students completed the questionnaire at the beginning of laboratory classes in a different environment and in smaller groups. Prior to completing the questionnaire on occasion 2, students participated in a 10-minute guided mindfulness exercise called ‘body scan’, which is a standard component of Mindfulness-Based Cognitive Therapy (MBCT) (Segal et al. 2013). It was expected that the mindfulness exercise would increase or at least influence mindfulness levels of the participants. To ensure the same conditions across lab classes and to minimize experimenter effects, the ‘body scan’ exercise instructions were played to the participants from the audio CD included in the book ‘Mindfulness: Finding Peace in a Frantic World’ (Williams and Penman 2011). On the third occasion, which occurred after a one-month interval after the first data collection, students completed the questionnaire in the lecture theatre before the lecture. This occasion was a week before an important class test, and the lecture included the test overview and relevant discussion. It was expected that students would have higher stress levels on this occasion, which might impact on their mindfulness levels. The students were asked to create a unique ID containing letters and numbers (e.g. ABC123), which could not be used to identify them but could be used to anonymously match the questionnaires completed by the same person on three different occasions. The authors’ university ethics committee had approved this study.

Instrument

The Toronto Mindfulness Scale (TMS) (Lau et al. 2006) is a 13-item self-report questionnaire designed to measure two dimensions of state mindfulness: curiosity and decentering. The former is defined as present-moment awareness with a quality of curiosity, while the latter refers to awareness of one’s experience from a distant observer perspective and thus without identifying oneself with the content of one’s thoughts and feelings and getting carried away by them (Lau et al. 2006). Meditators scored higher on both TMS subscales compared to those without meditation experience, and Decentering scores were shown to reflect meditation experience (Davis et al. 2009) and changes in psychological symptoms (Lau et al. 2006). Both TMS subscales displayed increased scores after mindfulness training, which provide support for their construct validity, although no test-retest reliability scores were reported (Park et al. 2013). The TMS includes a 6-item Curiosity subscale (Cronbach’s alpha .86-.91) and a 7-item Decentering subscale (Cronbach’s alpha .85-.87) (Park et al. 2013). Both subscales use a 5-point Likert-scale response format (0=‘Not at all’ to 4=‘Very much’). The total subscale scores are calculated by adding responses to individual subscale items with higher scores corresponding to higher levels of state mindfulness.

Data Analysis

Descriptive statistics together with Cronbach's alpha coefficients and test-retest bivariate correlations for the Curiosity and Decentering subscales of the TMS were computed using IBM SPSS version 23 at each of the three assessment occasions. Test-retest reliability scores for a state measure were expected to be in the range from .16 to .57 (Ramanaiah et al., 1983; Spielberger, 1999).

Rasch analysis was conducted using RUMM2030 software (Andrich, Sheridan, & Luo, 2009) to assess psychometric properties of the TMS subscales at the overall and the individual item level. Stacked data from three occasions were used for Rasch analysis to account for score variability due to state changes (Wright, 2003). In particular, both TMS subscale items were screened for disordered thresholds and fit residuals, which should be in the range between -2.50 and +2.50. The Rasch analysis followed the sequential steps described in Chapter Two, but did not include generation of conversion tables because its purpose was to test appropriateness of the raw data for the following parametric GT method.

GT analyses were conducted using EduG 6.1-e software (Swiss Society for Research in Education Working Group 2006) that produces an extended output, which is easier to interpret in practical terms. We employed a random effects design with two crossed facets for both G and D-study: persons (P), by occasion (O), by item (I), expressed as P x O x I, where the P and O facets are infinite and the I facet is fixed. The facets were defined from the trait perspective with persons as the object of measurement, which is a facet of differentiation, and items and occasions as instrumentation facets (Gardinet et al. 2009). States are expected to vary across occasions reflected by person-item interaction, but not across items. Here, the error variance attributed to interaction between person and occasion (P x O) will be indicative of a state component in a scale score, which is expected to be relatively strong for a state measure.

Conventional ANOVA was used to compute sums of squares, mean squares, variance components, variance percentages associated with each facet including standard errors. Variance components were estimated for each effect based on their mean squares and samples to assess measurement error due to each of the sources using formulas developed by Brennan (1977, 1992). Variance components are estimated by EduG after applying a kWhimbey's correction to classical ANOVA estimates that accounts for facets, which are not sampled from infinite universes (e.g. scale items) (Gardinet et al. 2009). It is expressed as ((N(f)-1) / N(f)), where N(f) is the universe size of the f facet in the G-study design and has no effect on merely random facets.

Generalizability analysis was applied to estimate contribution of each facet to variance of universe scores including relative and absolute error variance and to calculate relative and absolute G-coefficients for the object of measurement (persons). Relative G-coefficient only accounts for variance directly influencing a relative measurement tool (e.g. person-occasion and person-item interactions) (Shavelson et al. 1989) and may express commonly used ρ^2 , ϖ^2 or intermediate value by virtue of using Wimberley's correction (Gardinet et al. 2009):

$$G_{relative} = \frac{\sigma_p^2}{\sigma_p^2+\sigma_\delta^2} \tag{11}$$

Here, σ_p^2 is variance due to object of measurement (persons) and $\sigma_\delta^2 \ = \sigma_{po}^2 \ + \sigma_{pi}^2 + \sigma_{poi}^2$ is relative error variance. Absolute G-coefficient ($G_{absolute}$) is similar to the commonly used Phi (Φ) coefficient after applying Wimberley's correction. It accounts for absolute error variance (σ_Δ^2=$\sigma_o^2 + \sigma_i^2 + \sigma_{io}^2 + \sigma_{po}^2 \ + \sigma_{pi}^2 + \sigma_{poi}^2$) that includes other factors (e.g. items and occasions) influencing an absolute measure (Gardinet et al. 2009):

$$G_{absolute} \approx \Phi = \frac{\sigma_p^2}{\sigma_p^2+\sigma_\Delta^2} \tag{12}$$

Also, the SCI to estimate relationship between state and trait variance components was computed using the formulae proposed in the introduction. D-study included facets analyses of every individual item to estimate variance components and G-coefficients associated with the object of measurement (persons or trait), and variance due to person-occasion interaction as a state marker. It also involved testing various facet designs by manipulating their levels to optimize the instrument.

Results

All data distributions met normality assumptions with skewness and kurtosis values fairly close to zero and non-significant Shapiro-Wilk normality tests. Repeated-measures ANOVA indicated that the effect of occasion was significant for both facets of state mindfulness: Curiosity (*F*(2,54)=6.88, *p*=.002, η^2=.11) and Decentering (*F*(2,54)=12.46, *p*=.001, η^2=.19). Post-hoc tests showed that the mean Curiosity and Decentering levels on Occasion 2 (one week, after mindfulness exercise) were significantly higher compared to both other occasions. Table 31 presents descriptive statistics together with Cronbach's alpha coefficients and test-retest bivariate correlations for the Curiosity and Decentering subscales of the TMS at each of the three assessment occasions. While the Curiosity subscale showed good internal consistency at all three occasions, the Decentering alpha coefficients varied but in the acceptable range from .70 to .80. According to expectations, test-retest reliability coefficients for both subscales at one week and one month intervals ranged from .38 to .46 (Table 31).

Table 31. *Means, standard deviations (SD), internal and test-retest reliability estimates for the TMS[a] Curiosity and Decentering subscales (n=55).*

Subscale / Measurement	Baseline	1 Week	1 Month
Curiosity			
Mean (SD)	10.04 (5.08)	12.05* (5.73)	8.91 (5.36)
Cronbach's alpha	.83	.87	.88
Test-retest (*r*)[b]	--	.38	.34
Decentering			
Mean (SD)	10.09 (4.80)	13.44* (5.62)	10.36 (5.12)
Cronbach's alpha	.70	.80	.79
Test-retest (*r*)[b]	--	.44	.46

Note:*mean is significantly different from other means (p<.05); [a]TMS=The Toronto Mindfulness Scale; [b] Test-retest correlations between the baseline scores and scores after 1 Week and 1 Month intervals.

Rasch Analysis of TMS subscales

Rasch analysis was conducted to check the suitability of the TMS subscales for application of GT. The TMS Curiosity subscale instantly fitted to Rasch model expectations with acceptable reliability measured by person separation index (PSI=.84) and non-significant Chi Square of item-trait interaction ($\chi^2(12)=12.79$, $p=.38$). The mean of the item-fit residuals was acceptable (0.33, SD=.51), and the mean of the person-fit residuals indicated a moderate fit (-0.59, SD=1.67). There were no misfitting items, and no disordered thresholds were identified as can be seen in the top panel of Figure 23, which shows the thresholds map for the Curiosity items ordered by location.

The TMS Decentering had acceptable reliability (PSI=.77) and fit residuals for item (0.46, SD=1.12) and person (-0.42, SD=1.42). Similar to the Curiosity subscale, all items displayed good model fit and no disordered thresholds were evident (Figure 23, bottom panel). Even though the Chi Square value was relatively low, item-trait interaction was still significant ($\chi^2(14)=25.11$, $p=0.03$). However, the exact amount of variance attributed to item-trait interaction is better assessed and evaluated in the GT analysis that follows, and we thus decided to avoid any modifications of the original subscale. Thus, Rasch analysis confirmed appropriateness to use the TMS data for generalisability (i.e. parametric statistical) analysis.

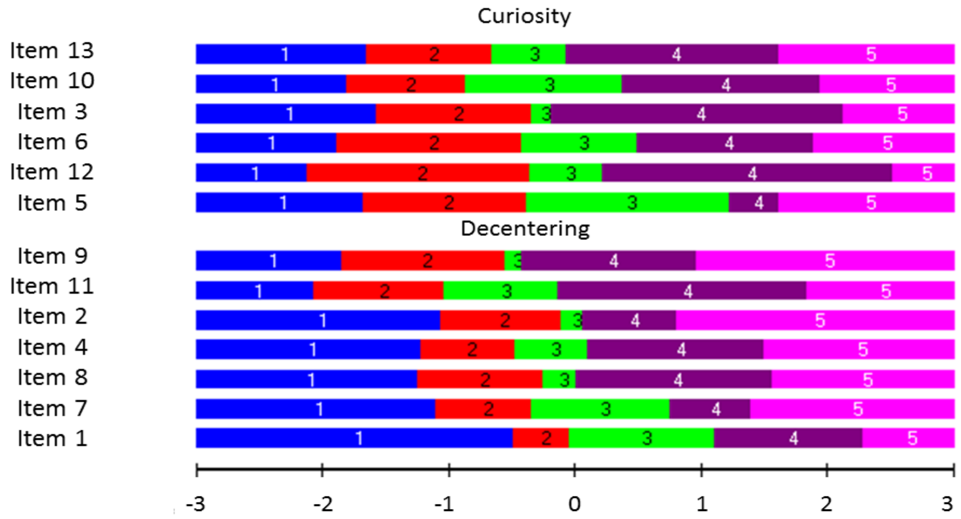


Figure 23. Threshold maps for the TMS Curiosity (above) and Decentering (below) subscales.

G-Study

ANOVA results for the TMS Curiosity and Decentering subscales together with variance components attributed to person (P), item (I) and occasion (O), and interactions between them are included in Table 32 and provide basic estimates for the G-study. Corrected variance components included in columns seven and eight (in %) are computed by applying Whimbey's correction. Relative and absolute contribution of the percentage values presented in column 8 (Table 32) were estimated from a GT perspective and are presented in Table 33. The largest amount of variance of both subscales scores was explained by person-occasion interactions, which is a marker of individual state changes in domains of curiosity and decentering across three different occasions.

The results of a generalizability analysis of both Curiosity and Decentering TMS subscales are presented in Table 33. Components that cannot be computed (as they did not exist) in the current design are represented as a row of dots. As predicted for a valid state measure, person-occasion (P x O) interaction is the main source of error variance for both subscales explaining over 90% of relative and absolute error variance.

Table 32. *ANOVA for the Curiosity (above) and Decentering (below) subscales of the TMS including sum of squares (SS), degrees of freedom (df), mean squares (MS), variance components (in %) and standard errors (SE) for the Person (P) x Occasion (O) x Item (I) design including interactions (n=55).*

				Curiosity Variance Components				
Source	SS	df	MS	Random	Mixed	Corrected [a]	%	SE [b]
P	341.44	54	6.32	0.10	0.10	0.10	6.70	0.07
O	1.53	2	0.76	0.00	0.00	0.00	0.00	0.00
I	11.97	5	2.39	0.01	0.01	0.01	0.70	0.01
P x O	**489.80**	**108**	**4.54**	**0.64**	**0.69**	**0.69**	**46.50**	**0.10**
P x I	179.20	270	0.66	0.00	0.00	0.00	0.00	0.02
O x I	6.79	10	0.68	0.00	0.00	0.00	0.00	0.01
P x O x I	371.88	540	0.69	0.69	0.69	0.69	46.10	0.04
Total	1402.61	989					100%	
				Decentering Variance Components				
P	267.87	54	4.96	0.06	0.06	0.06	3.70	0.05
O	1.28	2	0.64	0.00	0.00	0.00	0.00	0.00
I	105.89	6	17.65	0.10	0.10	0.09	6.10	0.05
P x O	**408.63**	**108**	**3.78**	**0.41**	**0.54**	**0.54**	**31.60**	**0.07**
P x I	280.02	324	0.86	0.00	0.00	0.00	0.00	0.03
O x I	11.34	12	0.94	0.00	0.00	0.00	0.10	0.01
P x O x I	576.76	648	0.89	0.89	0.89	0.89	58.40	0.05
Total	1651.78	1154					100%	

Note: [a] Corrected components are computed by applying Whimbey's correction to the ANOVA estimates.
[b] SE in the right column is related to the mixed effects presented in the column 6.

The final results show that relative and absolute G-coefficients are both below the acceptable level of .80 recommended for assessment of traits and in line with expectations for a state measure. The proposed SCI values were calculated based on differentiation variance of person (trait: $\sigma_t^2 = \sigma_p^2$) and absolute error variance of person-occasion interaction (state: σ_s^2=σ_{po}^2) for the Curiosity (SCI=.70) and for the Decentering (SCI=.75) subscales. These values indicate that, after accounting for all sources of error identifiable in the data, both subscales mainly reflect variance associated with state changes in line with expectations for a valid state measure.

Table 33. *Estimated variance components of the TMS Curiosity and Decentering subscales with standard errors (SE) and G-coefficients for the G-study P x O x I design.*

Source of Variance	Differentiation variance	Relative error variance	% relative	Absolute error variance	% absolute
Curiosity TMS Subscale[a]					
P	**0.10**	.....		.....	
O	.....	.....		0.00	0.00
I	.....	.....		0.00	0.40
P x O	.....	**0.23**	**91.20**	**0.23**	**90.90**
P x I	.....	0.00	0.00	0.00	0.00
O x I	.....	…..	…..	0.00	0.00
P x O x I	…..	0.02	8.80	0.02	8.80
Sum of variances	0.10	0.25	100%	0.25	100%
Standard deviation	0.32	Relative SE:	0.50	Absolute SE:	0.50
G relative 0.28					
G absolute 0.28					
Decentering TMS Subscale[b]					
P	**0.06**	.....		.....	
O	.....	.....		0.00	0.00
I	.....	.....		0.00	0.00
P x O	.....	**0.18**	**100.00**	**0.18**	**100.00**
P x I	.....	0.00	0.00	0.00	0.00
O x I	.....	…..	…..	0.00	0.00
P x O x I	.....	0.00	0.00	0.00	0.00
Sum of variances	0.06	0.18	100%	0.18	100%
Standard deviation	0.24	Relative SE:	0.43	Absolute SE:	0.43
G relative 0.24					
G absolute 0.24					

Note: [a]Curiosity (n=55, Grand mean: 1.72, SE of the grand mean: 0.09);
[b]Decentering (n=55, Grand mean: 1.61, SE of the grand mean: 0.11).

D-Study

Facets analysis was conducted first, to obtain variance estimates for every individual item by excluding all other items. The estimates for a differentiation facet of a person together with estimates for person-item interaction and G-coefficients are included in Table 34. In line with expectations for items measuring state, most of the items show a high amount of variance attributed to person-item interaction and typically above 0.4 with the exception of items 1 and 11, which are just below this benchmark. Low differentiation estimates (P) were found for most of the items consistently reflected by the low values of the G-coefficients in the right column, which are both expected to be high for a trait measure (i.e. G-coefficient above .80). However, two items, 4 and 7 in the Decentering subscale did not reflect any variance attributed to a trait (person) and consequently had generalizability coefficients of zero. Therefore, we tested the relative contribution of these items to the Decentering subscale by removing them. After removing those two items the proportion of variance due to person-occasion interaction decreased from 100% (Table 34) to 79.1% and produced an additional 19.80% error variance attributed to person-occasion-item interaction, which is a threat to scale reliability. Also, removing those items did not affect the G-coefficients remaining at the same level of 0.24 (relative) and 0.23 (absolute). This illustrates that items 4 and 7 contribute to the overall reliability of the Decentering subscale in discriminating between state levels.

Table 34. *Estimated person and person x occasion (P x O) interaction variance components together with G-coefficients for the items of TMS Curiosity and Decentering subscales.*

TMS Subscales and Items	P Variance	P x O Variance[a]	G Coefficients[a]
Curiosity subscale			
3. curious to learn about myself by noticing my reactions	0.06	0.51	0.11
5. curious to see what my mind was up to from moment to moment	0.04	0.41	0.09
6. curious about each of the thoughts and feelings I was having	0.07	0.44	0.13
10. curious about the nature of each experience as it arose	0.09	0.43	0.17
12curious about my reactions to things	0.09	0.41	0.19
13. curious to learn about myself by noticing my attention focus	0.20	0.46	0.30
Decentering subscale			
1. experienced myself separate from thoughts & feelings	0.15	0.29	0.34
2. more concern with being open to experiences than controlling	0.12	0.52	0.19
4. experienced my thoughts more as events than as reflection	0.00	0.49	0.00
7. observing unpleasant thoughts and feelings without interfering	0.00	0.44	0.00
8. more invested in watching my experiences than analysing them	0.04	0.46	0.08
9. trying to accept each experience, pleasant or unpleasant	0.04	0.48	0.08
11. aware of thoughts and feelings without overidentifying with them	0.03	0.35	0.07

Note: [a] There is no difference between relative and absolute P x O variance components and G-coefficients in P x O design because there are no finite populations.

Removing individual items from each subscale did not result in an increase but in some cases decreased the overall generalizability coefficients. Finally, removing Occasion 3 (before the class test) slightly increased G-coefficients in the Curiosity subscale up to .44 (absolute and relative) and removing occasion 1 (Baseline, after the holiday) decreased the overall G-coefficients of both subscales just below .10 (absolute and relative). Removing Occasion 2 (1 Week, mindfulness exercise) did not result in any substantial changes of the overall G-coefficients.

Discussion

The aim of this study was to demonstrate the application of GT to distinguish between state and trait variance components in a measure using the TMS as an example. This study has demonstrated that Generalizability Theory can be applied to distinguish between state and trait components in a measure, and it is recommended as the most appropriate psychometric method to validate state and trait measurement tools. The method and the sequence of analysis illustrated in the Results section allows researchers to assess the validity and reliability of any psychometric measure of a state or a trait using GT.

Currently, the only statistical method used to distinguish between state and trait measures is a single correlation between total test scores at two different occasions (test-retest). The proposed GT method is based on an accurate estimation of variance components of both state and trait that accounts for various sources of error variance and provides an advanced alternative for validation of state and trait measures. It is particularly powerful in its ability to examine the 'stateness' or 'traitness' of each individual item.

To demonstrate the application of GT, we used a state measure of mindfulness, the TMS (Lau et al. 2006). We chose this measure because, while GT has already been used to assess the reliability of trait measures (Arterberry et al. 2014), it has not previously been used to distinguish between state and trait mindfulness. Before using the TMS to illustrate the application of GT methods, reliability and construct validity of the instrument were tested using more traditional methods and supported by the results (Table 31). Prior to GT analysis we also ensured that the data met assumptions of normality and confirmed acceptable psychometric properties of the TMS subscales and individual items using Rasch analysis. Although, not the main purpose of the study, the results provide support for construct validity of the TMS as a state measure as the scores followed predicted changes, namely increased mindfulness after a brief mindfulness exercise and decreased mindfulness during a stressful pre-exam period. These findings are consistent with Lau et al. (2006) who reported an increase of the TMS scores following a mindfulness-based intervention.

In this G-study two-way repeated measures ANOVA was used first to extract the variance due to the object of measurement (persons) reflecting a trait, person-occasion interaction reflecting a state, and other sources of error variance such as occasion, item and interactions of the TMS subscales. Such ANOVA results are important because they provide basic estimates for further analysis. In terms of a state-trait distinction, a trait measure should have the largest amount of variance explained by the person and a state measure, as in this case, by the person-occasion interaction (Table 32). However, traditional ANOVA is not precise enough to identify such individual contributions. For instance, it can be seen that variances due to person-item and occasion-item interactions are close to zero for both TMS subscales, suggesting that the variance due to person-occasion-item interaction is mainly explained by person-occasion interaction or a state. Therefore, subsequent G-analysis is necessary to estimate the unique contribution of each variance component available in the data together with G-coefficients.

G-analysis estimates variance components and G-coefficients in both relative and absolute terms. The essential difference between them is that absolute estimates will account for all possible error variances assuming that all samples are drawn from infinite populations but relative estimates will account for finite populations in the G-study design (e.g. items). In other words, if all populations are considered as drawn from infinite populations absolute and relative variance estimates and G coefficients will have the same values. In the current analysis (Table 33) G-coefficients are the same because error variance due to item, which is the only finite dimension, is close to zero.

One of the possible reasons why GT has not been widely used to validate state measures is possibly because person-occasion interaction is considered as a measurement error in common G-designs with persons representing the principle object of measurement. This common design was used in the current study to demonstrate its limitations and the advantages of introducing the SCI to assess 'stateness' of a state scale along with the TCI to assess 'traitness' for a measure of a trait. For instance, G-analysis (Table 33) shows error variance estimates due to different sources after accounting for the person (trait) variance. Here, error variance in both TMS subscales was mainly attributed to person-occasion interaction reflecting state changes, which is expected for a valid state measure. In the current G-analyses, person (trait) variance is assessed by G-coefficients showing values below .30 indicating that the TMS scores were unstable across occasions, which is consistent with expectations for a state measure. The G-analysis results mirror the traditional test-retest reliability findings and were consistent with those reported earlier for other state measures, such as a range of *r* values from .34 to .46 for the State Trait Anxiety Inventory (Ramanaiah et al. 1983; Spielberger 1970, 1999). In the case of a valid trait measure, where persons (traits) explain the most variance and show stability over time, G-coefficients of .80 and higher would be expected (Arterberry et al. 2014).

The proposed SCI is particularly useful to assess the degree of 'stateness' of a measure especially if a common G-design with persons as objects of measurement is used because person-occasion interaction (state) is often treated as a measurement error in such designs. Similar to other G-estimates, the SCI was calculated based on the corrected variance components from the ANOVA (Table 32). The SCI for the Curiosity subscale was .70 and for the Decentering .75, which is consistent with expectations for a valid state measure and arguably provides the first benchmark to distinguish between instruments measuring state and trait. An SCI below .60 would suggest that there are items in a scale, which are not sensitive to state changes (i.e. measuring a trait). In this case modifications of an instrument should be undertaken using D-study. Similarly, the TCI can be computed to assess validity of a trait measure and modifications could be conducted if a value below .60 is obtained.

Besides exploration of state and trait variance components, GT analysis is also useful to identify potential sources of measurement error. In our example the results show that error variance due to items and person-item interaction did not exceed 1%. Overall, the error variances were close to zero with the exception of interaction between person, occasion, and item in the Curiosity subscale, which constitute only 8.80% with the other 92.20% explained by the state (person-occasion) component. However, both person-item and occasion-item errors were nearly zero suggesting that this error is due to state-item interaction only. If this GT method is applied to other measures, identifying sources of measurement errors can be useful especially if the values exceed 5% and hence affect the precision of a measurement. In this case, a source of measurement error (e.g. items) could be investigated in a D-study and necessary adjustments could be made to resolve the issue.

A D-study can be used to improve measurement design and to address potential issues contributing to measurement error, which is especially useful at the individual item level. Our D-study examined state and trait variance components of every individual item (Table 34) and showed that all items displayed a higher proportion of variance attributed to state compared to trait and low generalizability of scores across occasions. These findings are generally consistent with the G-study results for the complete subscales. However, items 4 and 7 in the Decentering subscale showed no signs of differentiating between individual's trait levels reflected by lack of generalizability in measuring trait. Typically, a moderate or at least a weak relationship between state and trait components is expected in a state measure (Ramanaiah et al. 1983; Spielberger 1970, 1999). Excluding those two items from the subscale was associated with a decrease in state-related variance and increase of the error variance affecting the reliability of the subscale. Therefore, items 4 and 7 were found to measure state changes only and contributed to the overall reliability of the Decentering subscale. These findings challenge the assumptions that the trait component cannot be entirely excluded in a state measure because it is the basic predictor of a state (Hamaker et al. 2007; Kenny and Zautra 1995). Assessing variance components at the individual item level could be useful because a measure may include items measuring predominantly a trait, a state or both. In this case state and trait items could be combined into a state and a trait subscale respectively, and neutral items excluded from the measure, which will improve accuracy in assessing state and trait.

A D-study is also useful to evaluate the appropriateness of a G-study design and the individual contribution of occasions on variability of states. For instance, removing the baseline (after holiday condition) produced a decrease of generalizability of both subscales across occasions below 0.10. This result is expected if state changes are manipulated at both occasions in the opposite direction (mindfulness exercise vs class test) and supports the appropriateness of the G-study design. Finally, attempts to optimize subscales by removing items did not yield any psychometric benefits suggesting that the TMS is an adequate measure of state mindfulness in its present form.

Limitations and Conclusions

The following limitations have to be acknowledged. The proposed SCI and TCI indices for validation of state and trait measures are based on the results of this study and need to be extensively tested with different instruments to establish benchmarks and cut-off points. More accurate criteria for state and trait distinctions might evolve as a result of further GT analyses of other psychometric instruments. This study was conducted with a sample of university students that has a degree of homogeneity, and the results should be replicated with larger and more diverse samples. Generalizing the results of this study (state vs trait) to the rest of the population may be limited without a truly representative sample.

In summary, the current study developed and introduced a novel and promising method to distinguish between state and trait measures using GT. The application of this method was demonstrated by generalizability analysis of the TMS - state measure of mindfulness and provided supporting evidence for reliability and validity of the instrument. The current application of GT is recommended as the appropriate psychometric method to validate state and trait measurement tools and has the potential to open new avenues for future psychometric work.

Chapter Eleven. Integrated Conclusion

The principal aim of this work was to test and improve the reliability and validity of leading mindfulness and health outcome measures used in research and health practice by applying Rasch analysis and GT. Rasch analysis was applied to investigate and enhance the psychometric properties of seven mindfulness and three health outcome measures up to an interval level scale. Application of GT resulted in the development of a novel method for more accurate distinction of state from trait in psychometric measurement. Although CTT methods were useful in the original development of the psychometric instruments investigated in this work, both Rasch and GT methods are now essential for rigorous validation of mindfulness and outcome measures.

Contributions to Mindfulness and Outcome Measurement

One major contribution of this work was the provision of practical solutions to improve the reliability and structural validity of seven ordinal scales: four measuring trait mindfulness and three measuring health outcomes. Four of the measures that were investigated (the KIMS, the FFMQ, the CHIME and the UK FIM+FAM) were multidimensional and three (the MAAS, the PSS, the OHQ) unidimensional. Rasch analysis applied state-of-the-art methodology to evaluate and improve the psychometric properties of both these scales and their individual items and contributed valuable information about their reliability and structural (construct) validity. Therefore, necessary modifications were made to modify scales in order to solve identified issues. As a result, ordinal-to-interval conversion algorithms were produced for all these measures, which were published as ordinal-to-interval conversion tables that are relatively easy to apply and that do not require any modification of the original scales' formats.

Contribution by Applying Advanced Methodology

A methodological strength of this work was to apply Rasch analysis using modern strategies, which were developed over the last 21 years of experience reported in Rasch studies (Lundgren Nilsson & Tennant, 2011; Lundgren Nilsson et al., 2013). The aim was to improve the psychometric properties of the measures while keeping modifications to a minimum to retain their clinical face validity. Historically Rasch analysis studies have often relied on removing misfitting items in order to achieve a satisfactory model fit (Stewart-Brown et al., 2009; Goh et al., 2015). It was argued in Chapter One that IRT and Rasch methods produce shorter scales compared to CTT techniques, but reduction of a scale should not be achieved at the expense of reliability and construct validity of the instrument (Cohen & Swerdlik, 2010). Therefore, items should only be removed in cases of clear redundancy, semantic and/or conceptual inappropriateness, and misfit that cannot be corrected without removing misfitting item (Nunnally & Bernstein, 1994; Lundgren Nilsson, & Tennant, 2011; Lundgren Nilsson et al., 2013).

In this work, removing an item was considered as the last resort to achieve a satisfactory model fit and it was carefully considered by examining both individual item fit and the residual correlation matrix for evidence of local dependency among items. In addition, the appropriateness of an item's semantic item content was also closely examined. For instance, the strategies used to minimise modifications of original scales include the use of analytical pathways with and without rescoring, and the creation of subtests (testlets) combining two or more items to solve local dependency issues. As a result, in three (the CHIME, the PSS and the UK FIM+FAM) out of the seven analysed scales, the best fit to the Rasch model was achieved without removing any of the original items. In other scales, on average, only three misfitting items per scale were removed to improve their psychometric properties and the overall model fit. When conceptually important items did not work well psychometrically (e.g. the KIMS, the FFMQ), re-wording of the relevant items was recommended. Interestingly, the best Rasch model fit for the UK FIM+FAM was achieved using testlets and without rescoring disordered thresholds, which suggests that rescoring is not necessarily contributing to a better fit if testlets are applied. This can be explained by the advantages of using combined items, such as more scale points that contribute to accuracy of measurement and higher reliability compared to individual items (Little et al., 2002; Rushton et al., 1983).

Contribution by Establishing Structural Construct Validity

Testing internal construct validity of unidimensional measures such as the MAAS, the PSS and the OHQ is straight-forward, if the data fits the Rasch model and unidimensionality is confirmed then internal construct validity is supported (Tennant and Conaghan, 2007). Examining structural construct validity of multidimensional measures is more complicated (Chapters Four to Six and Nine) and involves fitting the full scale to the Rasch model by treating individual subscales as subtests using the methodology of Lundgren Nilsson et al. (2013). Essentially, this method tests the hypothesis that, if an overarching latent construct (e.g. mindfulness) is accurately defined by facets/subscales and items of each subscale are combined into a subtest, then the full scale should fit the

Rasch model. Combining items of individual subscales is a way of dealing with local dependency at the full scale level because items measuring a specific mindfulness facet (e.g. describing) might often be expected to exhibit some degree of local dependency due to shared variance. This expectation is normally confirmed in Rasch analysis by examining the residual correlation matrix. Unlike the KIMS (Chapter Four), which failed to meet the expectations of the unidimensional Rasch model when the subscales were treated as subtests, the modified FFMQ demonstrated good model fit that supported its structural validity (Chapter Five). These findings allow researchers to calculate a reliable and valid interval-level mindfulness score for the total FFMQ using the generated conversion algorithms. This ultimately resolves considerable debate about reliability and validity of the total FFMQ score and the meaningfulness of using it in research (Baer et al., 2006; Park et al., 2013). Similar to the FFMQ, the CHIME and the UK FIM+FAM have demonstrated good psychometric properties of their individual subscales and the full scale.

Contribution by Evaluation of Psychometric Properties

The current Rasch analysis has demonstrated psychometric advantages of the FFMQ compared to the MAAS and the KIMS. Specifically, its subscales have better coverage of individual abilities and have overall higher reliability (PSI). Also, the FFMQ Act With Awareness subscale includes most of the MAAS items and thus, covers this aspect of mindfulness (Baer et al., 2006, 2008). Overall, two multidimensional mindfulness measures, namely the FFMQ and the CHIME, have demonstrated superior psychometric properties, strong correlations between their total scores and the corresponding subscale scores (Baer et al., 2006; Bergomi et al., 2014). The five facets of the FFMQ were developed by factor analysis of the best mindfulness measures available at that time (Baer et al., 2006). Since then, new mindfulness measures have been developed and their constructs were considered by Bergomi et al. (2014) in developing the eight facets CHIME. The CHIME and the FFMQ have similar facets measuring acting with awareness, non-judgmental acceptance and non-reactive attitude (which includes a decentering aspect in CHIME). However, the CHIME does not include a describing facet, which was identified as inconsistent with common mindfulness definitions (Table 1) but adds openness, relativity of thoughts, insightful understanding and external and internal awareness similar to observing of the FFMQ. Given that the three mindfulness facets (acting with awareness, non-judgmental acceptance and non-reactive attitude) shared by these measures were found as reliable predictors of psychological symptoms across different studies (Baer et al., 2006, 2008; Park et al., 2013; Bergomi et al., 2014), they seem to represent the core components of mindfulness. Even though the CHIME structure reflects the most relevant aspects of mindfulness represented by existing mindfulness measures, more psychometric work is necessary to validate these measures cross-culturally and with different populations. Before this work is done, the modified 37-item FFMQ arguably has superior psychometric properties compared to other measures of trait mindfulness available today with interval level conversions available for its subscales and the total score.

Rasch analysis of mindfulness and outcome measures also indicated that a 5-point Likert scale format used by the KIMS, the FFMQ, the PSS and the TMS is the most appropriate because most of their items displayed appropriate order of thresholds without the need for any modifications. However, a 6-point Likert scale format worked well in the CHIME only, but required uniform rescoring in the MAAS and the OHQ to correct disordered thresholds. Therefore, alternative empirically supported response formats were recommended to use for these measures in the future.

Contribution of Novel Methodology for Validation of State and Trait Scales

Another major contribution is the development of the novel GT-based methodology to distinguish between state and trait components in a measure, which is demonstrated with an empirical example and presented with step-by-step instructions and relevant interpretations of the results. This GT method also provides formulas to estimate state and trait component indices and is recommended for future studies as the most appropriate psychometric method to validate state and trait measurement tools. In addition to reliable and valid measures of trait mindfulness (e.g. the modified FFMQ) and outcomes, a state measure of mindfulness – the TMS was successfully validated using the proposed GT methodology.

Applying GT to distinguish between state and trait variance components using a state mindfulness measure - the TMS (Lau et al., 2006) has also contributed to its validation as a state measure. To increase state variability, data were collected on three separate occasions: ‘after a holiday’, ‘after a mindfulness exercise’ and ‘before a stressful event’. Person-occasion interaction is a marker of individual state changes and should explain the largest amount of variance in a valid state measure. The results support the reliability and construct validity of this instrument. As expected for a valid state measure, the highest amount of variance in both TMS subscales was attributed to person-occasion interaction reflecting state changes. In contrast to valid trait measures, where persons (traits) explain the most variance and show stability over time (Arterberry et al., 2014), the TMS scores were unstable across occasions, as evidenced by G-coefficients below .30. The construct validity of the TMS as a state measure was also supported by the proposed SCI of .70 for the Curiosity and .75 for the Decentering subscale. Additionally, it suggests that the Curiosity subscale is less sensitive to state changes compared to the Decentering subscale. Accordingly, curiosity appears to be a little more dispositional compared to decentering, which is defined as a shift from identifying oneself with emotions, thoughts and perceptions to a broader awareness of external and internal experiences without identifying personally with them (Teasdale et al., 2002).

Overall, the GT findings mirror the traditional test-retest reliability results, which were consistent with those reported earlier for other state measures, such as a range of *r*-values from .34 to .46 for the State Trait Anxiety Inventory (Ramanaiah et al., 1983; Spielberger et al., 1970; Spielberger, 1999). These results provide further support for the construct validity of the TMS as a state measure as the scores followed predicted changes, namely increased mindfulness after a brief mindfulness exercise and decreased mindfulness during a stressful exam period. Furthermore, variance due to persons found in this study was substantially lower for both subscales compared to state variance. Along with more traditional and GT methods, Rasch analysis was also used and supported good psychometric properties of the TMS subscales and individual items. In particular, Rasch analysis provided evidence for the appropriate choice of item response categories by the authors, the internal construct validity of the TMS subscales, and justified the use of a parametric statistical technique (ANOVA) in applying GT.

Using GT to distinguish between state and trait appears straightforward because it is the most accurate method available to date to estimate unique contributions of various sources to the total variance in a measure (Bloch & Norman, 2012). However, GT has not been used to validate psychometric state measures, and only few studies used GT to test temporal reliability of trait measures (Arterberry et al. 2014; Berggraf, Ulvenes, Wampold, Hoffart & McCullough, 2012). To date, GT is strongly recommended but not widely used in psychometric work (Brennan, 2001; Bloch & Norman, 2012). There are two possible reasons as why GT methods are not widely used in psychometrics: One refers to complexity of available software solutions and another to laborious data collection including three or more time points which can be associated with high attrition rates. Currently, GT analysis can be conducted using GENOVA (Crick & Brennan, 1983) and EduG (Gardinet et al., 2009) software especially designed for it as well as syntax developed for IBM SPSS and SAS programs (Mushquash & O'Connor, 2006). Although, these software packages can be downloaded for free, terminology and data input requirements differ across programs. For instance, it is necessary to learn specific commands and syntax to conduct GT analysis using GENOVA. If IBM SPSS or SAS is utilised for GT analysis, data should be written exactly following authors' instructions (Mushquash & O'Connor, 2006). There is no need to learn syntax here but the analysis process is not transparent, and there are no options to verify the output. Even though, EduG has a visual interface that makes analysis process easier to understand and thus more user friendly compared to other solutions (Yelboga, 2015), it requires data to be prepared in univariate format exactly matching the G-study design (Gardinet et al., 2009). Despite some difficulties in conducting GT analysis, the benefits of this work greatly outweigh the inconvenience, and the distinction between state and trait measures can now be accurately estimated based on variance components rather than relying on a single test-retest correlation. This work contributes a new method and perhaps a new 'gold standard' for distinguishing between instruments measuring state and trait.

Limitations and Directions for Further Research

The following limitations need to be acknowledged. The four analysed mindfulness scales were considered as trait measures based on available test-retest scores supporting their temporal stability. Given the limitations of using test-retest scores to distinguish between state and trait argued earlier, the GT method proposed in this work should be applied to evaluate the ability of these and other psychometric measures to accurately capture a state or a trait and to disaggregate their influences.

Some of the scales (e.g. the KIMS, the TMS) were analysed using student samples that have a degree of homogeneity and are not fully representative of the general adult population. Other studies reported here include both students and the general population (e.g. the FFMQ, the CHIME), but they were not tested with clinical samples meaning that the findings are only applicable for the populations from which these samples were collected. Similarly, the results of the UK FIM+FAM analysis are only applicable to stroke patients, but not for other diagnostic categories or the general population. Therefore, Rasch analysis of mindfulness and outcome measures should be conducted with different samples not covered by this work, and ordinal-to-interval conversion tables need to be produced for each specific sample where DIF with the current samples would be expected. For instance, widely used mindfulness measures such as the FFMQ needs to be Rasch analysed with clinical samples such as stroke or trauma to ensure their applicability in clinical practice. Similarly the G analysis method to distinguish state and trait needs to be tested with clinical samples.

Overall, the data reflects diversity of ethnic groups by country, although no effort was undertaken to sample underrepresented groups. Where comparisons across countries were made (e.g. the PSS, the OHQ), the ethnic groups were categorised more generally compared to the country of origin (e.g. Caucasian). Future research may wish to investigate DIF effects across specific ethnicities to make interval conversions applicable to specific ethnic groups.

The SCI and TCI indices developed for evaluation of state and trait measures need to be established with different instruments and samples to determine cut-off points and benchmarks. The results of this G-study should be replicated with larger and more diverse samples to increase generalisability. It was assumed, based on the reported test-retest scores, that analysed measures such as the MAAS, the KIMS, the FFMQ, the CHIME, the PSS, and the OHQ are all measuring a trait. However, this has to be tested using more appropriate GT-based methodology introduced in Chapter Ten.

In this study, no Rasch transformation was produced after the TMS was validated as a measure of state because the sample (n=55) is insufficient to produce conversion tables for general use. Rasch analysis of the TMS with a larger and more representative sample can be undertaken by future studies. Ideally, GT method should be applied first to examine a scale and to establish its validity either as a trait or as a state measure. At this stage, necessary modifications can be implemented such as identifying items measuring more of a state or of trait and allocating them in their respective subscale and removing ambiguous items. It is now recommended for test developers to prove that it is a trait or at least not a state measure with a G study. When a scale is validated as either state or a trait measure, Rasch analysis should be conducted to improve psychometric properties up to an interval level scale.

Summary

The current thesis applied Rasch analysis and GT to evaluate and improve the reliability, validity, and applicability of eight psychometric scales used in mindfulness and outcome measurement research by analysing data of 2,551 participants. Together the studies presented in this thesis contributed to a substantial improvement in the measurement of mindfulness and related outcomes by providing ordinal-to-interval transformation algorithms that enhance measurement precision and a newly developed GT methodology to distinguish reliably between state and trait components in a measure. Minor modifications of the investigated measures were implemented to improve their psychometric properties and produce algorithms to convert ordinal responses into interval-level data suitable for parametric statistics and thus accurate comparisons with interval measures. The novel GT-based methodology was developed to permit a reliable, detailed, quantiative distinction between state and trait measures, and its application was demonstrated using the example of the TMS - state mindfulness measure. This GT method is recommended as the most powerful for validation of state and trait measurement tools. The findings presented here have far-reaching implications to improve the accuracy of scales and the distinction between state and trait in mindfulness measurement and other areas of psychological assessment.

References

Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: a review. *Med Bio Eng Comput 44*, 1031-1051.

Allal, L., & Cardinet, J. (1976). *Application of generalizability theory: Estimation of errors and adaptation of measurement designs.* Neuchâtel: Institut Romand de Recherche et de documentation pédagogiques.

Allen, M. J., & Yen, W. M. (1979). Introduction on to Measurement Theory. Monterey, CA: Brooks/Cole.

Andrews, F. M., & McKennell, A. C. (1980). Measures of self-reported well-being: their affective, cognitive, and other components. *Social Indicators Research, 8*(2), 127-155.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrica, 43*, 561–573.

Andrich, D. & Hagquist, C. (2013). Real and artificial differential item functioning. *Journal of Educational and Behavioural Statistics 37*(3), 387–416.

Andrich, D., Sheridan, B., & Luo, G. (2009). *RUMM 2030.* Perth: RUMM Laboratory.

Argyle, M. (2001). *The psychology of happiness* (2nd ed.). New York: Taylor & Francis Inc.

Argyle, M., Martin, M., & Crossland, J. (Eds.). (1989). *Happiness as a function of personality and social encounters. In J. P. Forgas, & J. M. Innes (Eds.).* North-Holland: Elsevier.

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of Generalizability Theory to Big Five Inventory. *Personality and Individual Differences, 69,* 98-103

Bach, P., & Hayes, S. C. (2002). The use of Acceptance and Commitment Therapy to prevent the rehospitalisation of psychotic patients: a randomized controlled trial. *Journal of Consulting and Clinical Psychology, 70*, 1129-1139.

Baer, R. (2003). Mindfulness training as a clinical intervention: a conceptual and empirical review. *Clinical Psychology: Science and Practice, 10*(2), 125-142.

Baer, R., Samuel, D. B., & Lykins, E. L. B. (2010). Differential item functioning on the Five Facet Mindfulness Questionnaire is minimal in demographically matched meditators and nonmeditators. *Assessment, 20*(10), 1-8.

Baer, R., Smith, G. T., & Allen, K. B. (2004). Assessment of Mindfulness by Self-Report: The Kentucky Inventory of Mindfulness Skills. *Assessment, 11*(3), 191-206.

Baer, R., Smith, G., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment, 13*(1), 27-45.

Baer, R., Smith, G., Lykins, E. L. B., Button, D., Krietemeyer, J., Sauer, S., Walsh, E., Duggan, D. & Williams, J. M. G. (2008). Construct validity of the Five Facet

Mindfulness Questionnaire in meditating and nonmeditating samples. *Assessment,15*(3), 329-342.

Balducci, C., Romeo, L., Brondino, M., Lazzarini, G., Benedetti, F., Toderi, S., Fraccaroli, F., & Pasini M. (2015). The validity of the Short UK Health and Safety Executive Stress Indicator Tool for the assessment of the psychosocial work environment in Italy. *European Journal of Psychological Assessment.* doi: 10.1027/1015-5759/a000280.

Barbosa-Leiker, C., Kostick, M., Lei, M., McPherson, S., Roper, V., Hoekstra, T., & Wright, B. (2013). Measurement invariance of the Perceived Stress Scale and latent mean differences across gender and time. *Stress and Health 29*(3), 253-260.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist 54*, 462–479.

Barlow, J. S. (1985). Methods of analysis of nonstationary EEGs, with emphasis on segmentation techniques : a comparative review. *Journal of Clinical Neurophysiology, 2,* 267-304.

Barker, H. R., Wadsworth, A. P., & Wilson, W. (1976). Factor structure of the State-Trait Anxiety Inventory in a nonstressful situation. *Journal of Clinical Psychology, 32,* 595-598.

Barnhofer, T., Chittka, T., Nightingale, H., Visser, C., & Crane, C. (2010). State Effects of Two Forms of Meditation on Prefrontal EEG Asymmetry in Previously Depressed Individuals. *Mindfulness, 1,* 21-27.

Barry, R. J., Clark, A. R., Johnstone, S. J., Magee, C. A., & Rushby, J. A. (2007). EEG differences between eyes-closed and eyes-open conditions. *Clinical Neurophysiology, 118*, 2765-2773.

Baum, C., Kuyken, W., Bohus, M., Heidenreich, T., Michalak, J., & Steil, R. (2010). The Psychometric Properties of the Kentucky Inventory of the Mindfulness Skills in Clinical Populations. *Assessment, 17*(2), 220-229.

Beck, A. T. (1991). Cognitive therapy: a 30-year retrospective. *American Psychologist, 46*(4), 368-375.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory II.* San Antonio: Psychological Corporation.

Becker, D. E., & Shapiro, D. H. (1981). Physiological responses to clicks during Zen, Yoga, and TM meditation. *Psychophysiology, 18*, 694-699.

Bennet, K., & Dorje, D. (2015). The Impact of a Mindfulness-Based Stress Reduction Course (MBSR) on Well-Being and Academic Attainment of Sixth-form Students. *Mindfulness*, doi: 10.1007/s12671-015-0430-7

Berggraf, L., Ulvenes, P. G., Wampold, B. E., Hoffart, A., & McCullough, L. (2012). Properties of the Achievement of Therapeutic Objectives Scale (ATOS): A Generalizability Theory Study. *Psychotherapy Research, 22*(3), 327-347.

Bergomi, C., Tschacher, W., & Kupper, Z. (2013). The assessment of mindfulness with self-report measures: Existing scales and open issues. *Mindfulness*, *4*(3), 191-202.

Bergomi, C., Tschacher, W., & Kupper, Z. (2014). Konstruktion und erste Validierung eines Fragebogens zur umfassenden Erfassung von Achtsamkeit [Construction and Initial Validation of a Questionnaire for the Comprehensive Investigation of Mindfulness]. *Diagnostica*, *60*, 111–125.

Berkovich-Ohana, A., Glicksohn, J., & Goldstein, A. (2011). Mindfulness-induced changes in gamma band activity – Implications for the default mode network, self-reference and attention. *Clinical Neurophysiology*. doi:10.1016/j.clinph.2011.07.048

Bishop, S. R., Lau, M. A., Shapiro, S., Carlson, L. E., Anderson, N. D., Carmody, J., & Devins, G. (2006). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice, 11*, 230-241.

Bland, J. M., & Altman, D. G.(1995) Multiple significance tests: the Bonferroni method. *British Medical Journal, 310,* 170.

Bloch, G. J., Neeleman, L., & Aleamoni, L. M. (2004). The Salient Stressor Impact Questionnaire (SSIQ): A measurement of the intensity and chronicity of stress. *Assessment, 11*(4), 342-360.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher, 34*, 960-992.

Bond, F. W., & Bunce, D. (2000). Mediators of change in emotion-focused and problemfocused worksite stress management interventions. *Journal of Occupational Health Psychology, 5,* 156-163.

Bond, T., G., & Fox, C. M. (2007). *Applying the Rasch Model: fundamental measurement in the human sciences* (2nd Ed.). New York: Routledge.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Brandburn, N. M. (1969). *The structure of psychological well-being.* Chicago: Aldine.

Branstetter, A. D., Wilson, K. G., Hildebrandt, M., & Mutch, D. (2004. *Improving psychological adjustment among cancer patients: ACT and CBT.* presented at the meeting of the Paper presented at the Association for Advancement of Behavior Therapy, New Orleans.

Brennan, R. L. (1977). *Generalizability analysis: Principles and procedures*. Iowa City, Iowa: The American College Testing Program.

Brennan, R. L. (1992). *Elements of Generalizability Theory* (2nd edition). Iowa City: ACT Publications.

Brennan, R. (2001). *Generalizability theory.* New York: Springer Verlag.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika, 42*(4), 631-634.

Brown, K. W., & Kasser, T. (2005). Are psychological and ecological well-being compatible? The role of values, mindfulness, and lifestyle. *Social Indicators Research*, *74*(2), 349-368.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*(4), 822-884.

Brown, K. W., Ryan, R. M., & Creswell, J. D. (2007). Mindfulness: Theoretical foundations and evidence for its salutary effects. *Psychological Inquiry, 18*(4), 2011-2237.

Buchheld, N., Grossman, P., & Walach, H. (2001). Measuring mindfulness in insight meditation (vipassana) and meditation-based psychotherapy: The development of the Freiburg Mindfulness Inventory (FMI). *Journal for Meditation and Meditation Research, 1*, 11-34.

Buss, A. H. (1989). Personality as traits. *American Psychologist, 44*, 1378-1388.

Cahn, B. R., & Polich, J. (2006). Meditation states and traits : EEG, ERP, and neuroimaging studies. *Psychological Bulletin, 132*, 180-211.

Carr, A. (2004). *Positive psychology*. New York: Brunner-Routledge.

Cardaciotto, L., Herbert, J. D., Forman, E. M., Moitra, E., & Farrow, V. (2008). The assessment of present-moment awareness and acceptance: The Philadelphia Mindfulness Scale. *Assessment, 15*(2), 204-223.

Carlson, L. E., & Brown, K. W. (2005). Validation of the Mindful Attention Awareness Scale in a cancer population. *Journal of Psychosomatic Research, 58*(1), 29-33.

Cash, M., & Whittingham, K. (2010). What facets of mindfulness contribute to psychological well-being and depressive, anxious, and stress-related symptomatology? *Mindfulness, 1*(3), 177–182.

Cayoun, B. (2011). *Mindfulness-integrated CBT: principles and practice.* Oxford: John Wiley & Sons, Ltd.

Chadwick, P., Hember, M., Symes, J., Peters, E., Kuipers, E., & Dagnan, D. (2008). Responding mindfully to unpleasant thoughts and images: Reliability and validity of the Southampton Mindfulness questionnaire (SMQ). *The British Journal of Clinical Psychology, 47*(4), 451-455.

Chambers, R., Gullone, E., & Allen, N. B. (2009). Mindfulness emotion regulation: An integrative review. Clinical Psychology Review, 29(6), 560-572.

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of State and Trait: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology, 54*(4), 541-557.

Chavez-Korell, S., & Torres, L. (2014) Perceived stress and depressive symptoms among Latino adults: The moderating role of ethnic identity cluster patterns. *The Counseling Psychologist*, 42(2), 230-254.

Chiesa, A., & Serretti, A. (2010). A systematic review of neurobiological and clinical features of mindfulness meditations. *Psychological Medicine, 40*, 1239-1252.

Christensen, K. B., Makransky, G., & Horton, M. (2016). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, doi:10.1177/0146621616677520

Christopher, M. S., Charoensuk, S., Gilbert, B. D., Neary, T. J., & Pearce, K. L. (2009). Mindfulness in Thailand and the United States: A case of apples versus oranges? *Journal of Clinical Psychology, 65*(6), 590-612.

Christopher, M. S., & Gilbert, B. D. (2010). Incremental validity of components of mindfulness in the prediction of satisfaction with life and depression. *Current Psychology, 29*(1), 10-23.

Christopher, M. S., Neuser, N. J., Michael, P. G., & Baitmangalkar, A. (2012). Exploring the psychometric properties of the Five Facet Mindfulness Questionnaire. *Mindfulness, 3*(2), 124–131.

Coelho, H. F., Canter, P. H., & Ernst, E. (2007). Mindfulness-based cognitive therapy: evaluating current evidence and informing future research. *Journal of Consulting and Clinical Psychology, 75*(6), 1000-1005.

Cohen, S. (2013). *Frequently asked questions.* Retrieved from http://www.psy.cmu.edu/~scohen/

Cohen, S., Frank, E., Doyle, W. J., Skoner, D. P., Rabin, B. S., & Gwaltney, J. M. (1998). Types of stressors that increase susceptibility to the common cold in healthy adults. *Health Psychology, 17*(3), 214-223.

Cohen, R. J., & Swerdlik, M. E. (2010) *Psychological testing and assessment: An introduction to tests and measurement*. New-York: McGraw-Hill.

Cohen, S., & Williamson, G. M. (Eds.). (1988). *Perceived stress in a probability sample of the United States.* In S. Spacapan & S. Oskamp (Eds.) *The social psychology of health: Claremont Symposium on Applied Social Psychology.* Newbury Park, CA: Sage.

Cohen-Katz, J., Wiley, S. D., Capuano, T., Baker, D. M., & Shapiro, S. (2005). The Effects of Mindfulness-based Stress Reduction on Nurse Stress and Burnout, Part II: A Quantitative and Qualitative Study. *Holistic Nursing Practice, 19*(1), 26-35.

Cole, S. R. (1999). Assessment of differential item functioning in the Perceived Stress Scale-10. *Journal of Epidemiology & Community Health 53*(5), 319–320.

Compton, W. C. (2005). *Introduction to positive psychology.* Belmont, USA: Wadsworth.

Conte, H. R., Plutchik, R., Jung, B. B., Picard, S., Karasu, T. B., & Lotterman, A. (1990). Psychological mindedness as a predictor of psychotherapy outcome: A preliminary report. *Comprehensive Psychiatry, 31,* 426– 431.

Cordon, S. L., & Finney, S. J. (2008). Measurement invariance of the Mindful Attention Awareness Scale across adult attachment style. *Measurement & Evaluation in Counseling & Development, 40*(4), 228-245.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system.* (American College Testing Technical Bulletin 43). Iowa City: ACT, Inc; 1983.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, XVII*(2), 137-163.

Csikszentmihalyi, M. (1992). *The psychology of happiness*. Kent: Mackays of Chatham plc.

Dahl, J., Wilson, K. G., & Nilsson, A. (2004). Acceptance and commitment therapy and the treatment of persons at risk for long-term disability resulting from stress and pain symptoms: A preliminary randomized trial. *Behavior Therapy, 35,* 785-802.

Dalai Lama, Baron, R., & Gaffiney, P. (2004). *Dzogchen: Heart Essence of the great Perfection.* (2nd ed.). New York, USA: Snow Lion Publications.

Davids, R., & Stede, W. (Eds.). (1921/2001). *Pali-english dictionary*. New Delhi, India: Munshiram Manoharlal Publishers Pvt, Ltd.

Davidson, R. J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., & Santorelli, S. F. (2003). Alterations in brain and immune function produced by mindfulness meditation. *Psychosomatic Medicine, 65*(564-570).

Davis, K. M., Lau, M. A., & Cairns, D. R. (2009). Development and preliminary validation of a trait version of the Toronto Mindfulness Scale. *Journal of Cognitive Psychotherapy, 23*(3), 185–197.

DeVellis, R. F. (2006) Classical test theory. *Medical Care, 44*(Suppl), 50-59.

Didonna, F. (Eds.).(2009). *Clinical handbook of mindfulness*. doi:10.1007/978-0-387-09593-6

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*(3), 542-575.

Diener, E. (2006). Guidelines for national indicators of subjective well-being and ill-being. *Journal of Happiness Studies, 7*(4), 397-404.

Diener, E., Lucas, R. E., & Oishi, S. (Eds.). (2005). *Subjective well-being: the science of happiness and life satisfaction.* New York, USA: Oxford University Press.

Diener, E., Sapyta, J. J., & Suh, E. M. (1999). Subjective well-being is essential to well-being. *Psychological Inquiry, 9*(1), 33-37.

Dimidjan, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Gallop, R., McGlinchey, J. B., ... Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy and antidepressant medication in the acute treatment of adults with major depression. *Journal of Counsulting and Clinical Psychology, 74*(4), 658-670.

Dimidjian, S., & Linehan, M. M. (2003). Defining an agenda for future research on the clinical application of mindfulness practice. . *Clinical Psychology: Science and Practice, 10,* 166-171.

Dunn, B. R., Hartigan, J. A., & Mikulas, W. L. (1999). Concentration and Mindfulness Meditations: Unique Forms of Consciousness? *Applied Psychophysiology and Biofeedback, 24*(3), 147-165.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Reviws, 5,* 155-74.

Eid, M., & Larsen, R. J. (Eds.). (2008). *The science of subjective well-being.* New York, USA: The Guilford Press.

Ellis, A. (2002). *Overcoming resistance: a rational emotive behavior therapy integrated approach* (2nd ed.). New York: Springer Publishing Company, Inc.

Epstein, S. (1984). Trait theory as personality theory: Can a part be as great as the whole? *Psychological Inquiry, 5,* 120-122.

Embretson, S. E. (1996) The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Erwin, R. J., Mawhinney-Hee, M., Gur, R. C., & Gur, R. E. (1989). Effects of task and gender on EEG indices of hemispheric activation: similarities to previous rCBF findings. *Neuropsychiatry, Neuropsychology, & Behavioural Neurology, 2*(4), 248-260.

Feldman, G., Hayes, A., Kumar, S., Greeson, J., & Laurenceau, J. (2007). Mindfulness and emotion regulation: The development and initial validation of the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). *Journal of Psychopathology & Behavioral Assessment, 29*(3), 177-190.

Fisak, B., & von Lehe, A. C. (2012). The relation between the Five Facets of Mindfulness and Worry in a Non-clinical Sample. *Mindfulness, 3*(1), 15–21

Fisher, R. A. (1925, 2006). *Intraclass correlation and the analysis of variance. In: Statistical methods for research workers.* New Delhi: Cosmo Publications for Genesis Pub.

Fisher, W. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

Fordyce, M. W. (1988). A review of research on the happiness measures: a sixty second index of happiness and mental health. *Social Indicators Research, 20*(4), 355-381.

Frewen, P., Evans, E., Maraj, N., Dozois, D., & Partridge, K. (2008). Letting go: Mindfulness and negative automatic thinking. *Cognitive Therapy & Research, 32*(6), 758-774.

Gaudiano, B. A., & Herbert, J. D. (2006). Acute treatment of inpatients with psychotic symptoms using acceptance and commitment therapy: pilot results. *Behavioral Research Therapy, 44,* 415-437.

Gardinet, J., Johnson, S., & Pini, G. (2009). *Applying Generalizability Theory Using EduG.* New York: Routledge.

Gazzaniga, M. S., Ivry, R. V., Mangun, G. R., & Steven, M. S. (2009). *Cognitive neuroscience: the biology of the mind* (3rd ed.). New York: W. W. Norton & Company, Inc.

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., & Servera, M. (2015). Analyzing person, situation and person - situation interaction effects: Latent State-Trait Models for the Combination of Random and Fixed Situations. *Psychological Methods, 20*(2), 165-192.

Germer, C. K., Siegel, R. D., & Fulton, P. R. (Eds.). (2005). *Mindfulness and psychotherapy*. New York, USA: The Guilford Press.

Gitchel W. D., Roessler, R. T., & Turner, R. C. (2011). Gender effect according to item directionality on the Perceived Stress Scale for adults with multiple sclerosis. *Rehabilitation Counseling Bulletin, 55*(1), 20-28.

Goh, H. E., Marais, I., & Ireland, M. J. (2015). A Rasch model analysis of the Mindful Attention Awareness Scale. *Assesment*. doi:DOI: 10.1177/1073191115607043

Goldin, P. R., & Gross, J. J. (2010). Effects of Mindfulness-Based Stress Reduction (MBSR) on emotion regulation in social anxiety disorder. *Emotion, 10*(1), 89-91.

Govern, J. M., & Marsch, L. A. (2001). Development and validation of the situational selfawareness scale. *Consciousness and Cognition, 10*, 366–378.

Gunaratana, B. (2002). *Mindfulness in plain English.* Somerville, USA: Wisdom Publications.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology 33*, 205–233.

Haigh, E. A., Moore, M. T., Kashdan, T. B., & Fresco, D. M. (2011). Examination of the factor structure and concurrent validity of the Langer Mindfulness/Mindlessness Scale. *Assessment, 18*(1), 11-26.

Hall, K. M., Hamilton, B. B., Gordon, W. A., & Zasler, N. D. (1993) Characteristics and comparisons of functional assessment indices: Disability Rating Scale, Functional Independence Measure, and Functional Assessment Measure. *Journal of Head Trauma Rehabilitation, 8,* 60–74.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait-state model. *Journal of Research in Personality, 41*, 295-315.

Hamilton, B. B., Granger, C. V., Sherwin, F. S., Zielezny, M., & Tashman, J. S. (1987) *A uniform national data system for medical rehabilitation.* In Fuhrer, J. M., (Eds.). Rehabilitation outcomes: analysis and measurement, p. 137-47. Baltimore: Brookes.

Hawley, C. A., Taylor, R., Hellawell, D. J., & Pentland, B. (1999). Use of the functional assessment measure (FIM+FAM) in head injury rehabilitation: a psychometric analysis. *Journal of Neurological Neurosurgering Psychiatry, 67*, 749-754.

Hayes, S. C., Follette, V. M., & Linehan, M. M. (Eds.). (2004). *Mindfulness and acceptance: expanding cognitive-behavioral tradition*. New York: The Guilford Press.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes, and outcomes. *Behaviour Research and Therapy, 44,* 1-25.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptancenand commitment therapy: an experiential approach to behavior change.* New York: Guilford Press.

Helton, W. S., & Näswall, K. (2015) Short Stress State Questionnaire: Factor Structure and State Change Assessment. *European Journal of Psychological Assessment, 31*(1), 20–30.

Hill, E. M., Billington, R., & Krägeloh, C. (2014). Noise and diminished health: testing moderators and meditators of the relationship. *Noise & Health, 16*(68), 47-56.

Hillhouse, J. E., Kiecolt-Glaser, J. K., & Glaser, R. (Eds.). (1991). *Stress associated modulation of the immune response in humans.* In N. Plotnikoff, A. Murgo, R. Faith,&J.Wybran (Eds.), Stress and immunity. Boca Raton, FL: CRC Press.

Hills, P., & Argyle, M. (1998). Positive moods derived from leisure and their relationship to happiness and personality. *Personality and Individual Differences, 25* (3), 523-535.

Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: a compact scale for the measurement of psychological well-being. *Personality and Individual Differences, 33*, 1073-1082.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment,13*(12), 1-200.

Hobart, J. C., Cano, S. J., & Thompson, A. J. (2010). Effect sizes can be misleading: is it time to change the way we measure change? *Journal of Neurological & Neurosurgical Psychiatry, 81*,1044-8.

Hofmann, S. G., Sawyer, A. T., Witt, A., & Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of Consulting and Clinical Psychology,, 78*(2), 169-183.

Holland, P., & Wainer, H. (1993). *Differential item functioning*. NJ, Hollsdale: Lawrence Erlbaum.

Hölzel, B. K., Carmodyc, J., Vangela, M., Congletona, C., Yerramsettia, S. M., Garda, T., & Lazara, S. W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. *Psychiatry Research: Neuroimaging, 191*, 36-43.

Howells, F. M., Ives-Deliperi, V. L., Horn, N. R., & Stein, D. J. (2012). Mindfulness based cognitive therapy improves frontal control in bipolar disorder: a pilot study. *BMC Psychiatry, 12*(1), 15.

Ingram, P. B., Clarke, E. & Lichtenberg, J. W. (2016). Confirmatory factor analysis of the Perceived Stress Scale-4 in a community sample. *Stress and Health 32*(2), 173–176.

Inchausti, F., Prieto, G., & Delgado, A. R. (2013). Rasch analysis of the Spanish version of the Mindful Attention Awareness Scale (MAAS) in a clinical sample. *Revista de Psiquiatria y Salud Mental, 7*(1), 32-41.

Ivanovski, B., & Malhi, G. S. (2007). The psychological and neurophysiological concomitants of mindfulness forms of meditation. *Acta Neuropsychiatrica, 19,* 76-91.

Ives-Deliperi, V. L., Solms, M., & Meintjes, E. M. (2010). The neural substrates of mindfulness: An fMRI investigation. *Social Neuroscience, 6*(3), 231-242.

Jacobson, N. S., Martell, C. R., & Dimidjian, S. (2001). Behavioral activation treatment for depression: returning to contextual roots. *Clinical Psychology: Science and Practice, 8*(3), 255-270.

Johnson, C., Burke, C., Brinkman, S., & Wade, T. (2016). Development and validation of a multifactor scale in youth: the Comprehensive Inventory of Mindfulness Experiences – Adolescents (CHIME-A). *Psychological Assessment*. doi:10.1037/pas0000342

Josefsson, T., Lindwall, M., & Broberg, A. G. (2014). The Effects of a Short-term Mindfulness Based Intervention on Self-reported Mindfulness, Decentering, Executive Attention, Psychological Health, and Coping Style: Examining Unique Mindfulness Effects and Mediators. *Mindfulness, 5,* 18-35.

Joseph, S., & Lewis, C. A. (1998). The depression-happiness scale: reliability and validity of a bipolar self-reported scale. *Journal of Clinical Psychology, 54*(4), 537-544.

Joshanloo, M. (2015). Revisiting the empirical distinction between hedonic and eudaimonic aspects of well-being using exploratory structural equation modeling. *Journal of Happiness Studies,* 1-14. doi:10.1007/s10902-015-9683-z

Kabat-Zinn, J. (1982). An outpatient program in behavioural medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. *General Hospital Psychiatry, 4,* 33-47.

Kabat-Zinn, J. (1990). *Full catastrophe living: using the wisdom of your body and mind to face stress, pain and illness.* New-York, USA: Delacorte.

Kabat-Zinn, J. (1994). *Wherever you go, there you are: mindfulness meditation in everyday life.* New York, USA: Hyperion.

Kabat-Zinn, J. (2003). Mindfulness-based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice, 10,* 144-156.

Kabat-Zinn, J. (Ed.). (2000). *Indra's net at work: the mindstreaming of Dharma practice in society.* Nork Beach, USA: Weiser.

Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives, 20*(1), 3-24.

Kasamatsu, A., & Hirai, T. (1966). An electroencephalographic study on the zen meditation (Zazen). *Folia Psychiatrica et Neurologica Japonica, 20,* 315-336.

Kashdan, T. B. (2004). The assessment of subjective well-being (issues raised by the Oxford Happiness Questionnaire). *Personality and Individual Differences, 36* (5), 1225–1232.

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). The functional independence measure: a new tool for rehabilitation. *Advances in Clinical Rehabilitation, 1*, 6-18.

Keng, S. L., Smoski, M. J., & Robin, C. J. (2011). Effects of mindfulness on psychological health: A review of empirical studies. *Clinical Psychology Review, 31*(6), 1041-1056.

Kenny, D. A., & Zautra, A. (1995). The Trait-State-Error Model for Multivariate Data. *Journal of Councelling and Clinical Psychology, 63*(1), 52-59.

Keune, P. M., Bostanov, V., Hautzinger, M., & Kotchubey, B. (2011). Mindfulness-based cognitive therapy (MBCT), cognitive style, and the temporal dynamics of frontal EEG alpha asymmetry in recurrently depressed patients. *Biological Psychology, 88*(2-3), 243-252.

Khan, A, Chien, C. W., & Brauer, S. G. (2013). Rasch-based scoring offered more precision in differentiating patient groups in measuring upper limb function. *Journal of Clinical Epidemiology, 66*, 681-7.

Kim-Prieto, C., Diener, E., Tamir, M., Scollon, C. N., & Diener, M. (2005). Integrating the diverse definitions of happiness: a time-sequential framework of subjective well-being. *Journal of Happiness Studies, 6* (3), 261-300.

Korte, S. M., Koolhaas, J. M., Wingfield, J. C., & McEwen, B. S. (2005). The Darwinian concept of stress: benefits of allostasis and costs of allostatic load and the trade-offs in health and disease. *Neuroscience and Biobehavioral Reviews, 29*(1), 3-38.

Krägeloh, C. U., Kersten, P., Billington, D. R., Hsu, P. H.-C., Shepherd, D., Landon, J., & Feng, X. J. (2013). Validation of the WHOQOL-BREF quality of life questionnaire for general use in New Zealand: Confirmatory factor analysis and Rasch analysis. *Quality of Life Research, 22*(6), 1451-1457.

Kristoffersen, I. (2010). The metrics of subjective wellbeing: cardinality, neutrality and additivity. *The Economic Record, 86*(272), 98-123.

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., et al. (2006). The Toronto Mindfulness Scale: Development and validation. *Journal of Clinical Psychology, 62*(12), 1445–1467.

Law, J., Fielding, B., Jackson, D., & Turner-Stokes, L. (2009) The UK FIM+FAM Extended Activities of Daily Living module: evaluation of scoring accuracy and reliability. *Disability and Rehabilitation,31*, 825-30.

Lawrence, M., Booth, J., Mercer, S., & Crawford, E. (2013). A systematic review of the benefits of mindfulness-based interventions following transient ischemic attack and stroke. *International Journal of Stroke*, 8, 465-474.

Leary, M. R. (2004). *The curse of the self: Self-awareness, egotism, and the quality of human life.* New York: Oxford University Press.

Leary, M. R., Adams, C. E., & Tate, E. B. (2006). Hypo-egoic selfregulation: Exercising self-control by diminishing the influence of the self. *Journal of Personality, 74*, 1803–1831.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23,* 155-184.

Linacre, J. M. (1994) Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2011) *Winsteps Rasch measurement computer program, version 3.73.0.* Chicago: Winsteps.com.

Linehan, M. M. (1993a). *Cognitive-behavioural treatment of borderline personality disorder.* New York: Guilford Press.

Linehan, M. M. (1993b). *Skills training manual for treating bordeline personality disorder.* New York, USA: Guilford Press.

Linn, R. T., Blair, R. S., Granger, C.V., Harper, D. W., O'Hara, P. A., & Maciura, E. (1999). Does the Functional Assessment Measure (FAM) extend the Functional Independence Measure (FIMTM) instrument? A Rasch analysis of stroke inpatients. *Journal of Outcome Measurement, 3,* 339-359.

Little, T. D., Cunningham, W. A., Shahar, G., et al. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173.

Lord, F. M., & Novick, M. R. (1968) *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley.

Lundgren Nilsson, Å., Grimby, G., Ring, H., Tesio, L., Lawton, G., Slade, A., Penta, M.,Tripolski, M., Biering-Sørensen, F., Carter, J., Marincek, C., Phillips, S., Simone, A., & Tennant, A. (2005) Cross-cultural validity of functional independence measure items in stroke: a study using Rasch analysis. *Joiurnal of Rehabilitation Medicine, 37,* 23-31.

Lundgren Nilsson, Å., Jonsdottir, I. H., Ahlborg, G., & Tennant, A. (2013). Construct validity of the Psychological General Well Being Index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: A Rasch analysis. *Health and Quality of Life Outcomes, 11,* 2.

Lundgren Nilsson, Å., & Tennant, A. (2011) Past and present issues in Rasch analysis: The Functional Independence Measure (FIM™) revisited. *Journal of Rehabilitation Medicine, 43,* 884–891.

Lush, E., Salmon, P., Floyd, A., Studts, J. L., Weissbecker, I., & Sephton, S. E. (2009). Mindfulness Meditation for Symptom Reduction in Fibromyalgia: Psychophysiological Correlates. *Journal of Clinical Psychology inl Medical Settings, 16,* 200-207.

Lyvers, M., Makin, C., Toms, E., Thorberg, F. A., & Samios, C. (2014). Trait Mindfulness in Relation to Emotional Self-Regulation and Executive Function. Mindfulness, *5,* 619-625.

Mace, C. (2008). *Mindfulness and menthal health: therapy, theory and science.* New York, USA: Routladge.

MacKillop, J., & Anderson, E. J. (2007). Further psychometric validation of the Mindful Attention Awareness Scale (MAAS). *Journal of Psychopathology & Behavioral Assessment, 29*(4), 289-293.

Mallison, T. (2011). Rasch anlysis of repeated measures. *Rasch Measurement, 25*(1), 1317.

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105–124.

Marlatt, G. A., & Kristeller, J. L. (Eds.). (1999). *Mindfulness and meditation.* Washington DC, USA: American Psychological Association.

Masters G., A. (1982). Rasch model for partial credit scoring. *Psychometrica, 47*, 149–174.

Mayers, D. G. (1992). The secrets of happiness. *Psychology Today, 25*(4), 38-45

McEwen, B. S., & Stellar, E. (1993). Stress and the individual:mechanisms leading to disease. *Archives of Internal Medicine, 153*(18), 2093-2101.

Medvedev, O. N., Krägeloh, C. U., Hill, E. M., Billington, R., Siegert, R. J., Webster, C. S., Booth, R. J., & Henning, M. A. (2017b). Rasch analysis of the Perceived Stress Scale: Transformation from an ordinal to a linear measure. *Journal of Health Psychology,* doi:10.1177/1359105316689603

Medvedev, O. N., Krägeloh, C. U., Narayanan, A., & Siegert, R. J. (2017c). Measuring mindfulness: Applying Generalizability Theory to distinguish between state and trait. *Mindfulness,* doi: 10.1007/s12671-017-0679-0

Medvedev, O. N., Shepherd, D., & Hautus, M. J. (2015). The restorative potential of soundscapes: A physiological investigation. *Applied Acoustics, 96*, 20-26.

Medvedev, O. N., Siegert, R. J., Feng, X. J., Billington, D. R., Jang, J. Y., & Krägeloh, C. U. (2016a). Measuring trait mindfulness: how to improve the precision of the Mindful Attention Awareness Scale using a Rasch model. *Mindfulness,* 7(2), 384-395.

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (2016b). Rasch Analysis of the Kentucky Inventory of Mindfulness Skills. *Mindfulness*, 7(2), 466-478.

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (20117a). Improving the Precision of the Five Facet Mindfulness Questionnaire Using a Rasch Approach. *Mindfulness,* doi 10.1007/s12671-016-0676-8

Medvedev, O. N., Siegert, R. J., Mohamed, A. D., Shepherd, D., Landhuis, E., & Krägeloh, C. U. (2016c) The Oxford Happiness Questionnaire: Transformation from an ordinal to an interval measure using Rasch analysis. *Journal of Happiness Studies,* doi:10.1007/s10902-016-9784-3

Medvedev, O. N., Siegert, R. J., & Krägeloh, C. U. (May 2016c). *Measuring Trait Mindfulness –Rasch Approach.* Paper presented at the Second International Conference on Mindfulness, Rome, Italy.

Merbitz, C., Morris, J. & Grip, J. C. (1989). Ordinal scales and foundations misinference. *Archives of Physical Medicine and Rehabilitation, 70*(4), 308–312.

MiCBT Institute. (2011). *Mindfulness-integrated Cognitive Behaviour Therapy (MiCBT).* Retrieved 12.09, 2011, from http://www.mindfulness.net.au/conceptual-outline/

Mitchell, A. M., Crane, P. A., & Kim, Y. (2008). Perceived stress in survivors of suicide: Psychometric properties of the Perceived Stress Scale. *Research in Nursing & Health, 31*(6), 576–585.

Murata, T., Koshino, Y., & Omori, M. (1994). Quantitative EEG study on Zen meditation (zaZen). *Japanese Journal of Psychiatry and Neurology, 48,* 881-890.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

Nayar, M., Vanderstay, R., Siegert, R. J., & Turner-Stokes, L. (2016). The UK Functional Assessment Measure (UK FIM+FAM): Psychometric evaluation in patients undergoing specialist rehabilitation following a stroke from the Nation UK Clinical Dataset. *PLOS One,11*(1): e0147288 doi:10.1371/journal.pone. 0147288

Nesse, R. M. (1990). Evolutionary explanation of emotions. *Human Nature, 1* (3), 261-289.

Nickerson, R. S. (Ed.). (1978). *Attention and Performance VIII.* New Jersey: Lawrence Erbaum Associates, Inc., Publishers.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. 3rd edition. New-York: McGraw-Hill.

Norquist, J. M., Fitzpatrick, R., Dawson, J. &Jenkinson, C. (2004). Comparing Alternative Rasch-Based Methods vs Raw Scores in Measuring Change in Health. *Medical Care, 42*(1), 125-136

Nyanaponika, T. (1973). *The heart of Buddhist meditation.* New York: Weiser Books.

OECD. (2015). The Organisation for Economic Co-operation and Development (OECD) Better life Index. http://www.oecdbetterlifeindex.org/ - /55555555555

Olendzki, A. (Ed.). (2005). *The roots of mindfulness.* New York: Guilford.

Pagnoni, G., & Cekic, M. (2007). Age effects on gray matter volume and attentional performance in Zen meditation. *Neurobiology of Aging, 28,* 1623–1627.

Pagnoni, G., Cekic, M., & Guo, Y. (2008). ‘Thinking about notthinking’ : neural correlates of conceptual processing during Zen meditation. *PLoS One, 3*(9), e3083.

Park, T., Reilly-Spong, M., & Gross, C. R. (2013). Mindfulness: a systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research, 22,* 2639-2659.

Pearson, M. R., Brown, D. B., Bravo, A. J., & Witkiewitz, K. (2015). Staying in the moment and finding purpose: The associations of trait mindfulness, decentering, and purpose in life with depressive symptoms, anxiety symptoms, and alcohol-related problems. *Mindfulness,* doi: 10.1007/s12671-014-0300-8

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*, 335-355.

Ramanaiah, N. V., Franzen, M., & Schill, T. (1983). A psychometric study of the State-Trait Anxiety Inventory. *Personality Assessment, 47*, 531-535.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Symposium conducted at the meeting of the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California: University of California Press.

Robbins, M., Francis, L. J., & Edwards, B. (2010). Happiness as stable extraversion: Internal consistency reliability and construct validity of the Oxsford Happiness Questionnaire among undergraduate university students. *Current Psychology, 29* (2), 89-94.

Roberti, J. W., Harrington, L. N., & Storch, E. A. (2006). Further psychometric support for the 10-item version of the Perceived Stress Scale. *Journal of College Counseling 9*(2):135–147.

Rosenbaum, P. R. (1989). Criterion-related construct validity. *Psychometrica, 54*(4), 625-633.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin 94*, 18–38.

Sauer, S., Ziegler, M., Danay, E., Ives, J., & Kohls, N. (2013). Specific objectivity of mindfulness - a Rasch analysis of the Freiburg Mindfulness Inventory. *Mindfulness, 4*(1), 45-54.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2002). *Mindfulness based cognitive therapy for depression.* New York, USA: Guilford Press.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2013). *Mindfulness-Based Cognitive Therapy for depression* (2nd ed.). New York: Guilford Press.

Shavelson, R. G., Webb, N. M., & Rowley, G. L. (1989). Generalizability Theory. *American Psychologist, 44,* 599-612.

Siegert, J. R., Rowland, V., & Theadom A. A. (2016, May) *A critical review of mindfulness interventions for neurological conditions.* Paper presented on the Second International Conference on Mindfulness in Rome, Italy.

Siegert, R. J., Tennant, A., & Turner-Stokes, L. (2010). Rasch analysis of the Beck Depression Inventory-II in a neurological rehabilitation sample. *Disability and Rehabilitation, 32*(1), 8-17.

Siegling, A. B., & Petrides, K. V. (2014). Measures of trait mindfulness: Convergent validity, shared dimensionality, and linkages to the five-factor model. *Frontiers in Psychology, 5,* 1-8.

Simons, D. J., & Chabris, C. F. (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception, 28,* 1059-1074.

Skevington, S. M., MacArthur, P., & Somerset, M. (1997). Developing items for the WHOQOL: an investigation of contemporary beliefs about quality of life related to health in Britain. *British Journal of Health Psychology, 2* (1), 55-72.

Smith, E. V. (2002). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.

Specialised Neurorehabilitation Service Standards (2015) *Specialist neuro-rehabilitation services: providing for patients with complex rehabilitation needs.* London: British Society of Rehabilitation Medicine. Updated 2015. 2010. Retrieved from http://www.bsrm.org.uk/downloads/specialised-neurorehabilitation-service-standards--7-30-4-2015-forweb.pdf

Spielberger, C. D. (1999). *Manual for the State-Trait Anger Expression Inventory-2.* Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Test manual for the State Trait Anxiety Inventory.* Palo Alto, California.: Consulting Psychologists Press.

Sharp, L. K., Kimmel, L. G., Kee, R., Saltoun, C., & Chang, C. H. (2007). Assessing the Perceived Stress Scale for African American adults with asthma and low literacy. *Journal of Asthma, 44*(4), 311-316.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability Theory. *American Psychologist, 44*(6), 922-932. doi: 10.1037/0003-066x.44.6.922

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13,* 251-271.

Spruk, R., & Kešeljević, A. (2015). Institutional origins of subjective well-being: estimating the effects of economic freedom on national happiness. *Journal of Happiness Studies,* 1-54. doi:DOI 10.1007/s10902-015-9616-x

Stewart, M. E., Watson, R., Clark, A., Ebmeier, K. P., & Deary, I. J. (2010). A hierachy of happiness? Mokken scaling analysis of the Oxford Happiness Inventory. *Personality and Individual Differences, 48* (7), 845-848.

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich S. (2009). Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. *Health and Quality of Life Outcomes, 7*(15), 1-8.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and Traits in Psychological Assessment. *European Journal of Psychological Assessment, 8*(2), 78-98.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales.Apractical guide to their development and use.* 4th edition.Oxford: Oxford University Press.

Stucki, G., Daltroy, L., Katz J.N., Johannesson, M., Liang, M.H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology, 49*(7), 711-717.

Statistics New Zealand. (2013). *Census ethnic group profiles*. Wellington: New Zealand Government.

Takahashia, T., Murataa, T., Hamadab, T., Masao Omoria, Kosakaa, H., Kikuchic, M., ... Wadaa, Y. (2004). Changes in EEG and autonomic nervous activity during meditation and their association with personality traits. *International Journal of Psychophysiology, 55*, 199-207.

Tanay, G., & Bernstein, A. (2013). State Mindfulness Scale (SMS): Development and Initial Validation. *Psychological Assessment, 25*(4), 1286-1299.

Taylor, J. M. (2015). Psychometric analysis of the ten-item Perceived Stress Scale. *Psychological Assessment, 27*(1), 90-101.

Teasdale, J. D., Moore, R. G., Hayhurst, H., Pope, M., Williams, S., & Segal, Z. V. (2002). Meta-cognitive awareness and prevention of relapse in depression: Empirical evidence. *Journal of Councelling and Clinical Psychology, 70*(2), 275-287. doi: 10.1037/0022-006x.70.2.275

Teh, H. C., Archer, J. A., Chang, W., & Chen, S. H. A. (2015). Mental well-being mediates the relationship between perceived stress and perceived health. *Stress and Health, 31*(1), 71-77.

Tellegen, A. (1982). *Content categories: Absorption items* (Revised). Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A., Lykken, D. T., Bouchard, T. J., Wilcox, K. J., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology, 54*(6), 1031-1039.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism, 57*(8), 1358-1362.

Tennant, A., Pallant, J. F. (2006). Unidimensionality matters! (a tale of two Smiths?). *Rasch Measurement Transactions, 20*, 1048–1051.

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceralintegration perspective on self-regulation, adaptation and health. *The Society of Behavioral Medicine, 37*, 141-153.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*(3), 291-307.

Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology, 76*, 284–304.

Turner-Stokes, L., Fadyl, J., Rose, H., Williams, H., Schulter, P., & McPherson, K. M. (2014). The Work-ability Support Scale: Evaluation of Scoring Accuracy and Rater Reliability. *Journal of Rehabilitation Medicine, 24*(3), 511-24. doi: 10.1007/s10926-013-9

Turner-Stokes, L., Nyein, K., Turner-Stokes, T., & Gatehouse, C. (1999). The UK FIM+FAM: development and evaluation. *Clinical Rehabilitation*, *13,* 277-87.

Turner-Stokes, L., & Siegert, R. J. (2013). A comprehensive psychometric evaluation of the UK FIM + FAM. *Disability and Rehabilitation*, *35*, 1885-1895.

Turner-Stokes, L., Williams, H., Bill, A., Bassett, P., & Sephton, K. (2016). Cost-efficiency of specialist inpatient rehabilitation for working-aged adults with complex neurological disabilities: a multicentre cohort analysis of a national clinical data set. *British Medical Journal*, *6*(2), e010238.

Thurstone, L. L. (1931). The Measurement of Social Attitudes. *Abnormal and Social Psychology 27*, 249-269.

Twohig, M. P., Hayes, S. C., & Masuda, A. (2006). Increasing willingness to experience obsessions: Acceptance and Commitment Therapy as a treatment for obsessive compulsive disorder. *Behavior Therapy, 37*(1), 3-13.

Van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An Item Response Theory analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences, 49*, 805-810.

Van Dam, N. T., Earleywine, M., Danoff-Burg, S. (2009). Differential item function across meditators and non-meditators on the Five Facet Mindfulness Questionnaire. *Personality and Individual Differences, 47*, 516-521.

Visted, E., Vøllestad, J., Nielsen, M. B., & Nielsen, G. H. (2015). The Impact of Group-Based Mindfulness Trainingon Self-Reported Mindfulness: a SystematicReview and Meta-analysis. *Mindfulness, 6*, 501-522.

Wainer, H., & Kiely G. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement 24*(3), 185-201.

Walach, H., Buchheld, N., Buttenmuller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness-the Freiburg Mindfulness Inventory (FMI). *Personality and Individual Differences, 40*(8), 1543-1555.

Wallace, B. A. (1999). The Buddhist tradition of Samatha: Methods for refining and examining consciousness. *Journal of Consciousness Studies, 6*, 175-187.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measure of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Williams, M., & Penman, D. (2011) *Mindfulness: an eight-week plan to find peace in a frantic world.* New York: Rodale Inc.

Wilson, M. (2005). *Constructing measures*. Mahwah, NJ: LEA.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transaction, 10*, 509–511.

Wright, B., D. (2003) *Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test. Rasch Measurement Transactions, 17*(1), 905-906.

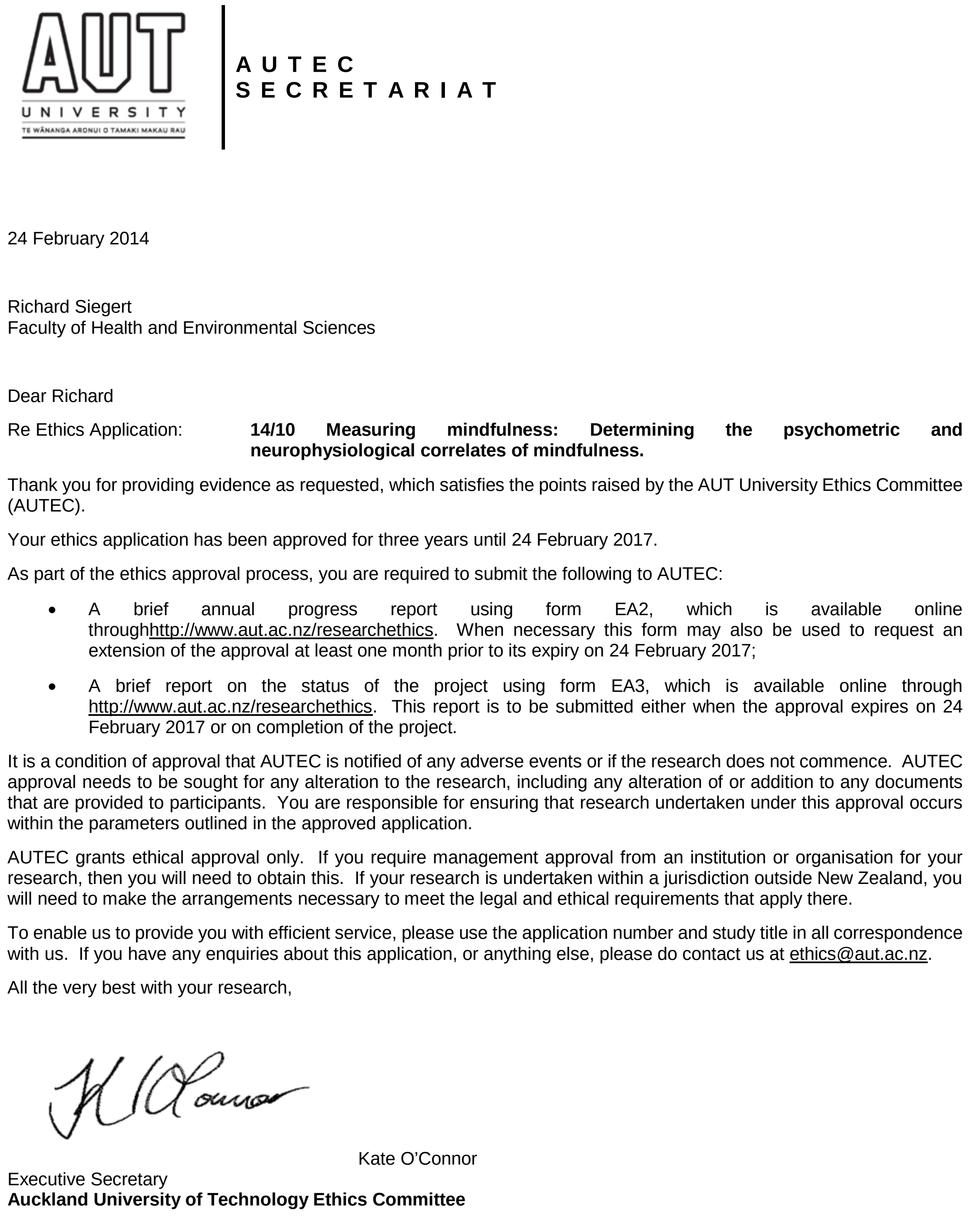
Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, USA: MESA Press.

Yelboga, A. (2015) Estimation of Generalizability Coefficient: An application with different programs. *Archives of Current Research International, 2*(1), 46-53.

Zettle, R. D., & Raines, J. C. (1989). Group cognitive and contextual therapies in treatment of depression. *Journal of Clinical Psychology, 45,* 438-445.

Zoogman, S., Goldberg, S. B., Hoyt, W. T., & Miller, L. (2015). Mindfulness interventions with youth: a meta-analysis. *Mindfulness, 6*(2), 290-302.

Appendix A1: AUTEC Approval (Chapters Three-Five, and Ten)



AUT UNIVERSITY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

AUTEC SECRETARIAT

24 February 2014

Richard Siegert
Faculty of Health and Environmental Sciences

Dear Richard

Re Ethics Application: **14/10 Measuring mindfulness: Determining the psychometric and neurophysiological correlates of mindfulness.**

Thank you for providing evidence as requested, which satisfies the points raised by the AUT University Ethics Committee (AUTEC).

Your ethics application has been approved for three years until 24 February 2017.

As part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online throughhttp://www.aut.ac.nz/researchethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 24 February 2017;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/researchethics. This report is to be submitted either when the approval expires on 24 February 2017 or on completion of the project.

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to obtain this. If your research is undertaken within a jurisdiction outside New Zealand, you will need to make the arrangements necessary to meet the legal and ethical requirements that apply there.

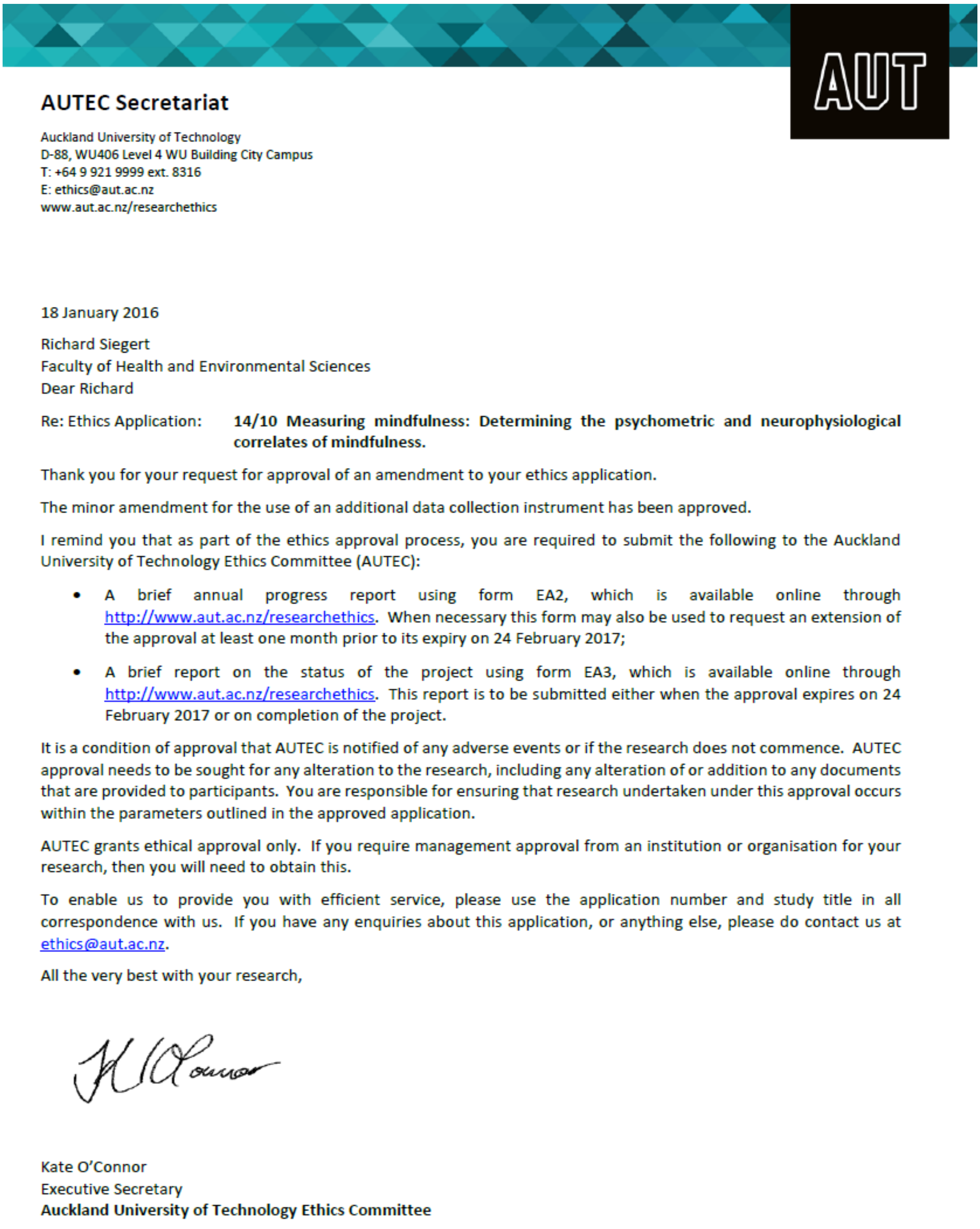
To enable us to provide you with efficient service, please use the application number and study title in all correspondence with us. If you have any enquiries about this application, or anything else, please do contact us at ethics@aut.ac.nz.

All the very best with your research,

K O'Connor

Kate O'Connor
Executive Secretary
Auckland University of Technology Ethics Committee

Appendix A2:AUTEC Approval for Amendments (Chapter Ten)



AUT

AUTEC Secretariat

Auckland University of Technology
D-88, WU406 Level 4 WU Building City Campus
T: +64 9 921 9999 ext. 8316
E: ethics@aut.ac.nz
www.aut.ac.nz/researchethics

18 January 2016

Richard Siegert
Faculty of Health and Environmental Sciences
Dear Richard

Re: Ethics Application: **14/10 Measuring mindfulness: Determining the psychometric and neurophysiological correlates of mindfulness.**

Thank you for your request for approval of an amendment to your ethics application.

The minor amendment for the use of an additional data collection instrument has been approved.

I remind you that as part of the ethics approval process, you are required to submit the following to the Auckland University of Technology Ethics Committee (AUTEC):

- A brief annual progress report using form EA2, which is available online through http://www.aut.ac.nz/researchethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 24 February 2017;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/researchethics. This report is to be submitted either when the approval expires on 24 February 2017 or on completion of the project.

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to obtain this.

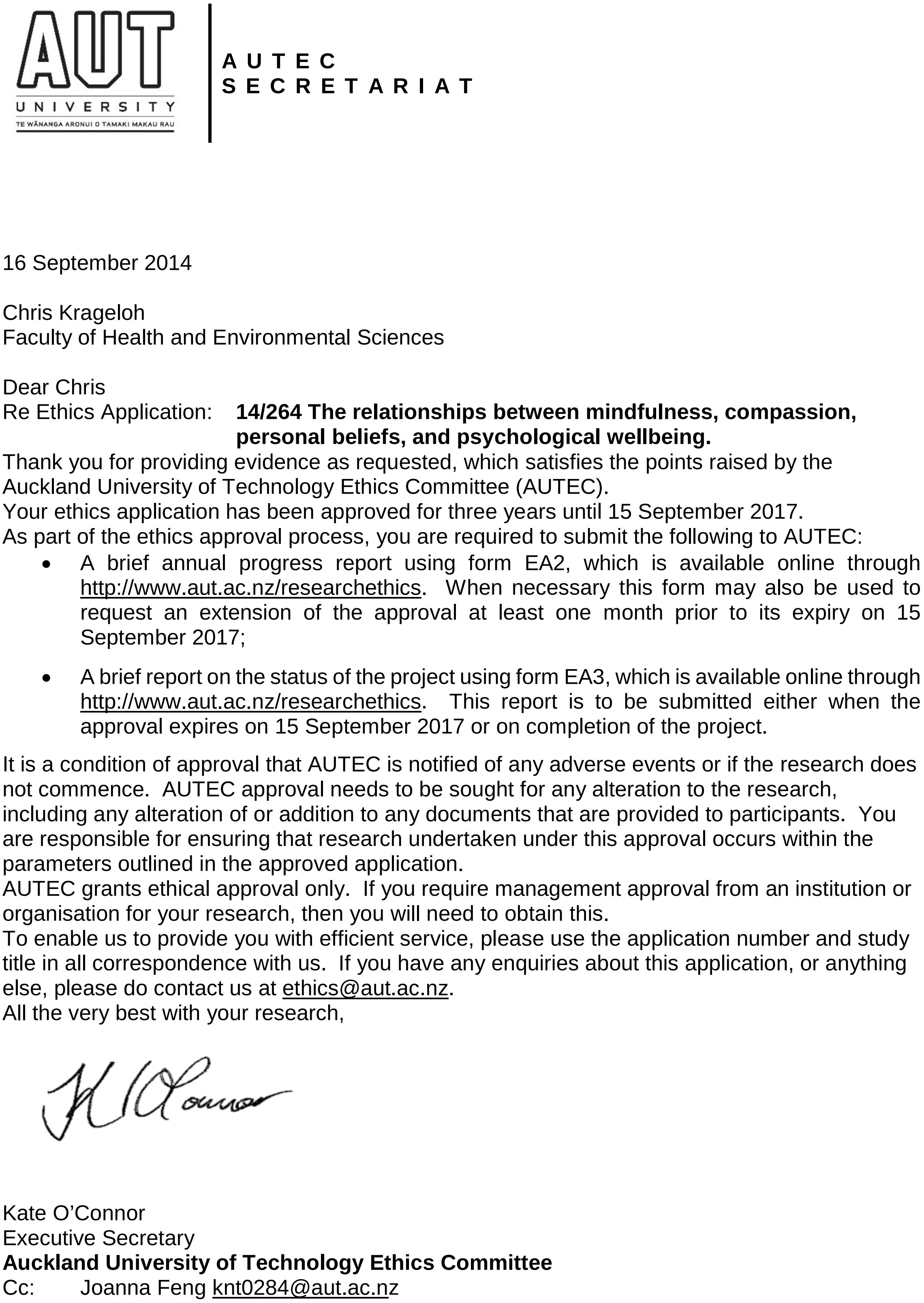
To enable us to provide you with efficient service, please use the application number and study title in all correspondence with us. If you have any enquiries about this application, or anything else, please do contact us at ethics@aut.ac.nz.

All the very best with your research,

K O'Connor

Kate O'Connor
Executive Secretary
Auckland University of Technology Ethics Committee

Appendix A3: AUTEC Approval for the additional data (Chapter Three)



AUTEC
SECRETARIAT

16 September 2014

Chris Krageloh
Faculty of Health and Environmental Sciences

Dear Chris
Re Ethics Application: **14/264 The relationships between mindfulness, compassion, personal beliefs, and psychological wellbeing.**

Thank you for providing evidence as requested, which satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC).

Your ethics application has been approved for three years until 15 September 2017.

As part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online through http://www.aut.ac.nz/researchethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 15 September 2017;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/researchethics. This report is to be submitted either when the approval expires on 15 September 2017 or on completion of the project.

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to obtain this.

To enable us to provide you with efficient service, please use the application number and study title in all correspondence with us. If you have any enquiries about this application, or anything else, please do contact us at ethics@aut.ac.nz.

All the very best with your research,

Kate O’Connor
Executive Secretary
Auckland University of Technology Ethics Committee
Cc: Joanna Feng knt0284@aut.ac.nz

Appendix A4: AUTEC Approval (Chapter Seven)

AUT UNIVERSITY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

MEMORANDUM
Auckland University of Technology Ethics Committee (AUTEC)

To: Daniel Shepherd
From: **Dr Rosemary Godbold and Madeline Banda** Executive Secretary, AUTEC
Date: 16 May 2011
Subject: Ethics Application Number 10/271 **Noise, health and environmental perceptions among Auckland residents.**

Dear Daniel

Thank you for providing written evidence as requested. We are pleased to advise that it satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC) at their meeting on 8 November 2010 and that on 18 February 2011, we approved your ethics application. This delegated approval is made in accordance with section 5.3.2.3 of AUTEC's *Applying for Ethics Approval: Guidelines and Procedures* and is subject to endorsement at AUTEC's meeting on 13 June 2011.

Your ethics application is approved for a period of three years until 18 February 2014.

We advise that as part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online through http://www.aut.ac.nz/research/research-ethics/ethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 18 February 2014;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/research/research-ethics/ethics. This report is to be submitted either when the approval expires on 18 February 2014 or on completion of the project, whichever comes sooner;

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are reminded that, as applicant, you are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

Please note that AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to make the arrangements necessary to obtain this.

When communicating with us about this application, we ask that you use the application number and study title to enable us to provide you with prompt service. Should you have any further enquiries regarding this matter, you are welcome to contact Charles Grinter, Ethics Coordinator, by email at ethics@aut.ac.nz or by telephone on 921 9999 at extension 8860.

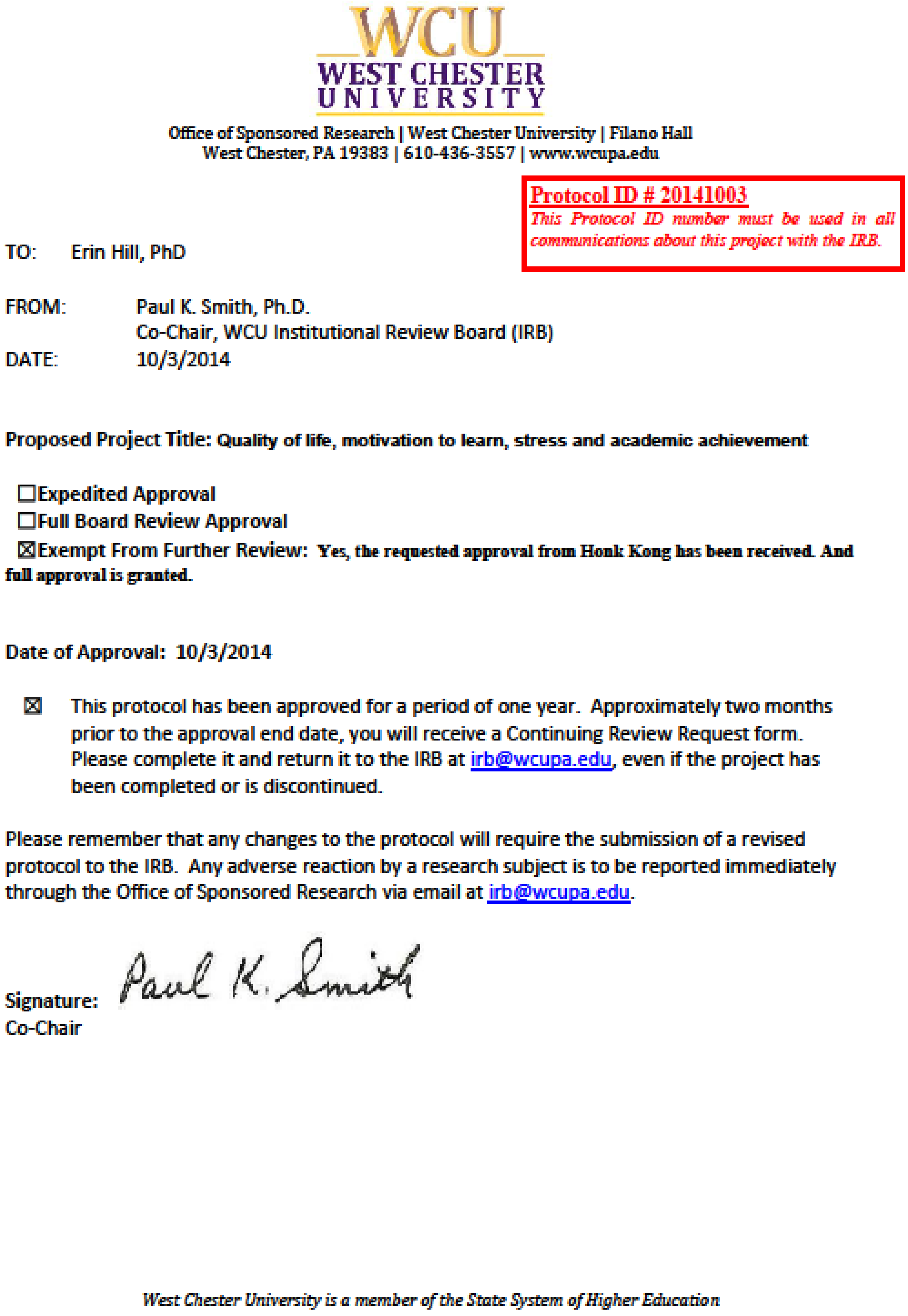
On behalf of AUTEC and ourselves, we wish you success with your research and look forward to reading about it in your reports.

Yours sincerely

Dr Rosemary Godbold and Madeline Banda
Executive Secretary
Auckland University of Technology Ethics Committee

Cc: Erin Hill Erin.hill@aut.ac.nz

Appendix A5: WCU Approval for the data used in the Study (Chapter Seven)



WCU
WEST CHESTER
UNIVERSITY

Office of Sponsored Research | West Chester University | Filano Hall
West Chester, PA 19383 | 610-436-3557 | www.wcupa.edu

Protocol ID # 20141003
This Protocol ID number must be used in all communications about this project with the IRB.

TO: Erin Hill, PhD

FROM: Paul K. Smith, Ph.D.
Co-Chair, WCU Institutional Review Board (IRB)
DATE: 10/3/2014

Proposed Project Title: **Quality of life, motivation to learn, stress and academic achievement**

☐Expedited Approval
☐Full Board Review Approval
☒Exempt From Further Review: **Yes, the requested approval from Honk Kong has been received. And full approval is granted.**

Date of Approval: 10/3/2014

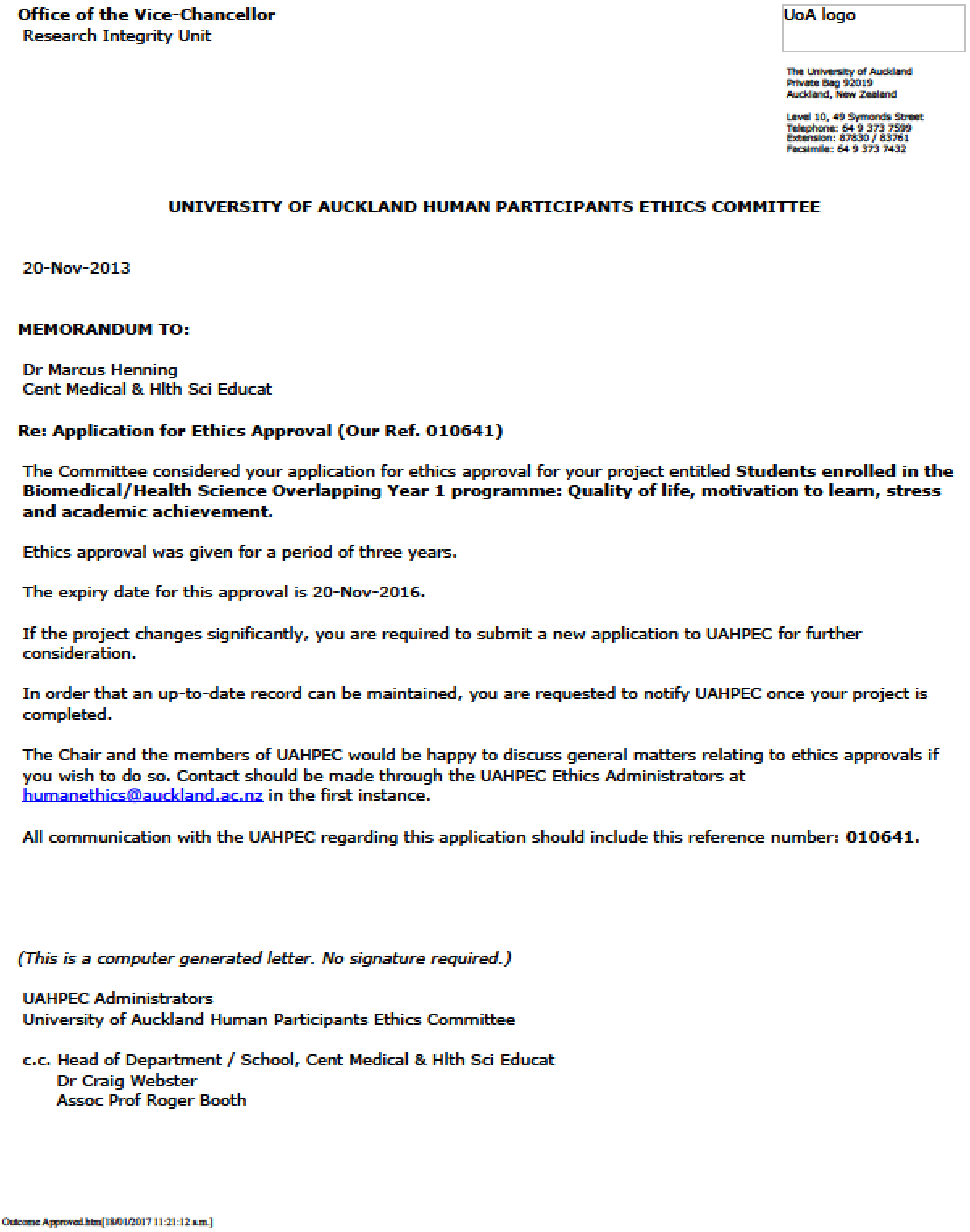
☒ This protocol has been approved for a period of one year. Approximately two months prior to the approval end date, you will receive a Continuing Review Request form. Please complete it and return it to the IRB at irb@wcupa.edu, even if the project has been completed or is discontinued.

Please remember that any changes to the protocol will require the submission of a revised protocol to the IRB. Any adverse reaction by a research subject is to be reported immediately through the Office of Sponsored Research via email at irb@wcupa.edu.

Signature: Paul K. Smith
Co-Chair

West Chester University is a member of the State System of Higher Education

Appendix A6: UoA Approval for the data used in the Study (Chapter Seven)



Office of the Vice-Chancellor
Research Integrity Unit

UoA logo

The University of Auckland
Private Bag 92019
Auckland, New Zealand

Level 10, 49 Symonds Street
Telephone: 64 9 373 7599
Extension: 87830 / 83761
Facsimile: 64 9 373 7432

UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE

20-Nov-2013

MEMORANDUM TO:

Dr Marcus Henning
Cent Medical & Hlth Sci Educat

Re: Application for Ethics Approval (Our Ref. 010641)

The Committee considered your application for ethics approval for your project entitled **Students enrolled in the Biomedical/Health Science Overlapping Year 1 programme: Quality of life, motivation to learn, stress and academic achievement.**

Ethics approval was given for a period of three years.

The expiry date for this approval is 20-Nov-2016.

If the project changes significantly, you are required to submit a new application to UAHPEC for further consideration.

In order that an up-to-date record can be maintained, you are requested to notify UAHPEC once your project is completed.

The Chair and the members of UAHPEC would be happy to discuss general matters relating to ethics approvals if you wish to do so. Contact should be made through the UAHPEC Ethics Administrators at humanethics@auckland.ac.nz in the first instance.

All communication with the UAHPEC regarding this application should include this reference number: **010641.**

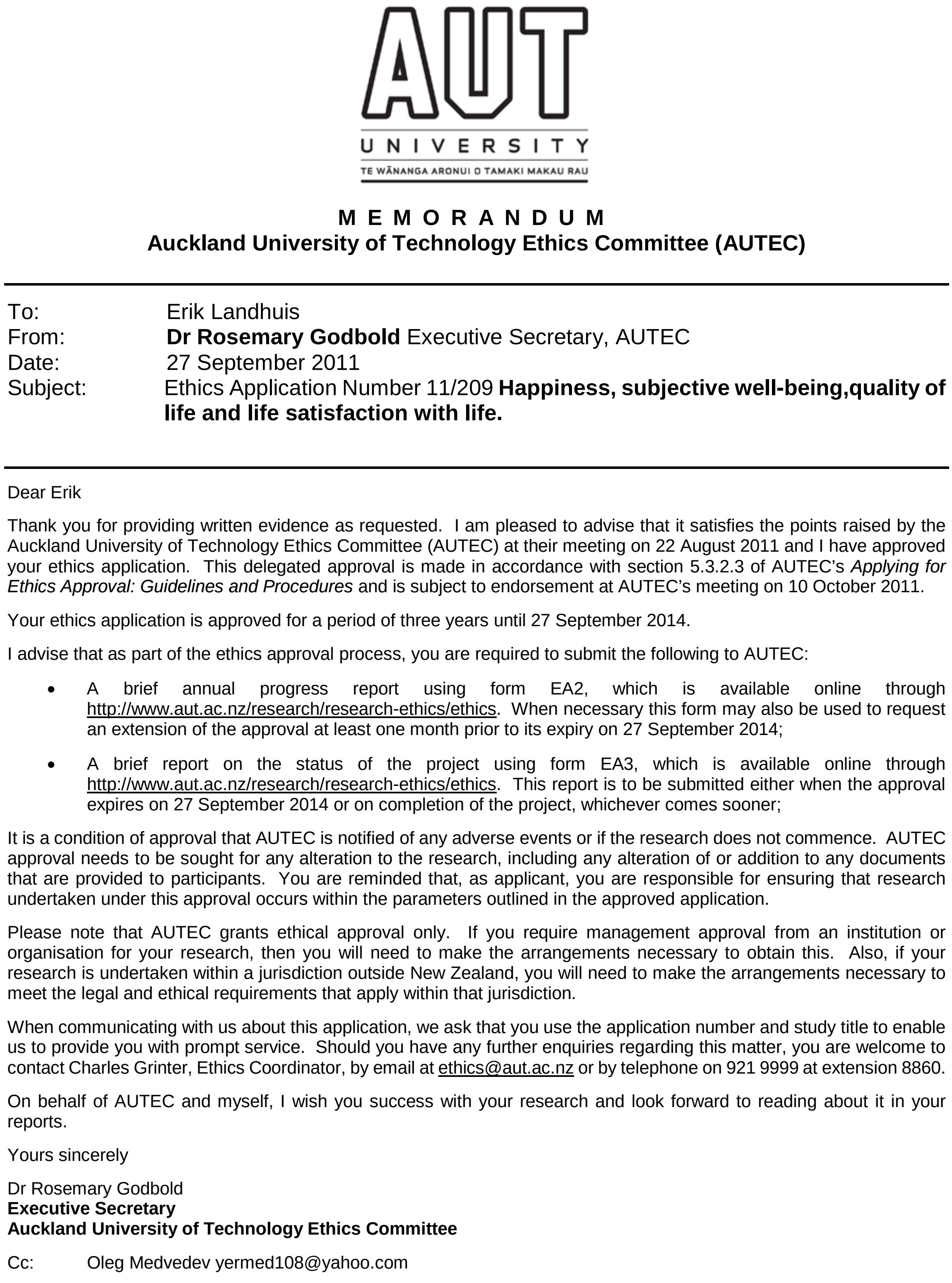
(This is a computer generated letter. No signature required.)

UAHPEC Administrators
University of Auckland Human Participants Ethics Committee

c.c. Head of Department / School, Cent Medical & Hlth Sci Educat
Dr Craig Webster
Assoc Prof Roger Booth

Outcome Approved.htm[18/01/2017 11:21:12 a.m.]

Appendix A7: AUTEC Approval (Chapter Eight, New Zealand)



AUT UNIVERSITY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

MEMORANDUM
Auckland University of Technology Ethics Committee (AUTEC)

To:	Erik Landhuis
From:	**Dr Rosemary Godbold** Executive Secretary, AUTEC
Date:	27 September 2011
Subject:	Ethics Application Number 11/209 **Happiness, subjective well-being,quality of life and life satisfaction with life.**

Dear Erik

Thank you for providing written evidence as requested. I am pleased to advise that it satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC) at their meeting on 22 August 2011 and I have approved your ethics application. This delegated approval is made in accordance with section 5.3.2.3 of AUTEC's *Applying for Ethics Approval: Guidelines and Procedures* and is subject to endorsement at AUTEC's meeting on 10 October 2011.

Your ethics application is approved for a period of three years until 27 September 2014.

I advise that as part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online through http://www.aut.ac.nz/research/research-ethics/ethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 27 September 2014;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/research/research-ethics/ethics. This report is to be submitted either when the approval expires on 27 September 2014 or on completion of the project, whichever comes sooner;

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are reminded that, as applicant, you are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

Please note that AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to make the arrangements necessary to obtain this. Also, if your research is undertaken within a jurisdiction outside New Zealand, you will need to make the arrangements necessary to meet the legal and ethical requirements that apply within that jurisdiction.

When communicating with us about this application, we ask that you use the application number and study title to enable us to provide you with prompt service. Should you have any further enquiries regarding this matter, you are welcome to contact Charles Grinter, Ethics Coordinator, by email at ethics@aut.ac.nz or by telephone on 921 9999 at extension 8860.

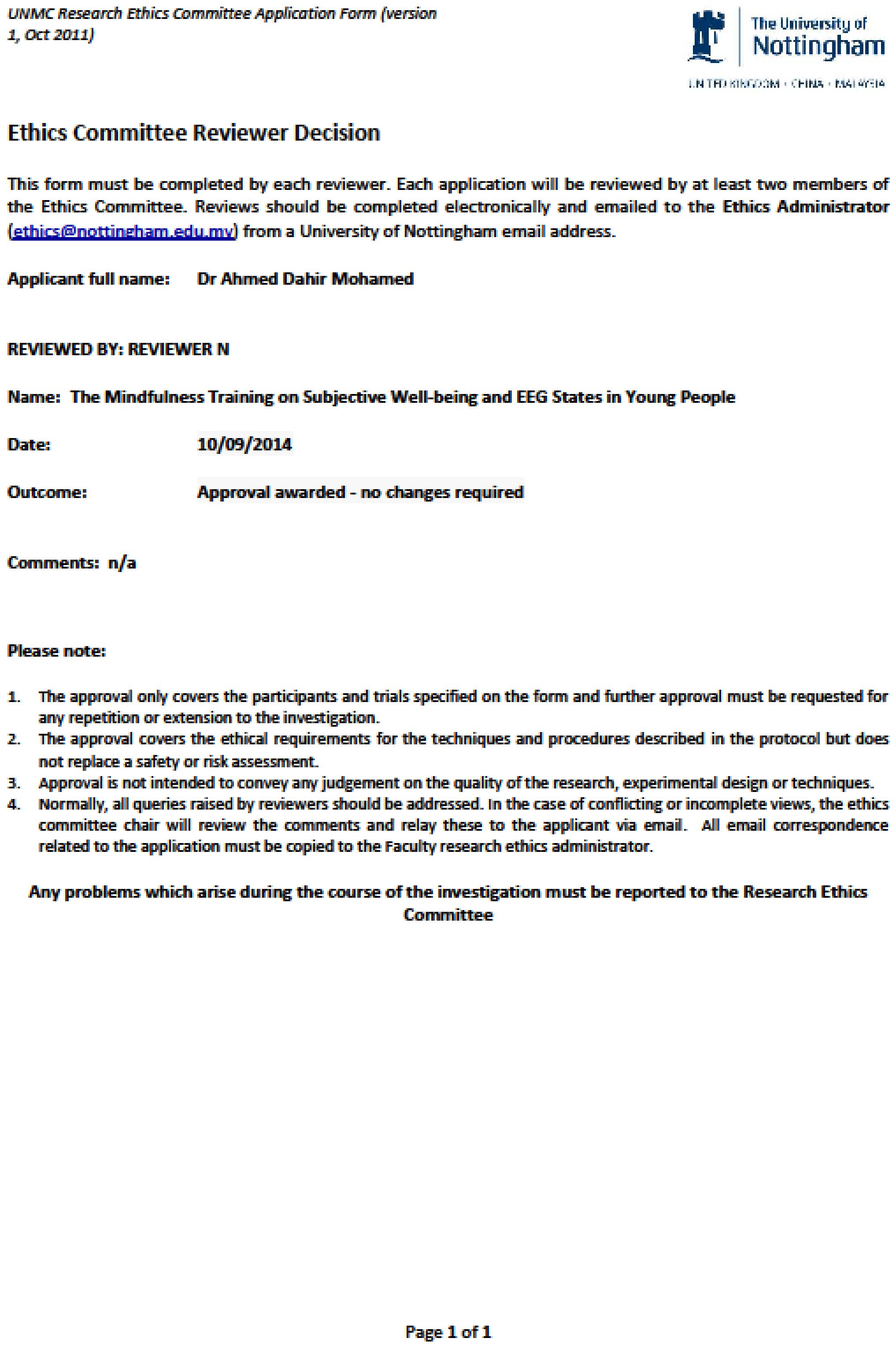
On behalf of AUTEC and myself, I wish you success with your research and look forward to reading about it in your reports.

Yours sincerely

Dr Rosemary Godbold
Executive Secretary
Auckland University of Technology Ethics Committee

Cc: Oleg Medvedev yermed108@yahoo.com

Appendix A8: University of Nottingham Approval (Chapter Eight, Malaysia)



UNMC Research Ethics Committee Application Form (version 1, Oct 2011)

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Ethics Committee Reviewer Decision

This form must be completed by each reviewer. Each application will be reviewed by at least two members of the Ethics Committee. Reviews should be completed electronically and emailed to the Ethics Administrator (ethics@nottingham.edu.my) from a University of Nottingham email address.

Applicant full name: **Dr Ahmed Dahir Mohamed**

REVIEWED BY: REVIEWER N

Name: The Mindfulness Training on Subjective Well-being and EEG States in Young People

Date: **10/09/2014**

Outcome: **Approval awarded - no changes required**

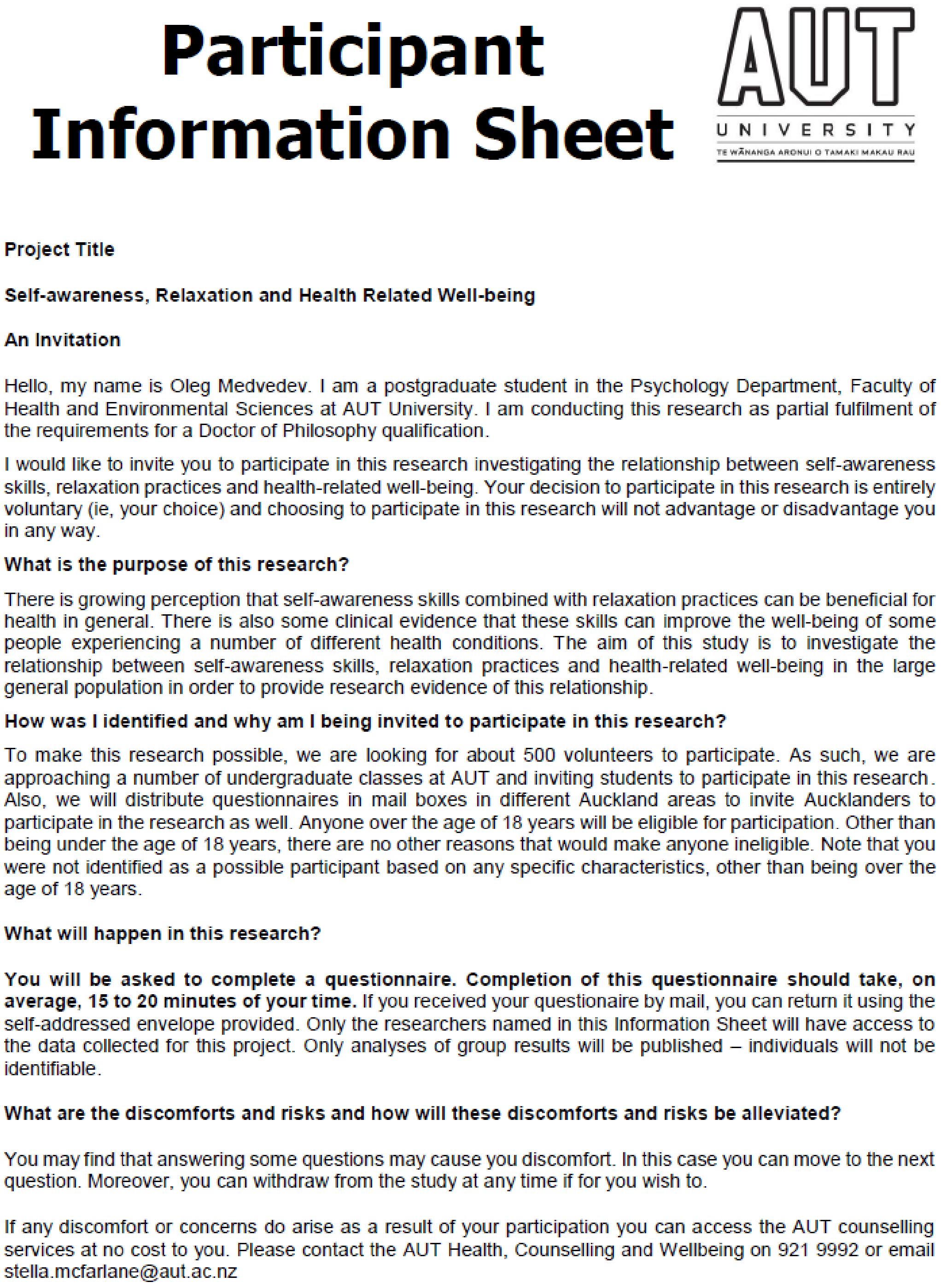
Comments: n/a

Please note:

1. The approval only covers the participants and trials specified on the form and further approval must be requested for any repetition or extension to the investigation.
2. The approval covers the ethical requirements for the techniques and procedures described in the protocol but does not replace a safety or risk assessment.
3. Approval is not intended to convey any judgement on the quality of the research, experimental design or techniques.
4. Normally, all queries raised by reviewers should be addressed. In the case of conflicting or incomplete views, the ethics committee chair will review the comments and relay these to the applicant via email. All email correspondence related to the application must be copied to the Faculty research ethics administrator.

Any problems which arise during the course of the investigation must be reported to the Research Ethics Committee

Appendix B: Participant Information Sheet (Chapters Three-Five and Ten)



Participant Information Sheet

AUT UNIVERSITY TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Project Title

Self-awareness, Relaxation and Health Related Well-being

An Invitation

Hello, my name is Oleg Medvedev. I am a postgraduate student in the Psychology Department, Faculty of Health and Environmental Sciences at AUT University. I am conducting this research as partial fulfilment of the requirements for a Doctor of Philosophy qualification.

I would like to invite you to participate in this research investigating the relationship between self-awareness skills, relaxation practices and health-related well-being. Your decision to participate in this research is entirely voluntary (ie, your choice) and choosing to participate in this research will not advantage or disadvantage you in any way.

What is the purpose of this research?

There is growing perception that self-awareness skills combined with relaxation practices can be beneficial for health in general. There is also some clinical evidence that these skills can improve the well-being of some people experiencing a number of different health conditions. The aim of this study is to investigate the relationship between self-awareness skills, relaxation practices and health-related well-being in the large general population in order to provide research evidence of this relationship.

How was I identified and why am I being invited to participate in this research?

To make this research possible, we are looking for about 500 volunteers to participate. As such, we are approaching a number of undergraduate classes at AUT and inviting students to participate in this research. Also, we will distribute questionnaires in mail boxes in different Auckland areas to invite Aucklanders to participate in the research as well. Anyone over the age of 18 years will be eligible for participation. Other than being under the age of 18 years, there are no other reasons that would make anyone ineligible. Note that you were not identified as a possible participant based on any specific characteristics, other than being over the age of 18 years.

What will happen in this research?

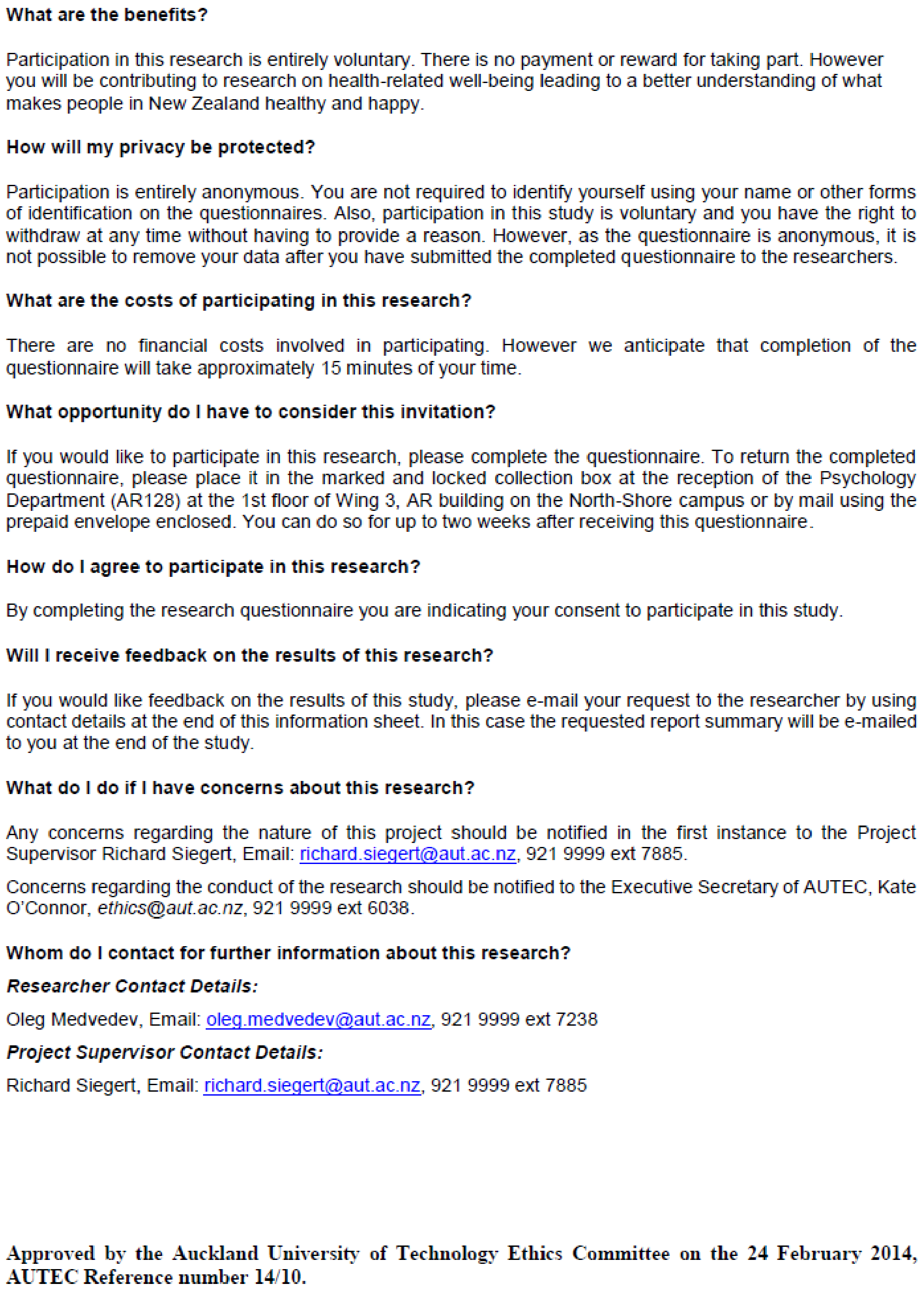
You will be asked to complete a questionnaire. Completion of this questionnaire should take, on average, 15 to 20 minutes of your time. If you received your questionaire by mail, you can return it using the self-addressed envelope provided. Only the researchers named in this Information Sheet will have access to the data collected for this project. Only analyses of group results will be published – individuals will not be identifiable.

What are the discomforts and risks and how will these discomforts and risks be alleviated?

You may find that answering some questions may cause you discomfort. In this case you can move to the next question. Moreover, you can withdraw from the study at any time if for you wish to.

If any discomfort or concerns do arise as a result of your participation you can access the AUT counselling services at no cost to you. Please contact the AUT Health, Counselling and Wellbeing on 921 9992 or email stella.mcfarlane@aut.ac.nz

Appendix B: Participant Information Sheet (Chapters Three-Five and Ten)



What are the benefits?

Participation in this research is entirely voluntary. There is no payment or reward for taking part. However you will be contributing to research on health-related well-being leading to a better understanding of what makes people in New Zealand healthy and happy.

How will my privacy be protected?

Participation is entirely anonymous. You are not required to identify yourself using your name or other forms of identification on the questionnaires. Also, participation in this study is voluntary and you have the right to withdraw at any time without having to provide a reason. However, as the questionnaire is anonymous, it is not possible to remove your data after you have submitted the completed questionnaire to the researchers.

What are the costs of participating in this research?

There are no financial costs involved in participating. However we anticipate that completion of the questionnaire will take approximately 15 minutes of your time.

What opportunity do I have to consider this invitation?

If you would like to participate in this research, please complete the questionnaire. To return the completed questionnaire, please place it in the marked and locked collection box at the reception of the Psychology Department (AR128) at the 1st floor of Wing 3, AR building on the North-Shore campus or by mail using the prepaid envelope enclosed. You can do so for up to two weeks after receiving this questionnaire.

How do I agree to participate in this research?

By completing the research questionnaire you are indicating your consent to participate in this study.

Will I receive feedback on the results of this research?

If you would like feedback on the results of this study, please e-mail your request to the researcher by using contact details at the end of this information sheet. In this case the requested report summary will be e-mailed to you at the end of the study.

What do I do if I have concerns about this research?

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor Richard Siegert, Email: richard.siegert@aut.ac.nz, 921 9999 ext 7885.

Concerns regarding the conduct of the research should be notified to the Executive Secretary of AUTEC, Kate O'Connor, *ethics@aut.ac.nz*, 921 9999 ext 6038.

Whom do I contact for further information about this research?

Researcher Contact Details:

Oleg Medvedev, Email: oleg.medvedev@aut.ac.nz, 921 9999 ext 7238

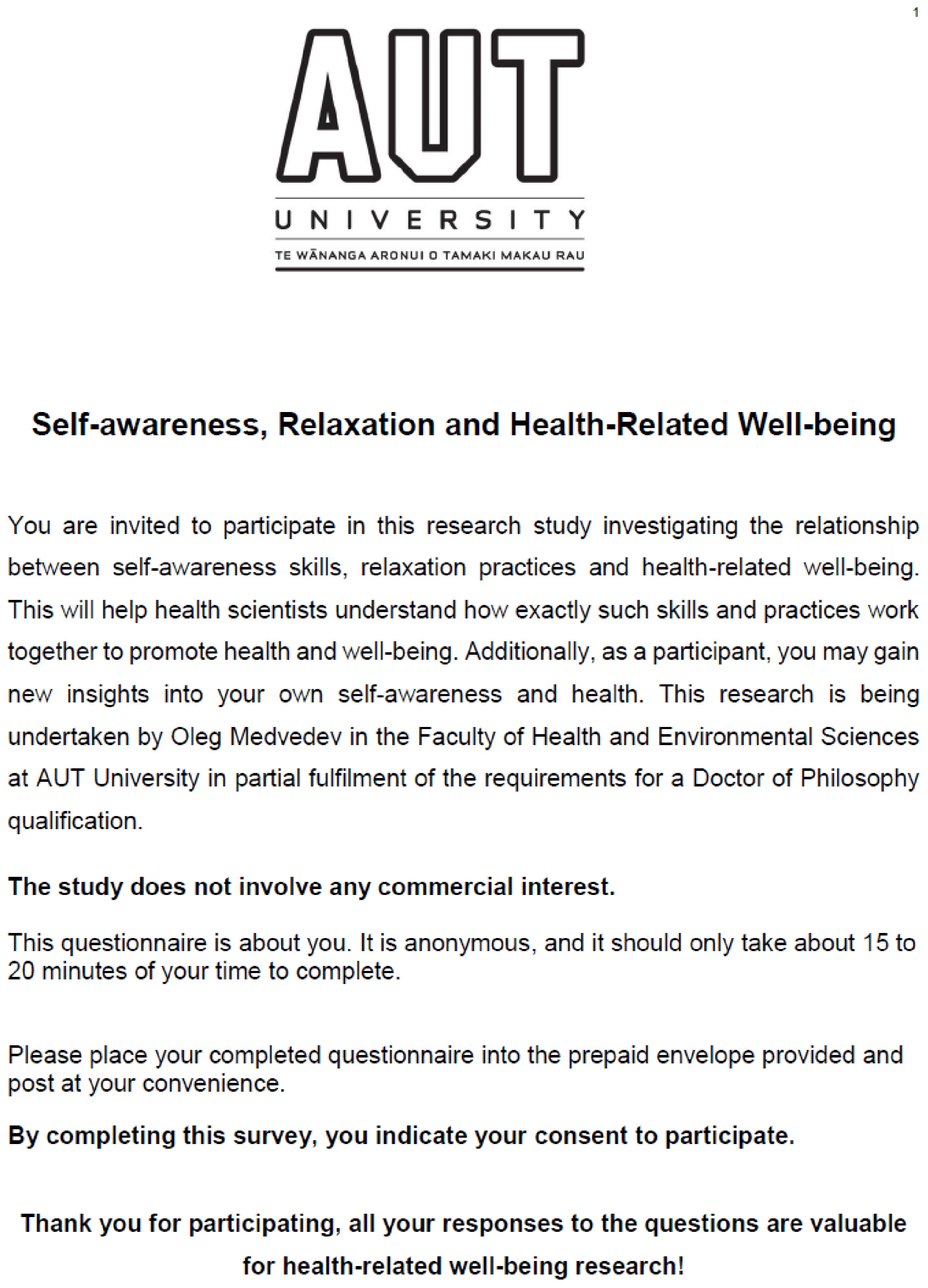
Project Supervisor Contact Details:

Richard Siegert, Email: richard.siegert@aut.ac.nz, 921 9999 ext 7885

Approved by the Auckland University of Technology Ethics Committee on the 24 February 2014, AUTEC Reference number 14/10.

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Title page: Post version



1

AUT

UNIVERSITY

TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Self-awareness, Relaxation and Health-Related Well-being

You are invited to participate in this research study investigating the relationship between self-awareness skills, relaxation practices and health-related well-being. This will help health scientists understand how exactly such skills and practices work together to promote health and well-being. Additionally, as a participant, you may gain new insights into your own self-awareness and health. This research is being undertaken by Oleg Medvedev in the Faculty of Health and Environmental Sciences at AUT University in partial fulfilment of the requirements for a Doctor of Philosophy qualification.

The study does not involve any commercial interest.

This questionnaire is about you. It is anonymous, and it should only take about 15 to 20 minutes of your time to complete.

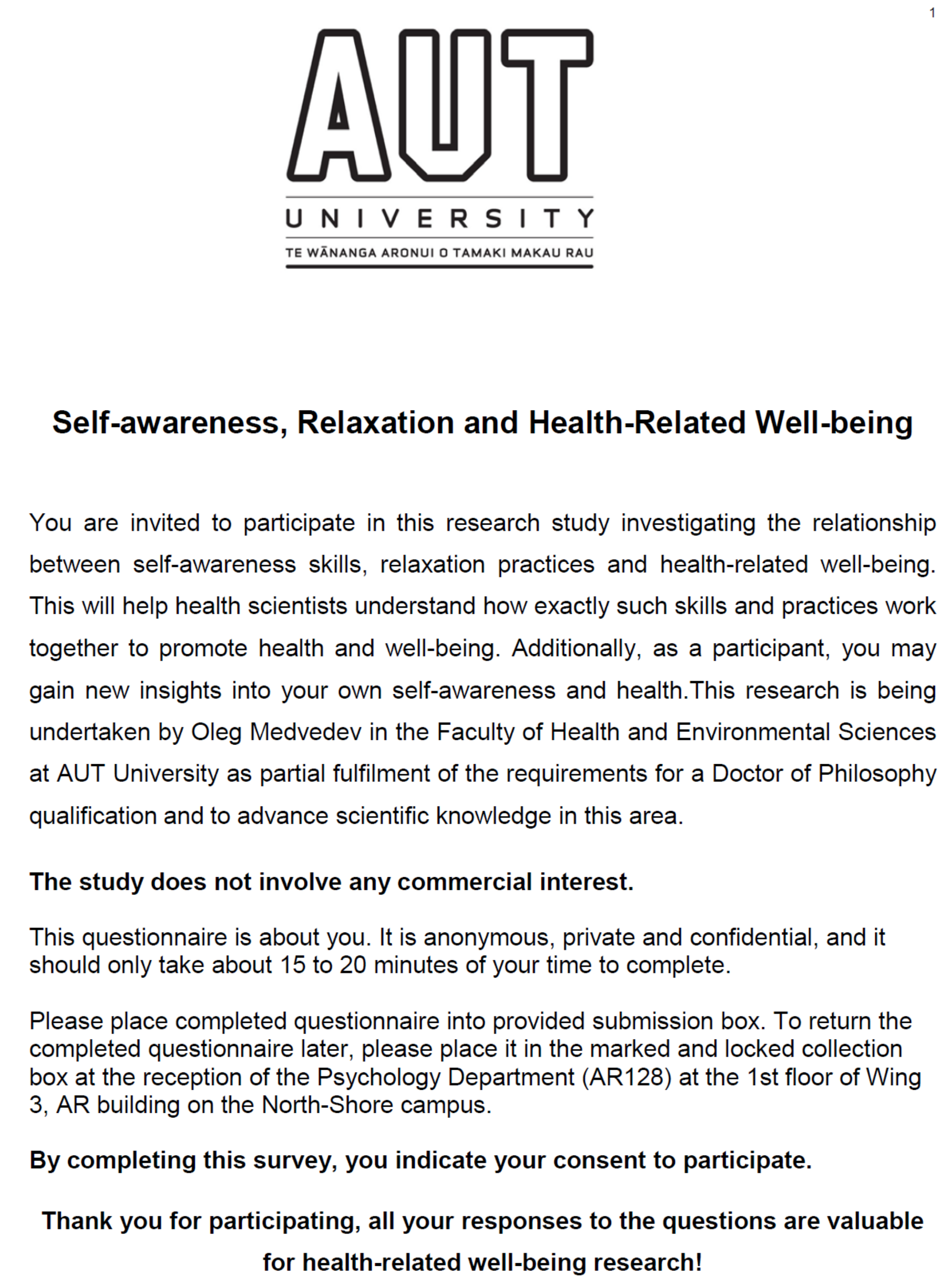
Please place your completed questionnaire into the prepaid envelope provided and post at your convenience.

By completing this survey, you indicate your consent to participate.

Thank you for participating, all your responses to the questions are valuable for health-related well-being research!

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Title page: Class version



Self-awareness, Relaxation and Health-Related Well-being

You are invited to participate in this research study investigating the relationship between self-awareness skills, relaxation practices and health-related well-being. This will help health scientists understand how exactly such skills and practices work together to promote health and well-being. Additionally, as a participant, you may gain new insights into your own self-awareness and health.This research is being undertaken by Oleg Medvedev in the Faculty of Health and Environmental Sciences at AUT University as partial fulfilment of the requirements for a Doctor of Philosophy qualification and to advance scientific knowledge in this area.

The study does not involve any commercial interest.

This questionnaire is about you. It is anonymous, private and confidential, and it should only take about 15 to 20 minutes of your time to complete.

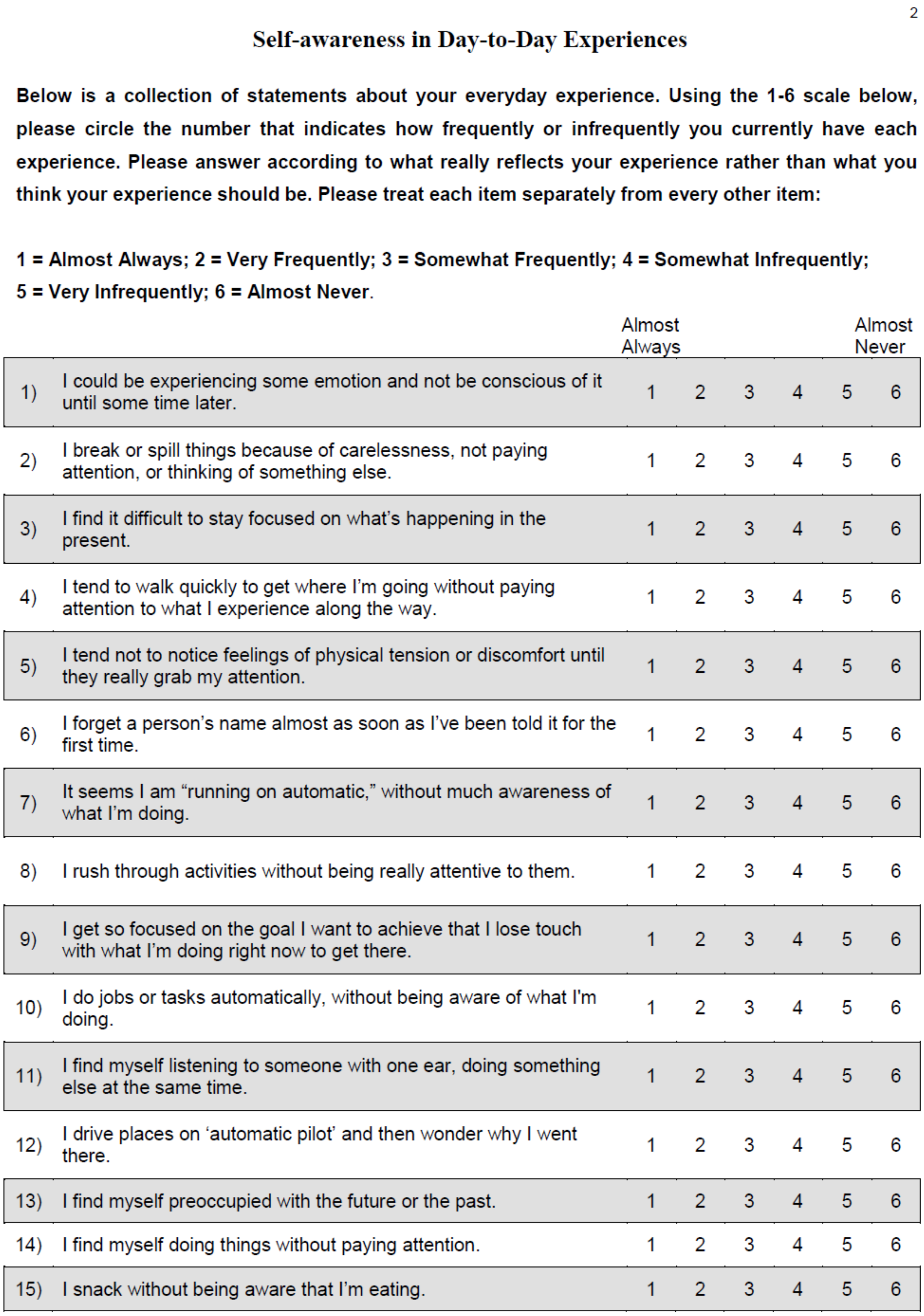
Please place completed questionnaire into provided submission box. To return the completed questionnaire later, please place it in the marked and locked collection box at the reception of the Psychology Department (AR128) at the 1st floor of Wing 3, AR building on the North-Shore campus.

By completing this survey, you indicate your consent to participate.

Thank you for participating, all your responses to the questions are valuable for health-related well-being research!

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Three: The MAAS



Self-awareness in Day-to-Day Experiences

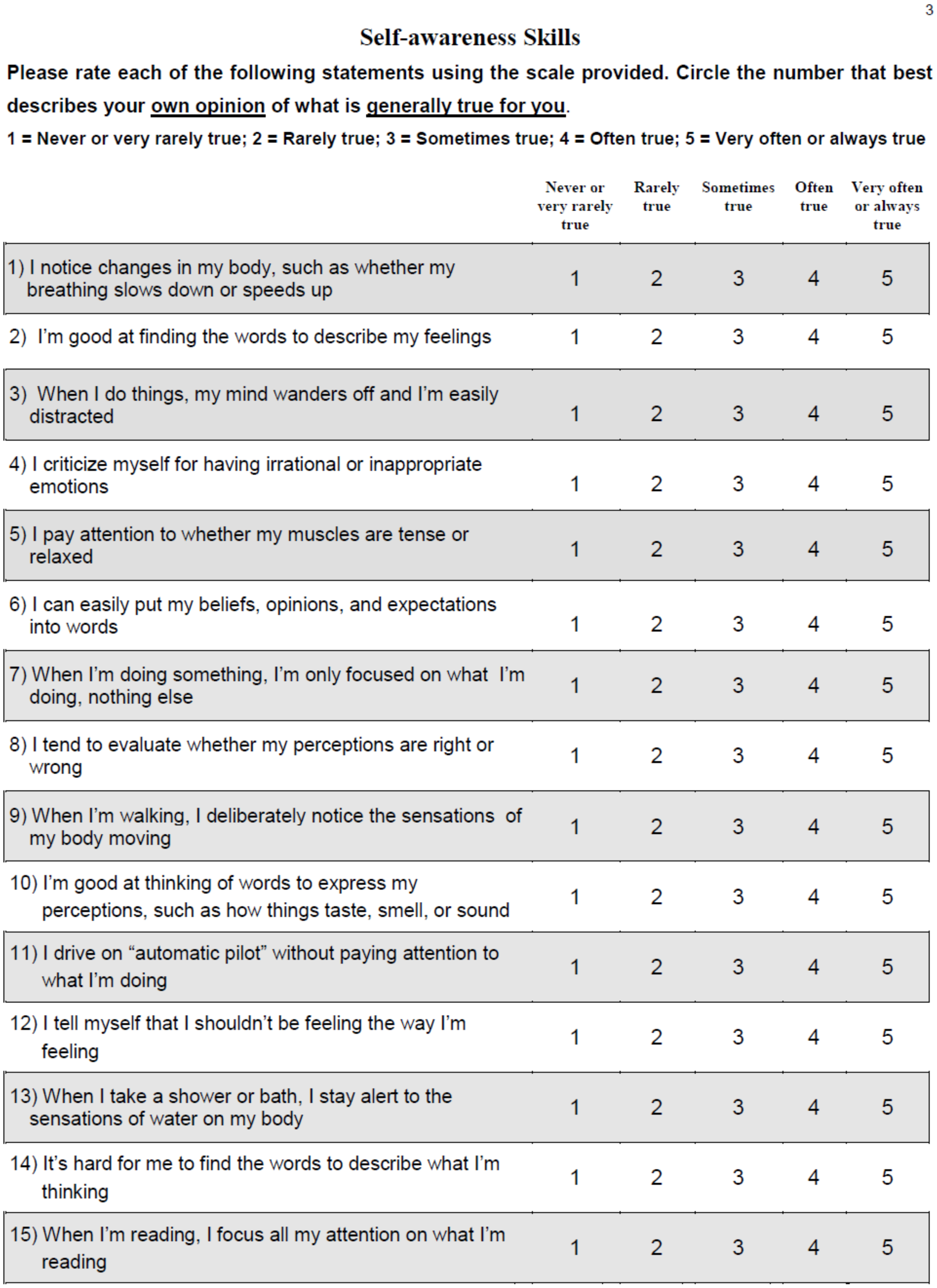
Below is a collection of statements about your everyday experience. Using the 1-6 scale below, please circle the number that indicates how frequently or infrequently you currently have each experience. Please answer according to what really reflects your experience rather than what you think your experience should be. Please treat each item separately from every other item:

1 = Almost Always; 2 = Very Frequently; 3 = Somewhat Frequently; 4 = Somewhat Infrequently; 5 = Very Infrequently; 6 = Almost Never.

		Almost Always					Almost Never
1)	I could be experiencing some emotion and not be conscious of it until some time later.	1	2	3	4	5	6
2)	I break or spill things because of carelessness, not paying attention, or thinking of something else.	1	2	3	4	5	6
3)	I find it difficult to stay focused on what’s happening in the present.	1	2	3	4	5	6
4)	I tend to walk quickly to get where I’m going without paying attention to what I experience along the way.	1	2	3	4	5	6
5)	I tend not to notice feelings of physical tension or discomfort until they really grab my attention.	1	2	3	4	5	6
6)	I forget a person’s name almost as soon as I’ve been told it for the first time.	1	2	3	4	5	6
7)	It seems I am “running on automatic,” without much awareness of what I’m doing.	1	2	3	4	5	6
8)	I rush through activities without being really attentive to them.	1	2	3	4	5	6
9)	I get so focused on the goal I want to achieve that I lose touch with what I’m doing right now to get there.	1	2	3	4	5	6
10)	I do jobs or tasks automatically, without being aware of what I'm doing.	1	2	3	4	5	6
11)	I find myself listening to someone with one ear, doing something else at the same time.	1	2	3	4	5	6
12)	I drive places on ‘automatic pilot’ and then wonder why I went there.	1	2	3	4	5	6
13)	I find myself preoccupied with the future or the past.	1	2	3	4	5	6
14)	I find myself doing things without paying attention.	1	2	3	4	5	6
15)	I snack without being aware that I’m eating.	1	2	3	4	5	6

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Four: The KIMS



3

Self-awareness Skills

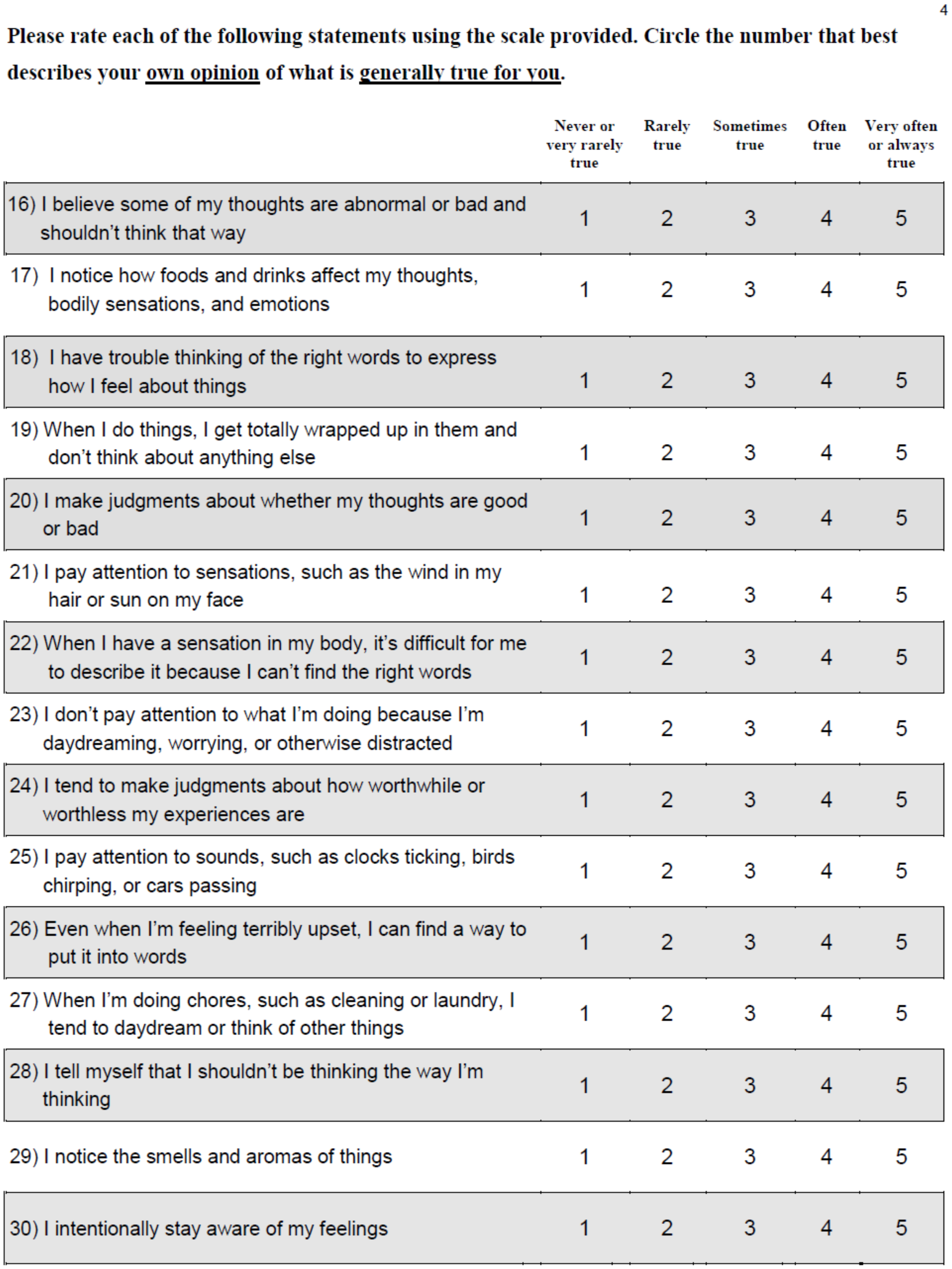
Please rate each of the following statements using the scale provided. Circle the number that best describes your <u>own opinion</u> of what is <u>generally true for you</u>.

1 = Never or very rarely true; 2 = Rarely true; 3 = Sometimes true; 4 = Often true; 5 = Very often or always true

	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
1) I notice changes in my body, such as whether my breathing slows down or speeds up	1	2	3	4	5
2) I'm good at finding the words to describe my feelings	1	2	3	4	5
3) When I do things, my mind wanders off and I'm easily distracted	1	2	3	4	5
4) I criticize myself for having irrational or inappropriate emotions	1	2	3	4	5
5) I pay attention to whether my muscles are tense or relaxed	1	2	3	4	5
6) I can easily put my beliefs, opinions, and expectations into words	1	2	3	4	5
7) When I'm doing something, I'm only focused on what I'm doing, nothing else	1	2	3	4	5
8) I tend to evaluate whether my perceptions are right or wrong	1	2	3	4	5
9) When I'm walking, I deliberately notice the sensations of my body moving	1	2	3	4	5
10) I'm good at thinking of words to express my perceptions, such as how things taste, smell, or sound	1	2	3	4	5
11) I drive on "automatic pilot" without paying attention to what I'm doing	1	2	3	4	5
12) I tell myself that I shouldn't be feeling the way I'm feeling	1	2	3	4	5
13) When I take a shower or bath, I stay alert to the sensations of water on my body	1	2	3	4	5
14) It's hard for me to find the words to describe what I'm thinking	1	2	3	4	5
15) When I'm reading, I focus all my attention on what I'm reading	1	2	3	4	5

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Four: The KIMS



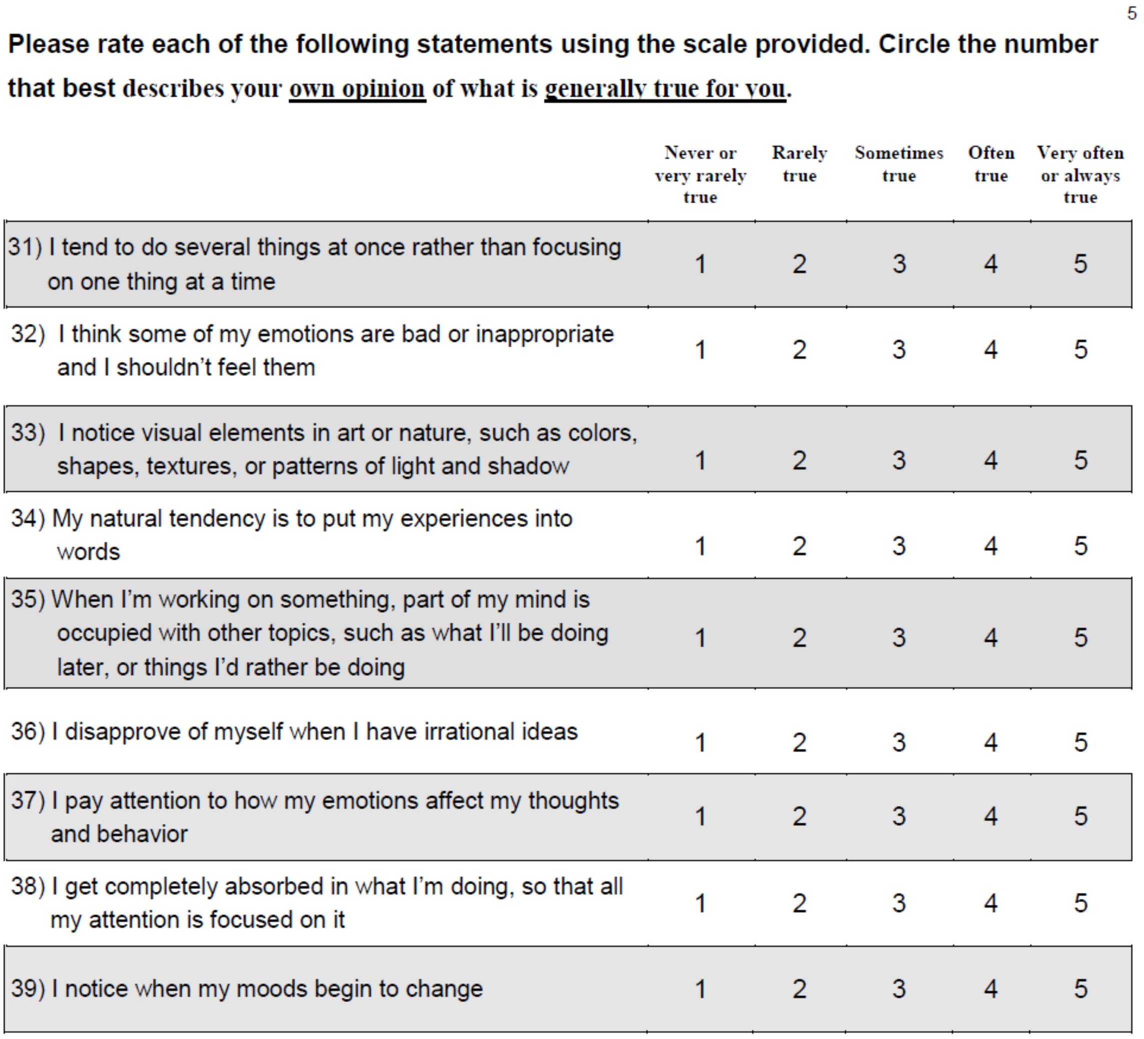
4

Please rate each of the following statements using the scale provided. Circle the number that best describes your own opinion of what is generally true for you.

	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
16) I believe some of my thoughts are abnormal or bad and shouldn't think that way	1	2	3	4	5
17) I notice how foods and drinks affect my thoughts, bodily sensations, and emotions	1	2	3	4	5
18) I have trouble thinking of the right words to express how I feel about things	1	2	3	4	5
19) When I do things, I get totally wrapped up in them and don't think about anything else	1	2	3	4	5
20) I make judgments about whether my thoughts are good or bad	1	2	3	4	5
21) I pay attention to sensations, such as the wind in my hair or sun on my face	1	2	3	4	5
22) When I have a sensation in my body, it's difficult for me to describe it because I can't find the right words	1	2	3	4	5
23) I don't pay attention to what I'm doing because I'm daydreaming, worrying, or otherwise distracted	1	2	3	4	5
24) I tend to make judgments about how worthwhile or worthless my experiences are	1	2	3	4	5
25) I pay attention to sounds, such as clocks ticking, birds chirping, or cars passing	1	2	3	4	5
26) Even when I'm feeling terribly upset, I can find a way to put it into words	1	2	3	4	5
27) When I'm doing chores, such as cleaning or laundry, I tend to daydream or think of other things	1	2	3	4	5
28) I tell myself that I shouldn't be thinking the way I'm thinking	1	2	3	4	5
29) I notice the smells and aromas of things	1	2	3	4	5
30) I intentionally stay aware of my feelings	1	2	3	4	5

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Four: The KIMS



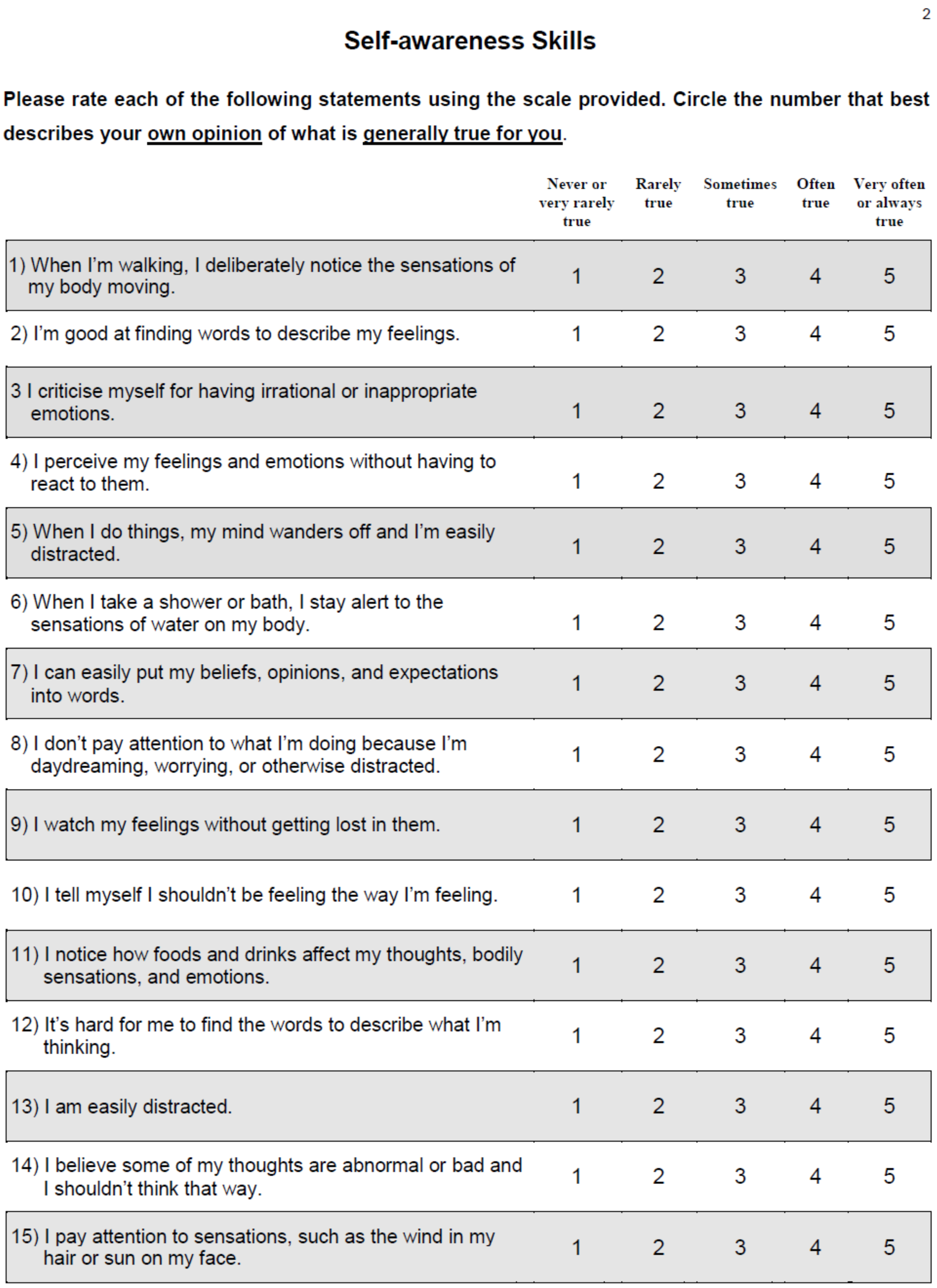
5

Please rate each of the following statements using the scale provided. Circle the number that best describes your <u>own opinion</u> of what is <u>generally true for you</u>.

	Never or very rarely true	**Rarely true**	**Sometimes true**	**Often true**	**Very often or always true**
31) I tend to do several things at once rather than focusing on one thing at a time	1	2	3	4	5
32) I think some of my emotions are bad or inappropriate and I shouldn't feel them	1	2	3	4	5
33) I notice visual elements in art or nature, such as colors, shapes, textures, or patterns of light and shadow	1	2	3	4	5
34) My natural tendency is to put my experiences into words	1	2	3	4	5
35) When I'm working on something, part of my mind is occupied with other topics, such as what I'll be doing later, or things I'd rather be doing	1	2	3	4	5
36) I disapprove of myself when I have irrational ideas	1	2	3	4	5
37) I pay attention to how my emotions affect my thoughts and behavior	1	2	3	4	5
38) I get completely absorbed in what I'm doing, so that all my attention is focused on it	1	2	3	4	5
39) I notice when my moods begin to change	1	2	3	4	5

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Five: The FFMQ



2

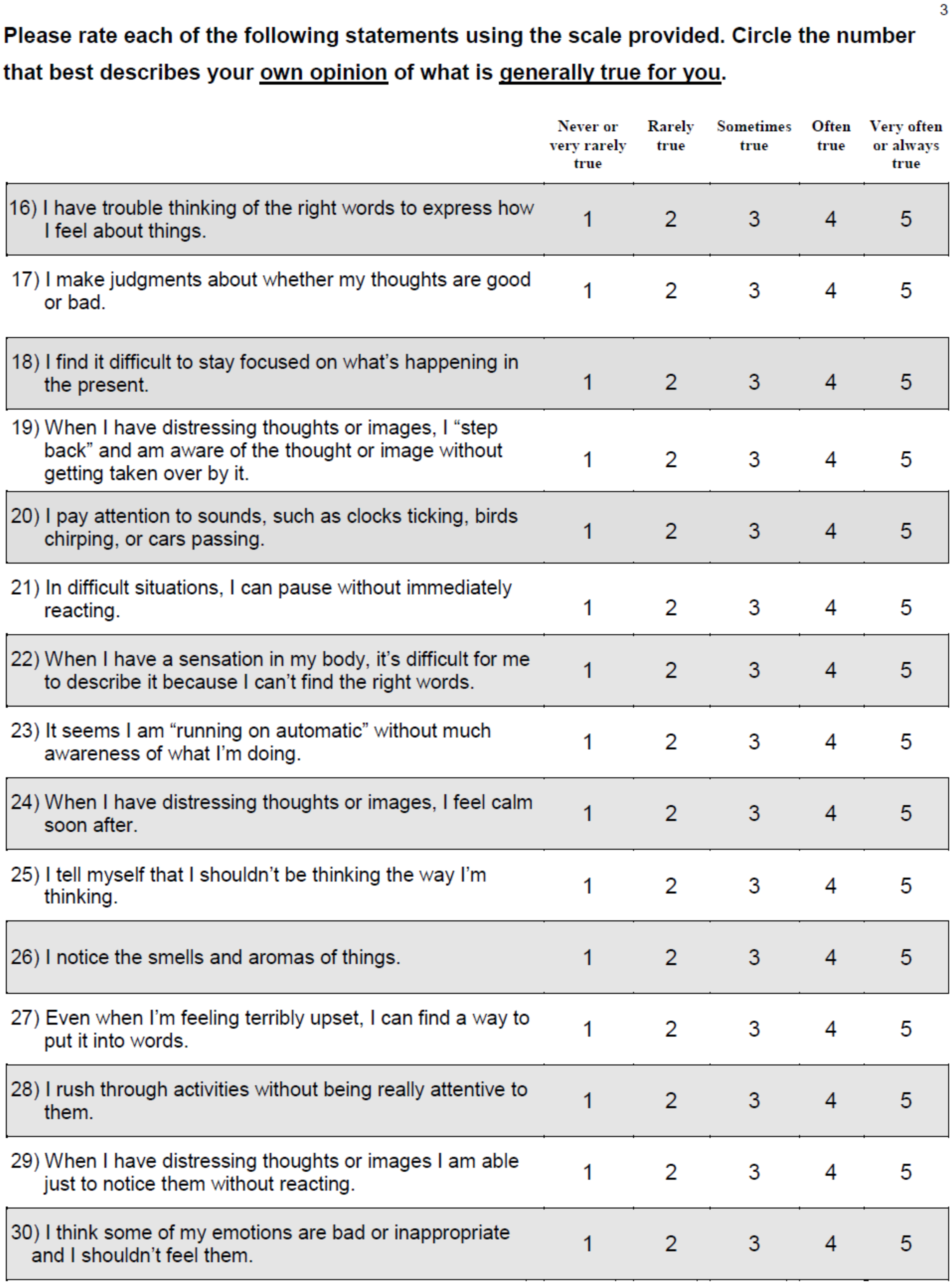
Self-awareness Skills

Please rate each of the following statements using the scale provided. Circle the number that best describes your <u>own opinion</u> of what is <u>generally true for you</u>.

	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
1) When I'm walking, I deliberately notice the sensations of my body moving.	1	2	3	4	5
2) I'm good at finding words to describe my feelings.	1	2	3	4	5
3 I criticise myself for having irrational or inappropriate emotions.	1	2	3	4	5
4) I perceive my feelings and emotions without having to react to them.	1	2	3	4	5
5) When I do things, my mind wanders off and I'm easily distracted.	1	2	3	4	5
6) When I take a shower or bath, I stay alert to the sensations of water on my body.	1	2	3	4	5
7) I can easily put my beliefs, opinions, and expectations into words.	1	2	3	4	5
8) I don't pay attention to what I'm doing because I'm daydreaming, worrying, or otherwise distracted.	1	2	3	4	5
9) I watch my feelings without getting lost in them.	1	2	3	4	5
10) I tell myself I shouldn't be feeling the way I'm feeling.	1	2	3	4	5
11) I notice how foods and drinks affect my thoughts, bodily sensations, and emotions.	1	2	3	4	5
12) It's hard for me to find the words to describe what I'm thinking.	1	2	3	4	5
13) I am easily distracted.	1	2	3	4	5
14) I believe some of my thoughts are abnormal or bad and I shouldn't think that way.	1	2	3	4	5
15) I pay attention to sensations, such as the wind in my hair or sun on my face.	1	2	3	4	5

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Five: The FFMQ



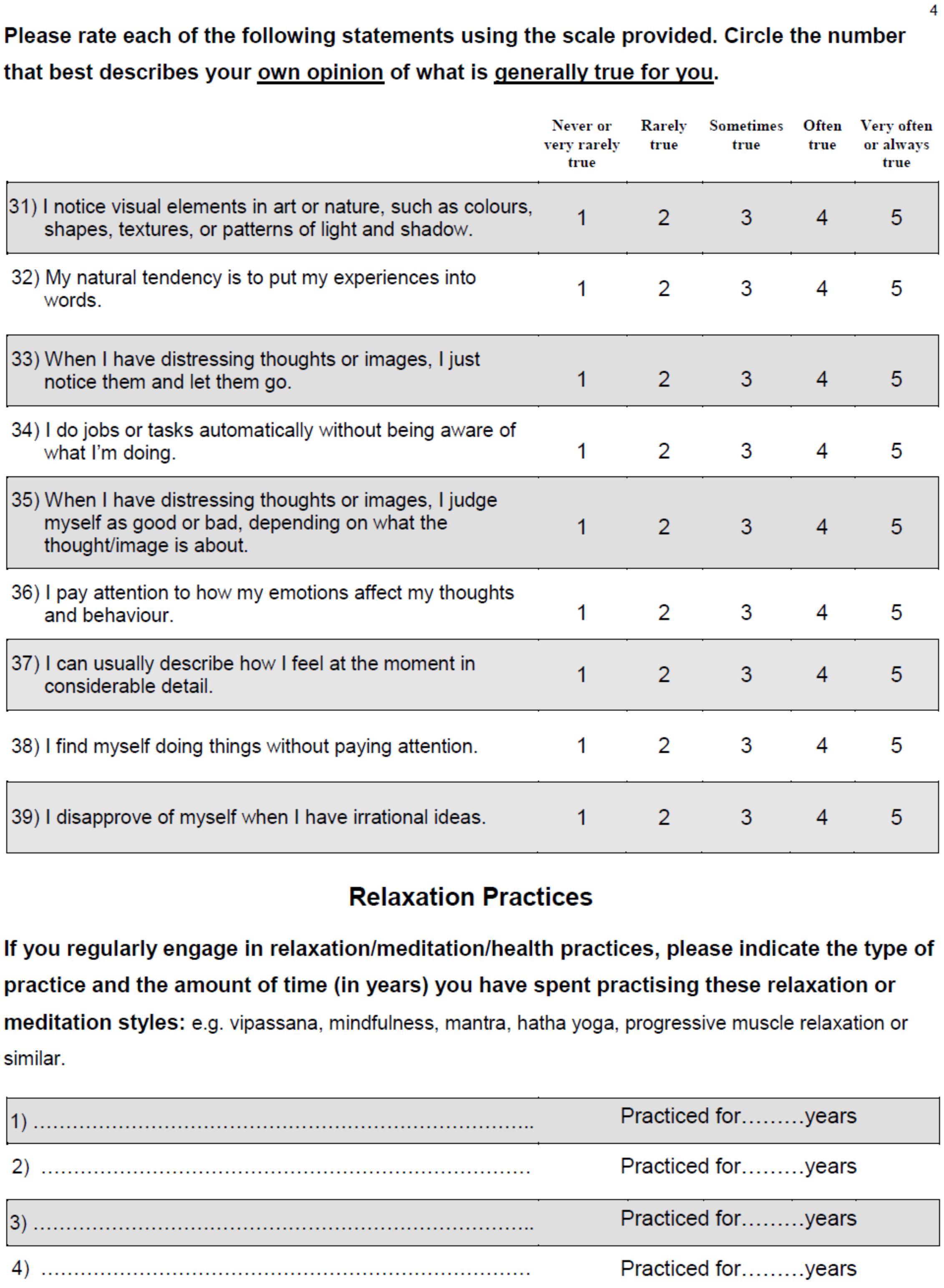
3

Please rate each of the following statements using the scale provided. Circle the number that best describes your own opinion of what is generally true for you.

	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
16) I have trouble thinking of the right words to express how I feel about things.	1	2	3	4	5
17) I make judgments about whether my thoughts are good or bad.	1	2	3	4	5
18) I find it difficult to stay focused on what's happening in the present.	1	2	3	4	5
19) When I have distressing thoughts or images, I "step back" and am aware of the thought or image without getting taken over by it.	1	2	3	4	5
20) I pay attention to sounds, such as clocks ticking, birds chirping, or cars passing.	1	2	3	4	5
21) In difficult situations, I can pause without immediately reacting.	1	2	3	4	5
22) When I have a sensation in my body, it's difficult for me to describe it because I can't find the right words.	1	2	3	4	5
23) It seems I am "running on automatic" without much awareness of what I'm doing.	1	2	3	4	5
24) When I have distressing thoughts or images, I feel calm soon after.	1	2	3	4	5
25) I tell myself that I shouldn't be thinking the way I'm thinking.	1	2	3	4	5
26) I notice the smells and aromas of things.	1	2	3	4	5
27) Even when I'm feeling terribly upset, I can find a way to put it into words.	1	2	3	4	5
28) I rush through activities without being really attentive to them.	1	2	3	4	5
29) When I have distressing thoughts or images I am able just to notice them without reacting.	1	2	3	4	5
30) I think some of my emotions are bad or inappropriate and I shouldn't feel them.	1	2	3	4	5

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Five: The FFMQ



4

Please rate each of the following statements using the scale provided. Circle the number that best describes your own opinion of what is generally true for you.

	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
31) I notice visual elements in art or nature, such as colours, shapes, textures, or patterns of light and shadow.	1	2	3	4	5
32) My natural tendency is to put my experiences into words.	1	2	3	4	5
33) When I have distressing thoughts or images, I just notice them and let them go.	1	2	3	4	5
34) I do jobs or tasks automatically without being aware of what I'm doing.	1	2	3	4	5
35) When I have distressing thoughts or images, I judge myself as good or bad, depending on what the thought/image is about.	1	2	3	4	5
36) I pay attention to how my emotions affect my thoughts and behaviour.	1	2	3	4	5
37) I can usually describe how I feel at the moment in considerable detail.	1	2	3	4	5
38) I find myself doing things without paying attention.	1	2	3	4	5
39) I disapprove of myself when I have irrational ideas.	1	2	3	4	5

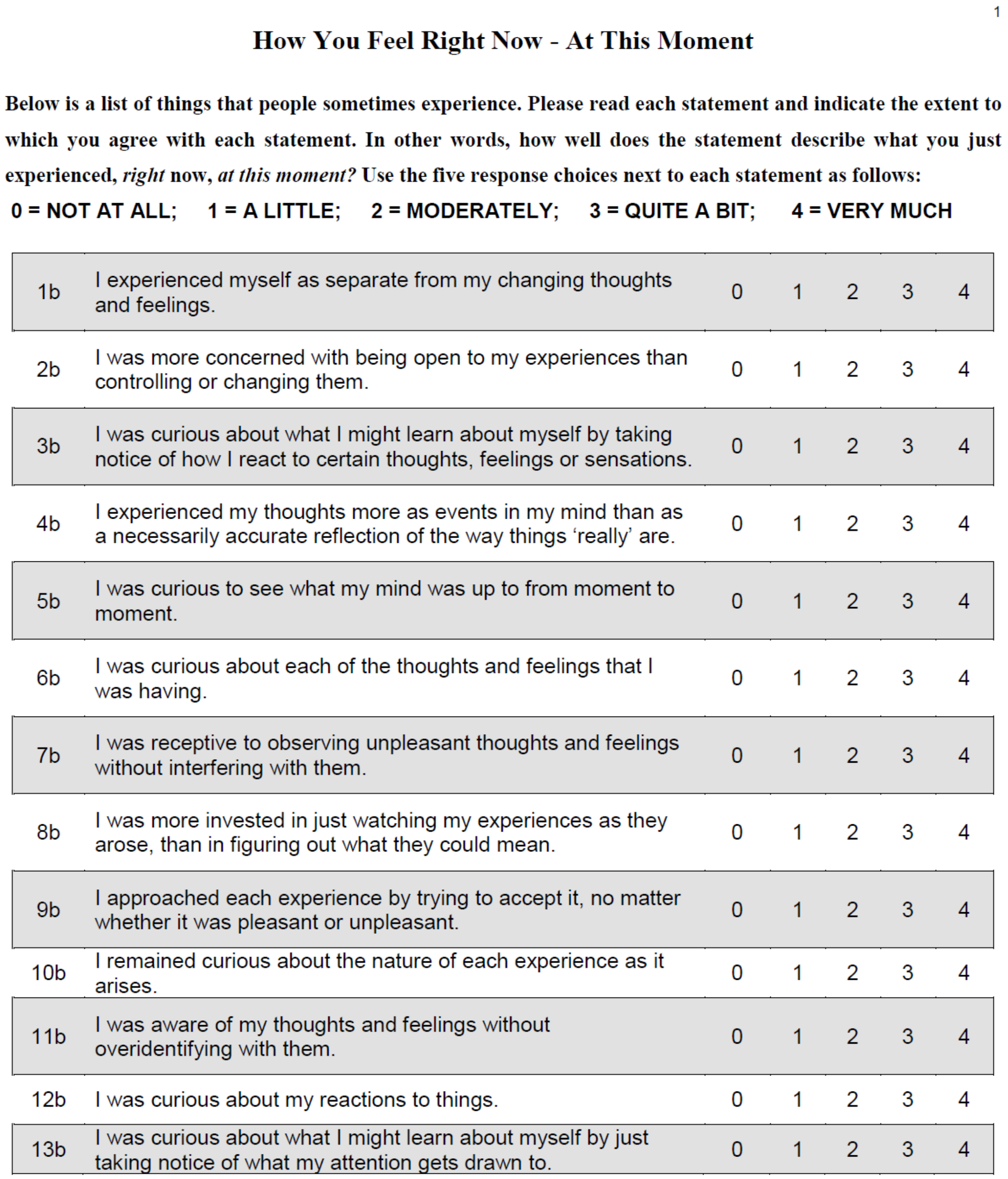
Relaxation Practices

If you regularly engage in relaxation/meditation/health practices, please indicate the type of practice and the amount of time (in years) you have spent practising these relaxation or meditation styles: e.g. vipassana, mindfulness, mantra, hatha yoga, progressive muscle relaxation or similar.

1) …………………………………………………………………..	Practiced for………years
2) ……………………………………………………………………	Practiced for………years
3) …………………………………………………………………..	Practiced for………years
4) ……………………………………………………………………	Practiced for………years

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)

Chapter Ten: The TMS



1

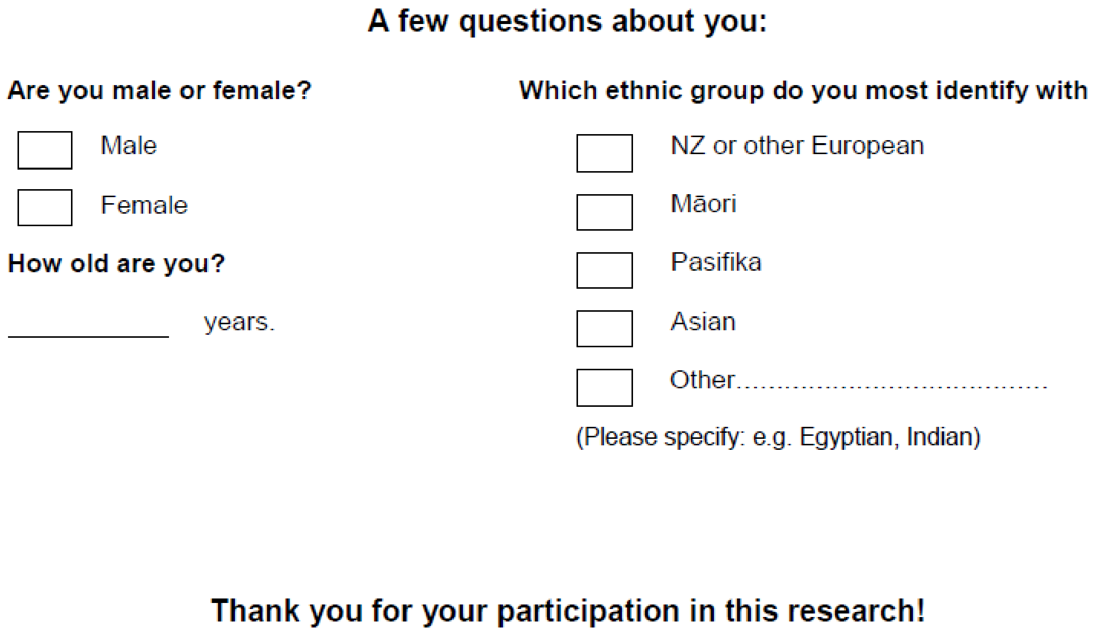
How You Feel Right Now - At This Moment

Below is a list of things that people sometimes experience. Please read each statement and indicate the extent to which you agree with each statement. In other words, how well does the statement describe what you just experienced, *right* now, *at this moment?* Use the five response choices next to each statement as follows:

0 = NOT AT ALL; 1 = A LITTLE; 2 = MODERATELY; 3 = QUITE A BIT; 4 = VERY MUCH

1b	I experienced myself as separate from my changing thoughts and feelings.	0	1	2	3	4
2b	I was more concerned with being open to my experiences than controlling or changing them.	0	1	2	3	4
3b	I was curious about what I might learn about myself by taking notice of how I react to certain thoughts, feelings or sensations.	0	1	2	3	4
4b	I experienced my thoughts more as events in my mind than as a necessarily accurate reflection of the way things ‘really’ are.	0	1	2	3	4
5b	I was curious to see what my mind was up to from moment to moment.	0	1	2	3	4
6b	I was curious about each of the thoughts and feelings that I was having.	0	1	2	3	4
7b	I was receptive to observing unpleasant thoughts and feelings without interfering with them.	0	1	2	3	4
8b	I was more invested in just watching my experiences as they arose, than in figuring out what they could mean.	0	1	2	3	4
9b	I approached each experience by trying to accept it, no matter whether it was pleasant or unpleasant.	0	1	2	3	4
10b	I remained curious about the nature of each experience as it arises.	0	1	2	3	4
11b	I was aware of my thoughts and feelings without overidentifying with them.	0	1	2	3	4
12b	I was curious about my reactions to things.	0	1	2	3	4
13b	I was curious about what I might learn about myself by just taking notice of what my attention gets drawn to.	0	1	2	3	4

Appendix C1: Participant Questionnaire (Chapters Three-Five, and Ten)



A few questions about you:

Are you male or female?

□ Male

□ Female

How old are you?

____________ years.

Which ethnic group do you most identify with

□ NZ or other European

□ Māori

□ Pasifika

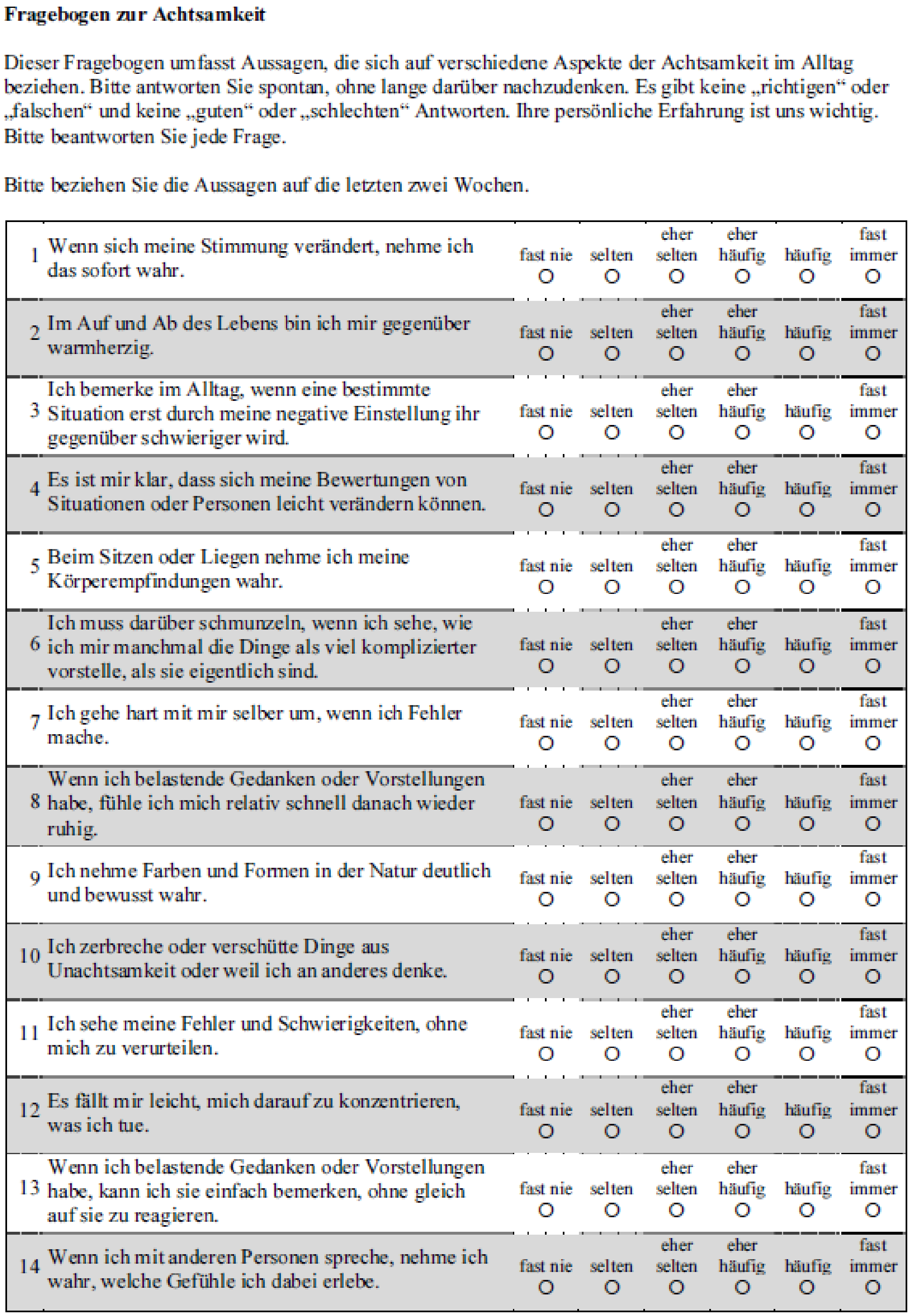
□ Asian

□ Other……………………………………

(Please specify: e.g. Egyptian, Indian)

Thank you for your participation in this research!

Appendix C2: Participant Questionnaire Chapter Six: The CHIME



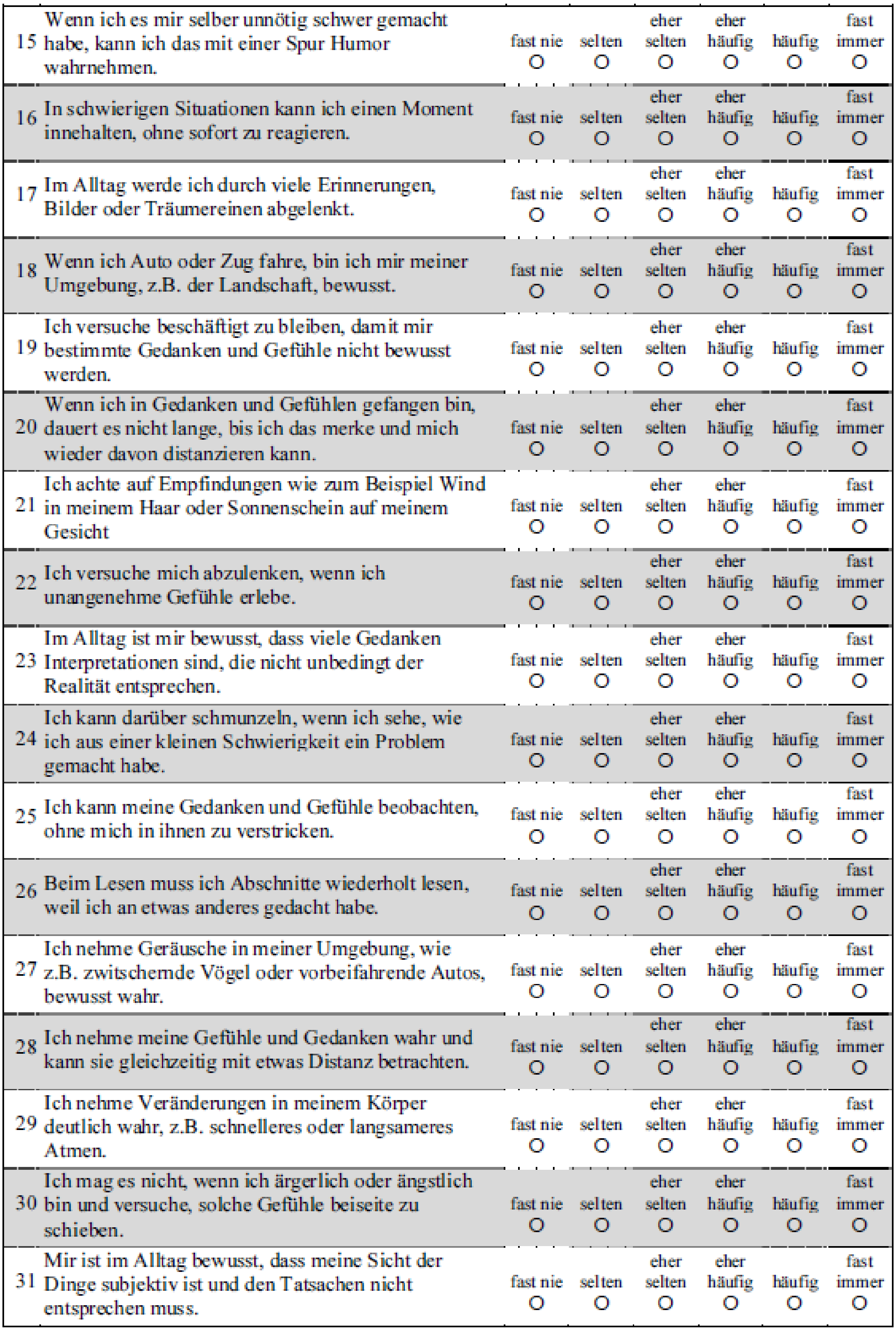
Fragebogen zur Achtsamkeit

Dieser Fragebogen umfasst Aussagen, die sich auf verschiedene Aspekte der Achtsamkeit im Alltag beziehen. Bitte antworten Sie spontan, ohne lange darüber nachzudenken. Es gibt keine „richtigen“ oder „falschen“ und keine „guten“ oder „schlechten“ Antworten. Ihre persönliche Erfahrung ist uns wichtig. Bitte beantworten Sie jede Frage.

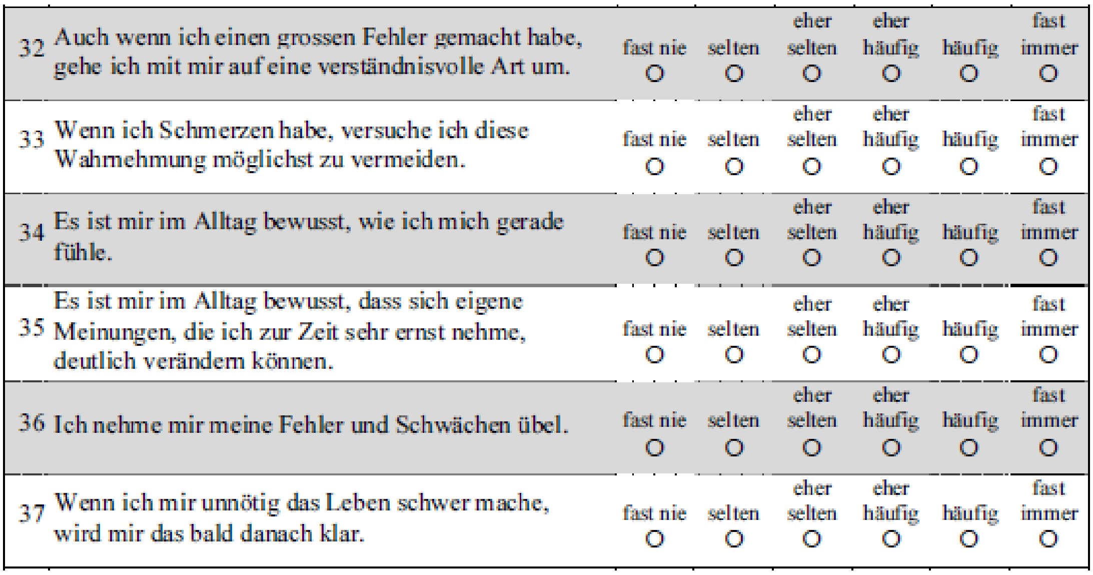
Bitte beziehen Sie die Aussagen auf die letzten zwei Wochen.

1	Wenn sich meine Stimmung verändert, nehme ich das sofort wahr.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
2	Im Auf und Ab des Lebens bin ich mir gegenüber warmherzig.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
3	Ich bemerke im Alltag, wenn eine bestimmte Situation erst durch meine negative Einstellung ihr gegenüber schwieriger wird.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
4	Es ist mir klar, dass sich meine Bewertungen von Situationen oder Personen leicht verändern können.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
5	Beim Sitzen oder Liegen nehme ich meine Körperempfindungen wahr.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
6	Ich muss darüber schmunzeln, wenn ich sehe, wie ich mir manchmal die Dinge als viel komplizierter vorstelle, als sie eigentlich sind.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
7	Ich gehe hart mit mir selber um, wenn ich Fehler mache.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
8	Wenn ich belastende Gedanken oder Vorstellungen habe, fühle ich mich relativ schnell danach wieder ruhig.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
9	Ich nehme Farben und Formen in der Natur deutlich und bewusst wahr.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
10	Ich zerbreche oder verschütte Dinge aus Unachtsamkeit oder weil ich an anderes denke.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
11	Ich sehe meine Fehler und Schwierigkeiten, ohne mich zu verurteilen.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
12	Es fällt mir leicht, mich darauf zu konzentrieren, was ich tue.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
13	Wenn ich belastende Gedanken oder Vorstellungen habe, kann ich sie einfach bemerken, ohne gleich auf sie zu reagieren.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
14	Wenn ich mit anderen Personen spreche, nehme ich wahr, welche Gefühle ich dabei erlebe.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○

Appendix C2: Participant Questionnaire Chapter Six: The CHIME

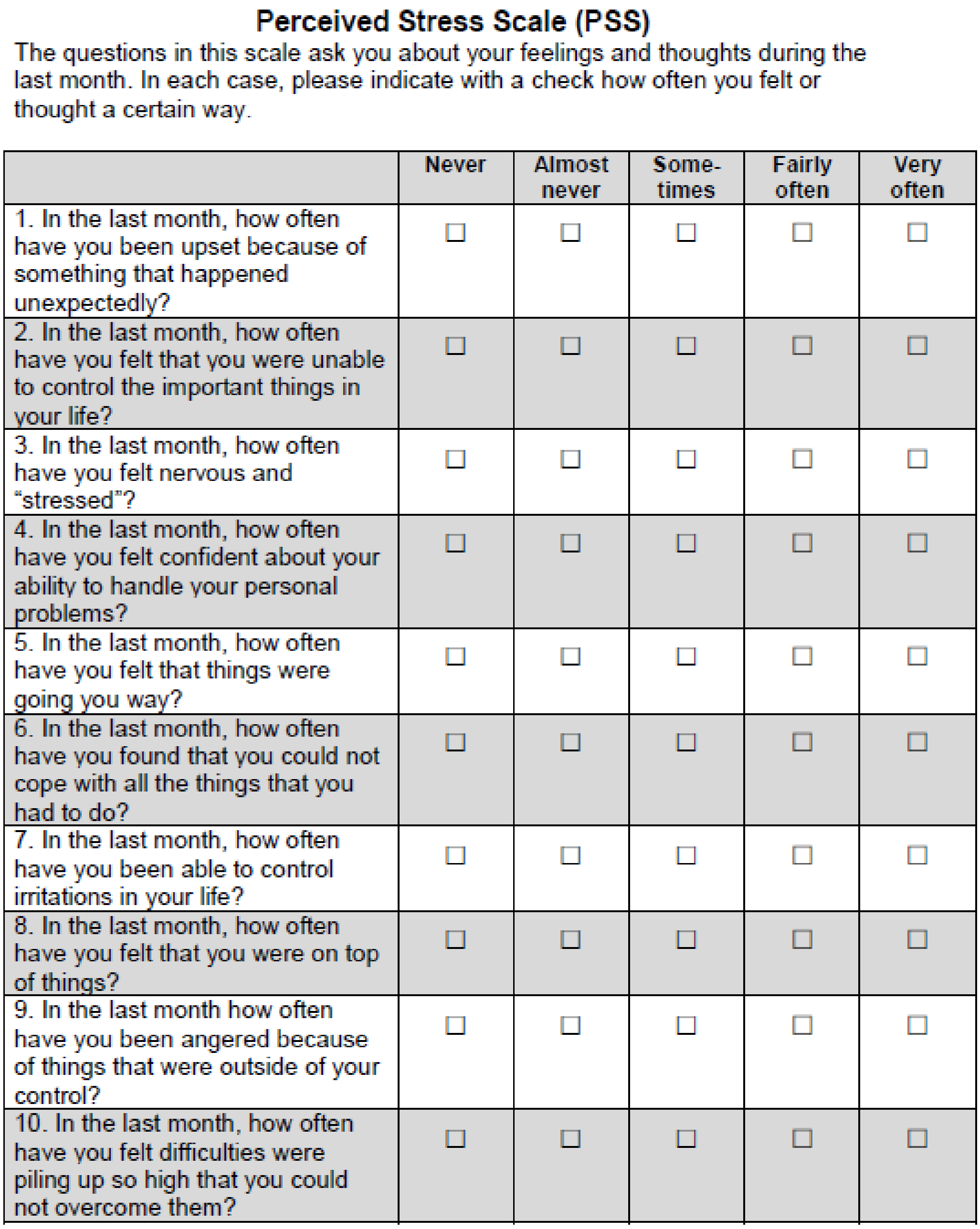


15	Wenn ich es mir selber unnötig schwer gemacht habe, kann ich das mit einer Spur Humor wahrnehmen.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
16	In schwierigen Situationen kann ich einen Moment innehalten, ohne sofort zu reagieren.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
17	Im Alltag werde ich durch viele Erinnerungen, Bilder oder Träumereinen abgelenkt.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
18	Wenn ich Auto oder Zug fahre, bin ich mir meiner Umgebung, z.B. der Landschaft, bewusst.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
19	Ich versuche beschäftigt zu bleiben, damit mir bestimmte Gedanken und Gefühle nicht bewusst werden.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
20	Wenn ich in Gedanken und Gefühlen gefangen bin, dauert es nicht lange, bis ich das merke und mich wieder davon distanzieren kann.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
21	Ich achte auf Empfindungen wie zum Beispiel Wind in meinem Haar oder Sonnenschein auf meinem Gesicht	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
22	Ich versuche mich abzulenken, wenn ich unangenehme Gefühle erlebe.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
23	Im Alltag ist mir bewusst, dass viele Gedanken Interpretationen sind, die nicht unbedingt der Realität entsprechen.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
24	Ich kann darüber schmunzeln, wenn ich sehe, wie ich aus einer kleinen Schwierigkeit ein Problem gemacht habe.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
25	Ich kann meine Gedanken und Gefühle beobachten, ohne mich in ihnen zu verstricken.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
26	Beim Lesen muss ich Abschnitte wiederholt lesen, weil ich an etwas anderes gedacht habe.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
27	Ich nehme Geräusche in meiner Umgebung, wie z.B. zwitschernde Vögel oder vorbeifahrende Autos, bewusst wahr.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
28	Ich nehme meine Gefühle und Gedanken wahr und kann sie gleichzeitig mit etwas Distanz betrachten.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
29	Ich nehme Veränderungen in meinem Körper deutlich wahr, z.B. schnelleres oder langsameres Atmen.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
30	Ich mag es nicht, wenn ich ärgerlich oder ängstlich bin und versuche, solche Gefühle beiseite zu schieben.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
31	Mir ist im Alltag bewusst, dass meine Sicht der Dinge subjektiv ist und den Tatsachen nicht entsprechen muss.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○



32	Auch wenn ich einen grossen Fehler gemacht habe, gehe ich mit mir auf eine verständnisvolle Art um.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
33	Wenn ich Schmerzen habe, versuche ich diese Wahrnehmung möglichst zu vermeiden.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
34	Es ist mir im Alltag bewusst, wie ich mich gerade fühle.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
35	Es ist mir im Alltag bewusst, dass sich eigene Meinungen, die ich zur Zeit sehr ernst nehme, deutlich verändern können.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
36	Ich nehme mir meine Fehler und Schwächen übel.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○
37	Wenn ich mir unnötig das Leben schwer mache, wird mir das bald danach klar.	fast nie ○	selten ○	eher selten ○	eher häufig ○	häufig ○	fast immer ○

Appendix C3: Participant Questionnaire Chapter Seven: The PSS

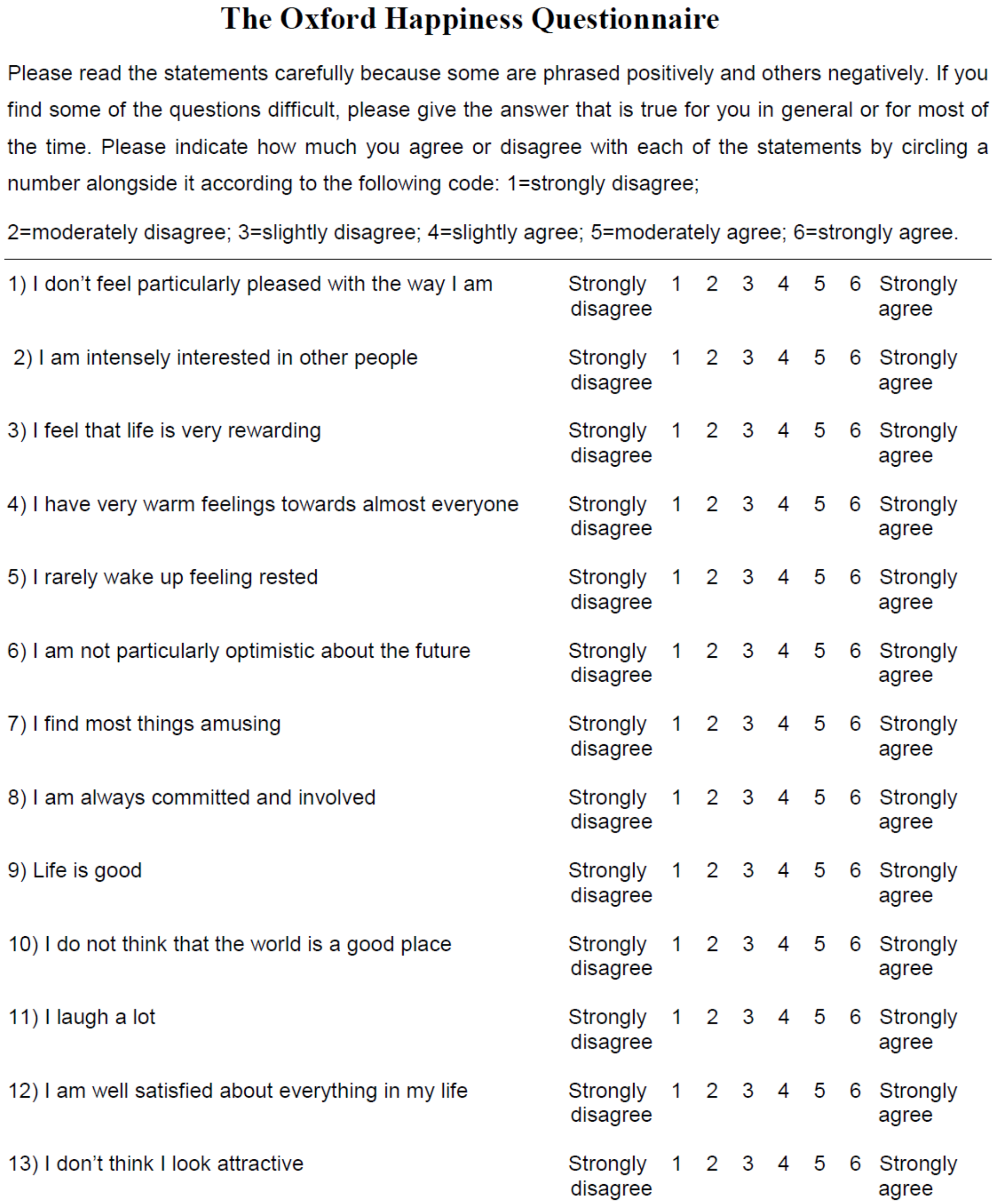


Perceived Stress Scale (PSS)

The questions in this scale ask you about your feelings and thoughts during the last month. In each case, please indicate with a check how often you felt or thought a certain way.

	Never	Almost never	Some-times	Fairly often	Very often
1. In the last month, how often have you been upset because of something that happened unexpectedly?	□	□	□	□	□
2. In the last month, how often have you felt that you were unable to control the important things in your life?	□	□	□	□	□
3. In the last month, how often have you felt nervous and "stressed"?	□	□	□	□	□
4. In the last month, how often have you felt confident about your ability to handle your personal problems?	□	□	□	□	□
5. In the last month, how often have you felt that things were going you way?	□	□	□	□	□
6. In the last month, how often have you found that you could not cope with all the things that you had to do?	□	□	□	□	□
7. In the last month, how often have you been able to control irritations in your life?	□	□	□	□	□
8. In the last month, how often have you felt that you were on top of things?	□	□	□	□	□
9. In the last month how often have you been angered because of things that were outside of your control?	□	□	□	□	□
10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?	□	□	□	□	□

Appendix C4: Participant Questionnaire Chapter Eight: The OHQ



The Oxford Happiness Questionnaire

Please read the statements carefully because some are phrased positively and others negatively. If you find some of the questions difficult, please give the answer that is true for you in general or for most of the time. Please indicate how much you agree or disagree with each of the statements by circling a number alongside it according to the following code: 1=strongly disagree;

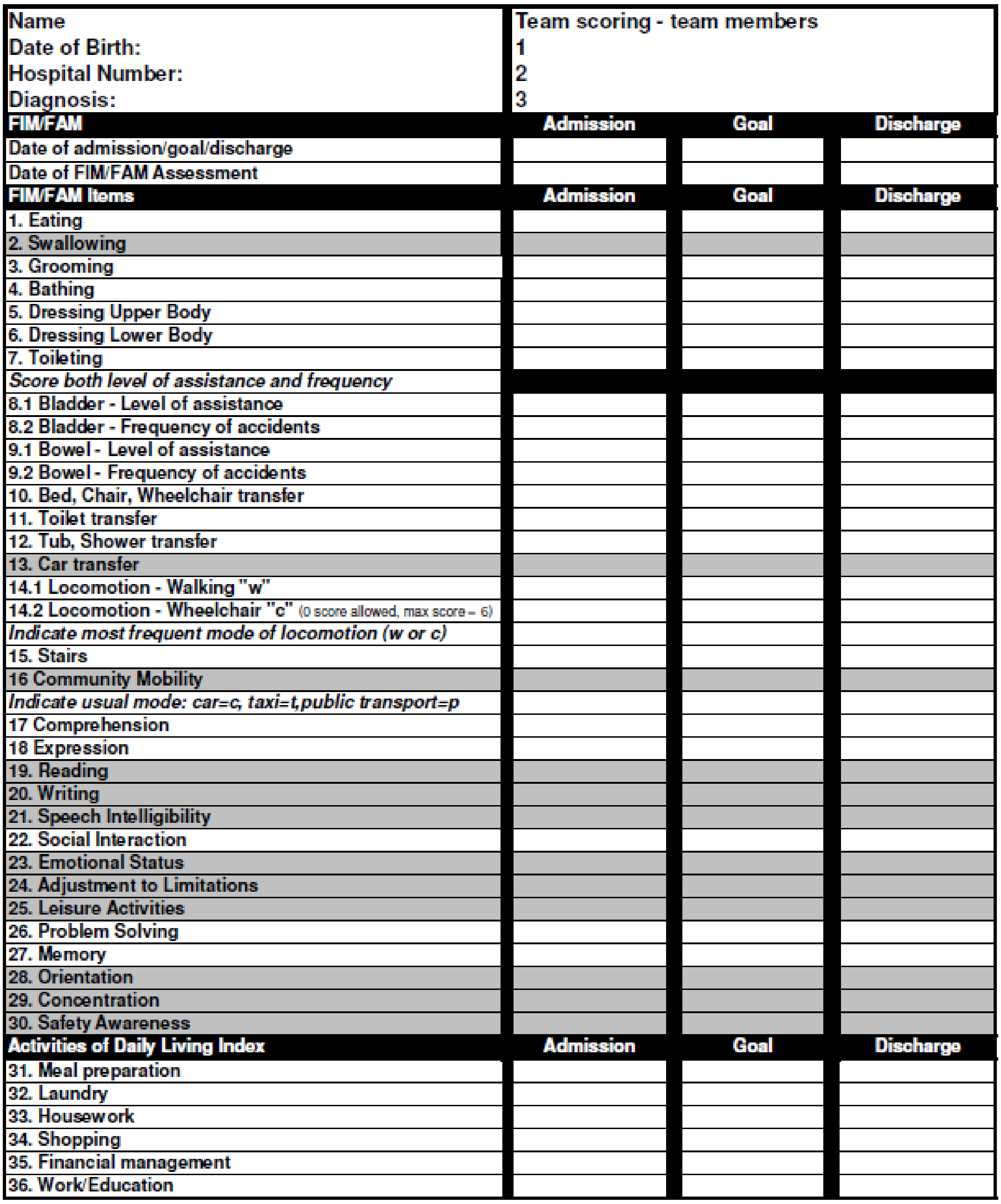
2=moderately disagree; 3=slightly disagree; 4=slightly agree; 5=moderately agree; 6=strongly agree.

1) I don't feel particularly pleased with the way I am	Strongly disagree	1	2	3	4	5	6	Strongly agree
2) I am intensely interested in other people	Strongly disagree	1	2	3	4	5	6	Strongly agree
3) I feel that life is very rewarding	Strongly disagree	1	2	3	4	5	6	Strongly agree
4) I have very warm feelings towards almost everyone	Strongly disagree	1	2	3	4	5	6	Strongly agree
5) I rarely wake up feeling rested	Strongly disagree	1	2	3	4	5	6	Strongly agree
6) I am not particularly optimistic about the future	Strongly disagree	1	2	3	4	5	6	Strongly agree
7) I find most things amusing	Strongly disagree	1	2	3	4	5	6	Strongly agree
8) I am always committed and involved	Strongly disagree	1	2	3	4	5	6	Strongly agree
9) Life is good	Strongly disagree	1	2	3	4	5	6	Strongly agree
10) I do not think that the world is a good place	Strongly disagree	1	2	3	4	5	6	Strongly agree
11) I laugh a lot	Strongly disagree	1	2	3	4	5	6	Strongly agree
12) I am well satisfied about everything in my life	Strongly disagree	1	2	3	4	5	6	Strongly agree
13) I don't think I look attractive	Strongly disagree	1	2	3	4	5	6	Strongly agree

Appendix C4: Participant Questionnaire Chapter Eight: The OHQ

14) There is a gap between what I would like to do and what I have done	Strongly disagree	1	2	3	4	5	6	Strongly agree
15) I am very happy	Strongly disagree	1	2	3	4	5	6	Strongly agree
16) I find beauty in some things	Strongly disagree	1	2	3	4	5	6	Strongly agree
17) I always have a cheerful effect on others	Strongly disagree	1	2	3	4	5	6	Strongly agree
18) I can fit in everything I want to	Strongly disagree	1	2	3	4	5	6	Strongly agree
19) I feel that I am not especially in control of my life	Strongly disagree	1	2	3	4	5	6	Strongly agree
20) I feel able to take anything on	Strongly disagree	1	2	3	4	5	6	Strongly agree
21) I feel fully mentally alert	Strongly disagree	1	2	3	4	5	6	Strongly agree
22) I often experience joy and elation	Strongly disagree	1	2	3	4	5	6	Strongly agree
23) I do not find it easy to make decisions	Strongly disagree	1	2	3	4	5	6	Strongly agree
24) I do not have a particular sense of meaning and purpose in my life	Strongly disagree	1	2	3	4	5	6	Strongly agree
25) I feel I have a great deal of energy	Strongly disagree	1	2	3	4	5	6	Strongly agree
26) I usually have a good influence on events	Strongly disagree	1	2	3	4	5	6	Strongly agree
27) I do not have fun with other people	Strongly disagree	1	2	3	4	5	6	Strongly agree
28) I don't feel particularly healthy	Strongly disagree	1	2	3	4	5	6	Strongly agree
29) I do not have particularly happy memories of the past	Strongly disagree	1	2	3	4	5	6	Strongly agree

Appendix C5: Participant Questionnaire Chapter Nine: The UK FIM+FAM



Name	Team scoring - team members		
Date of Birth:	1		
Hospital Number:	2		
Diagnosis:	3		
FIM/FAM	**Admission**	**Goal**	**Discharge**
Date of admission/goal/discharge			
Date of FIM/FAM Assessment			
FIM/FAM Items	**Admission**	**Goal**	**Discharge**
1. Eating			
2. Swallowing			
3. Grooming			
4. Bathing			
5. Dressing Upper Body			
6. Dressing Lower Body			
7. Toileting			
Score both level of assistance and frequency			
8.1 Bladder - Level of assistance			
8.2 Bladder - Frequency of accidents			
9.1 Bowel - Level of assistance			
9.2 Bowel - Frequency of accidents			
10. Bed, Chair, Wheelchair transfer			
11. Toilet transfer			
12. Tub, Shower transfer			
13. Car transfer			
14.1 Locomotion - Walking "w"			
14.2 Locomotion - Wheelchair "c" (0 score allowed, max score = 6)			
Indicate most frequent mode of locomotion (w or c)			
15. Stairs			
16 Community Mobility			
Indicate usual mode: car=c, taxi=t,public transport=p			
17 Comprehension			
18 Expression			
19. Reading			
20. Writing			
21. Speech Intelligibility			
22. Social Interaction			
23. Emotional Status			
24. Adjustment to Limitations			
25. Leisure Activities			
26. Problem Solving			
27. Memory			
28. Orientation			
29. Concentration			
30. Safety Awareness			
Activities of Daily Living Index	**Admission**	**Goal**	**Discharge**
31. Meal preparation			
32. Laundry			
33. Housework			
34. Shopping			
35. Financial management			
36. Work/Education			

Note: The original research articals are not included in this version of my thesis because these works are available online through various databases and the relevant references are provided in this thesis document.