Rigorous methods for the analysis, reporting and evaluation of ESM style data

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2016

Lesley-Anne Carter

School of Health Sciences

TABLE OF CONTENTS

List of tables 5

List of figures 7

Abstract of thesis 10

Declaration 11

Acknowledgements 12

List of abbreviations 13

1 Introduction and motivation for research 14

1.1 Experience sampling methodology 14

1.1.1 Benefits of methodology 16

1.1.2 ESM and technology 18

1.2 Data structure 18

1.3 Aims and objectives of this thesis 20

2 Introduction to multilevel models and motivating examples 22

2.1 An introduction to multilevel models 22

2.1.1 Random intercept models 22

2.1.2 Random coefficient models 25

2.1.3 Centring 27

2.1.4 Variation in a three-level model 28

2.1.5 Level 1 covariance structures 33

2.2 Motivating example: Recovery in psychosis (Bentall et al.) 36

2.2.1 Statistical models 37

2.2.2 Results of recovery data analysis 39

2.3 Motivating example: Bipolar disorder and cannabis use (Tyler et al. 2015) 44

3 Systematic Review 46

3.1 Aims of review 46

3.2 Design of review 46

3.3 Results of systematic review ... 47

3.3.1 Overview of studies ... 50

3.3.2 Sample size and adherence to protocol ... 52

3.3.3 Factor analysis and Cronbach's alpha ... 53

3.3.4 Types of research questions ... 53

3.3.5 Analysis method ... 55

3.4 Discussion of systematic review ... 57

4 Missing data in ESM research ... 59

4.1 Systematic review: Adherence to protocol and missing data ... 59

4.1.1 Examples of nonresponse in ESM data ... 61

4.2 Nonresponse in the recovery example ... 63

4.2.1 Complete nonresponse ... 64

4.2.2 Item nonresponse ... 64

4.2.3 Moment nonresponse ... 67

4.3 Statistical analysis with partially observed data ... 69

4.3.1 Missing data mechanisms ... 69

4.3.2 Approaches to analysis with missing data ... 72

4.4 Predictors of nonresponse in the recovery data ... 75

4.5 Summary ... 79

5 Momentary level variation and time trends ... 80

5.1 Introduction from systematic review ... 80

5.1.1 Papers identified in systematic review studying variation ... 80

5.2 Alternative approaches to investigating variation: Time trends ... 84

5.2.1 Two-level time trend models ... 85

5.2.2 Three-level paradigm for ESM data ... 88

5.3 Applications to data ... 97

5.3.1 Variance components model ... 97

5.3.2 Time trends in recovery data ... 99

5.4 Summary 112
6 Predicting momentary change 114
6.1 Introduction 114
6.1.1 Change models identified in the systematic review 115
6.2 Issues surrounding change models: Motivating example 120
6.2.1 Endogeneity 122
6.2.2 Solution to endogeneity: The first-difference method 123
6.2.3 Solution to endogeneity: The initial conditions problem 124
6.3 Recovery data analysis 130
6.3.1 First-difference approach 132
6.3.2 Initial conditions model 136
6.4 Summary 140
7 Power and sample size 141
7.1 Power in current ESM research 141
7.2 Considerations for power and sample size calculations 142
7.3 Power for longitudinal data 144
7.3.1 Power for three-level data 146
7.3.2 Empirical power 149
7.3.3 Available software 150
7.4 Determining sample size for ESM studies 151
7.4.1 Simulation protocol 152
7.4.2 Results of simulations 157
7.5 Summary 173
8 Discussion 175
8.1 Design - Power and sample size 175
8.1.1 Implications of missing data 178
8.1.2 Adaptation to more complex scenarios 179
8.1.3 Limitations of work on sample size and power 180

8.2 Reporting - Missing data 181

8.2.1 Predictors of nonresponse and missing data mechanisms 183

8.3 Analysis - Change models 186

8.3.1 Limitations of the initial conditions method 187

8.3.2 Future work on change models 189

8.4 Investigation – Time trends using random slope models 190

8.5 Applications outside of ESM 193

8.6 Summary 193

9 References 195

Appendix 1: Motivating examples 203

Appendix 2: Systematic review 205

Appendix 3: Missing data 208

Appendix 4: Predicting momentary change 210

Appendix 5: Power and sample size 211

Final word count: 53,840

LIST OF TABLES

Table 2:1 Complex variation model comparing momentary recovery between baseline groups 39
Table 2:2 Complex variation model comparing momentary recovery for continuous QPR scale 39
Table 2:3 Concurrent associations of Right now variables and recovery 40
Table 2:4 Lagged analysis between the Right now variables and recovery 42
Table 2:5 Change models for Right now variables and recovery 43
Table 3:1 Overview of published 2012 studies 51
Table 3:2 Research questions identified in the review 54
Table 3:3 Primary analysis method identified in the review 56
Table 4:1 Recovery data - Percentage missing data conditional on momentary response and that at least one item completed within the whole sampling period: IDs 37 and 51 dropped. 65
Table 4:2 Moment nonresponse definition comparison for recovery data 68
Table 4:3 Concurrent predictors of missing recovery item 76
Table 4:4 Lagged predictor of missing diary 77
Table 4:5 Table of observed and missing consecutive diaries displaying the relative proportions in response categories when the predictor is observed 78
Table 4:6 Within-day quadratic trend in event nonresponse 78
Table 4:7 Between-day quadratic trend in event nonresponse 79
Table 5:1 Three-level fixed and random time effects array 94
Table 5:2 Estimates of a two- and three-level variance components model for recovery ... 98
Table 5:3 Time trend models to be applied to the recovery data 99
Table 5:4 Fixed and random effect estimates of time trend models fitted to the recovery data 101
Table 5:5 Full and reduced quadratic random intercept models fitted to the recovery data 102
Table 5:6 Fixed and random effect estimates of reduced non-linear time trend models fitted to the recovery data – linear terms only in the random effects 104
Table 5:7 Within- and between-day random slope group difference models fitted to the recovery data with quadratic time trends and complex level 1 variance 107
Table 5:8 Within- and between-day random slope group difference models fitted to the recovery data with quadratic time trends and complex, time dependent level 1 variance 110

Table 6:1 Lagged correlation pairs for recovery data 121

Table 6:2 Naïve lagged outcome models for recovery data 131

Table 6:3 First-difference notation 132

Table 6:4 Change model results - first-difference method 134

Table 6:5 Comparison of naive model estimates with initial conditions method. Three equation specification joint modelled with $x0 = 1$ in initial equations........ 138

Table 6:6 Comparison of naive model estimates with initial conditions method. Three equation specification joint modelled with $x0 = x1$ in initial equations........ 139

Table 7:1 Possible scenarios of statistical hypothesis tests 143

Table 7:2 Design effects for estimating fixed effects........ 146

Table 7:3 Sample sizes for simulation procedure 153

Table 7:4 Effect sizes and variance estimates for simulation procedure 155

Table 7:5 Results of varying model and predictor variances for simple association model with fixed effect size 165

LIST OF FIGURES

Figure 1:1 Three-level data structure for ESM........19

Figure 2:1 Visual representation of the three-level random intercept model........24

Figure 2:2 Random coefficient model with positive covariance........26

Figure 2:3 Random coefficient model with negative covariance........26

Figure 2:4 Choice of centring in a random intercept model - x centered at 1 or 3........28

Figure 2:5 Choice of centring in a random slope models - x centred at 1 and 3........28

Figure 2:6 Homoskedastic errors around line of best fit........31

Figure 2:7 Constant level 1 variance with homoskedastic errors........31

Figure 2:8 Complex level 1 variance for binary x........32

Figure 2:9 Complex level 1 variance for continuous x........32

Figure 3:1 Systematic review flow of studies diagram........49

Figure 4:1 Compliance rates in systematic review (n=41)........60

Figure 4:2 Proportion of item nonresponse at each beep, conditional on event response. Recovery data........67

Figure 4:3 Proportion of item nonresponse on each day, conditional on event response. Recovery data........67

Figure 4:4 Proportion of item nonresponse at each beep, conditional on event response. Bipolar data........67

Figure 4:5 Proportion of item nonresponse on each day, conditional on event response. Bipolar data........67

Figure 4:6 Distribution of missed diaries in recovery data........68

Figure 4:7 Average proportion of missed diaries throughout the day in the recovery data.69

Figure 4:8 Proportion of missed diaries over the week in the recovery data........69

Figure 5:1 Box plot of Peters' vs Udachina's MSSD method for recovery data........84

Figure 5:2 Box plot of Udachina's vs Palmier-Claus' MSSD method for recovery data........84

Figure 5:3 Two-level model – Subject level random intercept, no fixed or random effect of time........86

Figure 5:4 Two-level model – Subject level random intercept, fixed effect of time........87

Figure 5:5 Two-level model – Subject level random intercept and random gradient for time........87

Figure 5:6 Two-level model – Subject level random intercept, both fixed and random effect of time........88

Figure 5:7 Three-level fixed and random time effect array........96

Figure 5:8 Variation in weekly recovery for the two baseline groups: population average trends plus variation at level 1.. 106
Figure 5:9 Variation in daily recovery for the two baseline groups: population average trends plus variation at level 1.. 106
Figure 5:10 Non-linear population average weekly trends in recovery for two groups...... 109
Figure 5:11 Time dependent heteroskedasticity between two groups: weekly trend........ 109
Figure 5:12 Non-linear population average daily recovery trends in two groups 111
Figure 5:13 Time dependent heteroskedasticity between two groups: daily trend 111
Figure 5:14 Time dependent heteroskedasticity between two groups: across-day variation in within-day trend 112
Figure 6:1 Parameter estimates of naïve change score models in the recovery data......... 120
Figure 7:1 Power for association with effect size 0.1 and small predictor and model variances – Association model. 158
Figure 7:2 Power for association with effect size 0.3 with small predictor and model variances – Association model 159
Figure 7:3 Power for effect size 0.1, large level 3 model variance, all remaining variances small – Association model... 160
Figure 7:4 Power for effect size 0.1, large level 2 model variance, all remaining variances small – Association model... 161
Figure 7:5 Power for effect size 0.1, large residual variance, all remaining variances small – Association model... 162
Figure 7:6 Power for effect size 0.1, large level 3 predictor variance, all remaining variances small – Association model... 163
Figure 7:7 Power for effect size 0.1, large level 2 predictor variance, all remaining variances small – Association model... 164
Figure 7:8 Power for effect size 0.1, large level 1 predictor variance, all remaining variances small – Association model... 165
Figure 7:9 Power for varying levels of missing data. Effect size 0.1, small model and predictor variances – Association model 166
Figure 7:10 Power for effect size 0.2, small model variances – Group difference model ... 168
Figure 7:11 Power for effect size 0.5, small model variances - Group difference model.... 169
Figure 7:12 Power for effect size 0.8, small model variances - Group difference model.... 169
Figure 7:13 Power for effect size 0.8, small model variances, larger n3 - Group difference model 170

Figure 7:14 Varying model variances for effect size 0.8 – Group difference model 171

Figure 7:15 Power under difference proportions of missing diary data - Group difference model 172

Figure 7:16 Power with missing data in both predictor and outcome - Group difference model 173

ABSTRACT OF THESIS

THE UNIVERSITY OF MANCHESTER

LESLEY-ANNE CARTER

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

RIGOROUS METHODS FOR THE ANALYSIS, REPORTING AND EVALUATION OF ESM STYLE DATA

2016

Experience sampling methodology (ESM) is a real-time data capture method that can be used to monitor symptoms and behaviours as they occur during everyday life. With measures completed multiple times a day, over several days, this intensive longitudinal data collection method results in multilevel data with observations nested within days, nested within subjects.

The aim of this thesis was to investigate the optimal use of multilevel models for ESM in the design, reporting and analysis of ESM data, and apply these models to a study in people with psychosis.

A methodological systematic review was conducted to identify design, analysis and statistical reporting practices in current ESM studies. Seventy four studies from 2012 were reviewed, and together with the analysis of a motivating example, four significant areas of interest were identified: power and sample size, missing data, momentary variation and predicting momentary change. Appropriate multilevel methods were sought for each of these areas, and were evaluated in the three-level context of ESM.

Missing data was found to be both underreported and rarely considered when choosing analysis methods in practice. This work has introduced a more detailed understanding of nonresponse in ESM studies and has discussed appropriate statistical methods in the presence of missing data.

This thesis has extended two-level statistical methodology for data analysis to accommodate the three-level structure of ESM. Novel applications of time trends have been developed, were time can be measured at two separate levels. The suitability of predicting momentary change in ESM data has been questioned; it is argued that the first-difference and joint modelling methods that are claimed in the literature to remove bias possibly induce more in this context.

Finally, Monte Carlo simulations were shown to be a flexible option for estimating empirical power under varying sample sizes at levels 3, 2 and 1, with recommendations made for conservative power estimates when a priori parameter estimates are unknown.

In summary, this work demonstrates how multilevel models can be used to examine the rich data structure of ESM and fully utilize the variation in measures captured at all levels.

DECLARATION

I declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

Acknowledgements

I would like to say that the completion of this thesis has been a solemn, dignified pursuit of knowledge, that I have fought alone these last four years, resolute in my endeavour to seek enlightenment through diligent, solitary study. Unfortunately, anyone who has ever met me would know that I am neither dignified nor enlightened, and that all this work would not have been possible without the masses of support from others.

Thank you to my supervisors Chris Roberts and Richard Emsley for their continued support and endless patience with all my passing deadlines. I can only hope I have absorbed at least some of their seemingly limitless knowledge along the way.

I could never have reached these lofty academic heights without the support of my friends. Thanks to Ava, Jen and Sam for keeping me fed and sane, thanks to Martina, Antonia, Matt and Ian for making me laugh and enduring my many moody episodes over the last four years. I would be remiss not to acknowledge the contribution one particular person has played: the daily shame and belittlement cast upon me by my far superior colleague and friend Jack Wilkinson has given me the motivation to strive to meet his excellence. I can only hope that one day I will.

Finally thank you to my family, to my brother and sister-in-law for showing me how it’s done and understanding my tears and tantrums, and to my parents for their constant love and support. Thank you Mum for all your offers to help and only suggesting a couple of times that maybe I’d rather go and earn lots of money working in the city.

LIST OF ABBREVIATIONS

AA	Ambulatory assessment
ANOVA	Analysis of Variance
AR	Autoregression
ARMA	Autoregression with moving average
BN	Bulimia nervosa
BPD	Borderline personality disorder
DE	Design effect
EMA	Ecological momentary assessment
ESM	Experience sampling methodology
GEE	Generalized estimating equations
HAR	Heterogeneous autoregression
ICC	Intraclass correlation coefficient
ID	Independent covariance structure
L1	Level 1
L2	Level 2
L3	Level 3
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
MD	Missing data
MDD	Major depressive disorder
MI	Multiple imputation
MI/LMM	Multiple imputation using linear mixed models
MI/NM	Multiple imputation using a normal model
MLM	Multilevel model
PDA	Personal digital assistant
PinT	Power analysis in two-level designs
TOEP	Toeplitz covariance structure

1 INTRODUCTION AND MOTIVATION FOR RESEARCH

This thesis will pursue statistical methods for the design and analysis of intensive longitudinal data, with specific applications to experience sampling methodology. This chapter will introduce experience sampling methodology, describing how and why it is used, and the resulting data structure. The aim and specific objectives of the thesis will then be defined.

1.1 EXPERIENCE SAMPLING METHODOLOGY

Experience sampling methodology (ESM) (Larson R and Csikszentmihalyi 1983; Delespaul 1995), also known Ecological Momentary Assessment (EMA) (Stone and Shiffman 1994; Shiffman, Stone et al. 2008), is a specialist diary based questionnaire used to gather momentary data from participants during their everyday life. ESM has been used to study symptoms, behaviours and attitudes in a range of settings including studies into psychosis, eating disorders, craving and addiction as well as research into quality of life and social relationships. In contrast to traditional questionnaires and clinical interviews delivered at the start and end of the study period, ESM is a self-reported assessment that is completed during participant's everyday life, continuously capturing symptoms and thoughts in real time.

One of the main disciplines conducting research using ESM is psychology and as will be discussed in Chapter 3, a broad range of topics within psychology are investigated using ESM. Although not the only field utilising this methodology, for continuity, psychology and mental health research will be the empirical area motivating this thesis.

A typical ESM questionnaire contains a collection of items designed to gather momentary information using short, unambiguous questions such as "Right now I feel cheerful", "Right now I see phenomena". Here, 'momentary' refers to participants rating their experience at the current time, capturing the variable cognitive state rather than the more stable trait (Larson R and Csikszentmihalyi 1983; Delespaul 1995; Csikszentmihalyi 2014) . Responses are given on both numerical scales and as open ended questions. Numerical scales include visual analogue scales where the participant marks a score on a continuous line representing a score of 1 – 100, and Likert scales which allow the participant to grade their answer using distinct categories, for example a 1 – 7 scale representing categories from "Not at all" to "Very much". Open ended questions can be used for items regarding current location or for recording present company. In the initial briefing session, participants are usually given a copy of the questionnaire which is completed with the researcher present to ensure the scales and questions are clear. Items are intended to capture the moment to moment variation in thoughts and symptoms and thus are designed in a way that simultaneously encourages variation in scores and limits ceiling and floor effects.

In contrast to other longitudinal methods, ESM is often employed to study short term variation and as such diaries are completed multiple times a day over several days or weeks. Though there is no fixed schedule attached to ESM, 10 diaries a day over a six day period is often cited with reference to Delespaul (1995) and the Maastricht University ESM research group. The justification for this figure is that 10 measurements a day should provide sufficient detail to study within-day variation and one week is considered long enough to capture the range of symptom and mood activity. Although this sampling routine is adopted by the Maastricht group and those looking to them for procedural guidance, there has been no published statistical justification behind this procedure. In the ESM literature the number of measurements varies depending on study design, from one moment a day (often referred to as daily diary studies rather than ESM studies) to several weeks of observation.

The sampling procedure for ESM studies can be partitioned into two main categories: signal contingent designs, where subjects are alerted to complete a diary via a programmed device, or an event contingent design in which subjects complete a diary following a predetermined event (Bolger, Davis et al. 2003). Signal contingent designs can be further divided into random or interval designs. Interval contingent designs require a participant to complete a diary at predetermined intervals throughout the day after being signalled by a programmed watch, PDA or mobile phone. For example, one might be signalled to complete a questionnaire at 7am, 11am, 2pm and 7pm each day. This design is effective for covering the desired study period; however, it has the potential for participants to become aware of the routine which could affect their mood or behaviour, skewing the results of the study. Random designs remove this problem, but may result in intervals between measurements that are hugely varied. This could lead to prompts in rapid succession which do not represent the full day's sampling period. Moreover, this is potentially very stressful for participants, which may be reflected in their answers to items, skewing the results or may result in uncompleted diaries. Conversely, measurements too far apart could be demoralizing for participants (Delespaul 1995). A compromise which benefits from the positive aspects of both designs is a block random sampling design, also referred to as a pseudo-random design, in which participants are signalled to complete a diary at random points within fixed time intervals, allowing the researcher to ensure samples are taken across the entire day while reducing any “anticipatory behavioural change” (Delespaul 1995) from participants’ expectation of the next moment.

In event contingent designs participants complete a diary each time a specific event occurs. Events might include each time the participant smokes a cigarette in smoking cessation studies, uses drugs in addiction studies or overeating events in studies exploring eating disorders. Event contingent designs are useful for capturing rarely occurring events that might not be observed using a signal contingent design: as items are often phrased “Right now I…” to capture behaviours in the current moment, behaviours in the time between measurements will not be recorded. Restricting questionnaires to be completed only after events, however, could lead to significant missing data: forgetting to complete one or two questionnaires when there are 10 a day may not have too great an impact on the findings but missing two rarely occurring events could lead to a participant having no data for the study period. This could lead to further problems if the reason the diary was not completed was linked to the event, to be discussed further in Chapter 4. To counteract this, one can use a mixed design with questionnaires completed at events and signalled prompts. This will record an event as it occurs but also gather more data pre- and post-event, providing the opportunity to record missed events at a later date and to gather contextual information in the moments leading up to and away from the event.

1.1.1 Benefits of methodology

The momentary nature of ESM allows researchers to identify mood, symptoms and behaviours as they occur in everyday life. A number of claims are made regarding the benefits of using ESM. Firstly, it allows for detailed monitoring of states which can help researchers understand conditions in a different way to traditional questionnaires, capturing short term variation and real-time reaction to events. Basic descriptive statistics can be used to visualize this moment to moment variation, such as simply plotting the variable of interest over time. While traditional measures may be used to study change over weeks or months, ESM can look at change throughout a day, and daily trends can be visualized over weeks.

It has been argued that ESM, with its momentary collection of data, can be used to memory effects (Larson R and Csikszentmihalyi 1983) or recall bias which can be a problem in traditional self-report questionnaires: when asked to summarize or average previous experience (or to use autobiographical memory (Bradburn, Rips et al. 1987)) substantial bias can be introduced; subject may only recall extremes and so provide overestimates of past experience; or where estimates of past experiences are influenced by their current state (Stone and Shiffman 1994; Shiffman, Hufford et al. 1997; Shiffman, Stone et al. 2008; Solhan, Trull et al. 2009).

Time varying associations can be also investigated using ESM, identifying patterns that can be used to further explore psychological disorders. These associations may be psychological or environmental; a person's current location or situational context may be used to explain variation in symptoms. For example, research has been conducted on the relationship between drug use and psychosis (Verdoux, Gindre et al. 2003; Henquet, van Os et al. 2010), where the temporal effect of drug use on symptoms is investigated, i.e. does drug use predict mood change or mood change predict drug use?

The effect of treatment or interventions can be studied using this methodology. ESM can be used to collect data pre and post treatment to allow comparison, such as in smoking cessation studies where craving is monitored before and after intervention (Bolt, Piper et al. 2012). ESM also has the potential to be used to monitor symptoms during a treatment or intervention. For example, it could be used to monitor adverse reactions and side effects to cancer or pain medication. Real-time intervention is an emerging field in electronic data collection studies such as ESM, the feasibility of which is currently under study. A review of the current literature on ESM interventions in psychiatry has recently been published (Myin-Germeys, Klippel et al. 2016) which found that, whilst still in its infancy, this method of intervention appears feasible and acceptable in the population with severe mental illness, though published results on efficacy are extremely limited.

Finally, it has been suggested that ESM can also be used to help patients understand their condition and could be used to assist medical professionals in giving more personalized care, using graphical examples of symptom variation and links to predictors, for example. This personalised feedback has been demonstrated to improve patient symptoms in the long term (Kramer, Simons et al. 2014), where patients given weekly feedback while using ESM were seen to have significantly improved symptoms at six month follow up compared to those completing ESM without feedback and to a control group. This form of personalised feedback has also been investigated to support of dementia caregivers (van Knippenberg, de Vugt et al. 2016), however with less promising results.

These suggested benefits all rely on two key points: that the questionnaire captures the construct intended and, importantly, that these constructs vary throughout the observation period and the measures used are sensitive enough to capture this moment to moment variability. Item selection, therefore, is a very important aspect of study design and should be carefully considered, though it is beyond the scope of this work.

1.1.2 ESM AND TECHNOLOGY

Originally administered via a paper booklet, developments in technology have allowed ESM to be delivered on PDAs and smartphones. These methods benefit from being able to time-stamp questionnaires allowing researchers to know exactly when (and potentially where) each questionnaire was completed. One drawback of ESM using paper booklets has been the uncertainty over compliance to protocol. ESM successfully captures momentary data if the questionnaires are completed as soon as the participant is alerted to complete it, with alerts for paper diaries usually administered via an alarm from a wristwatch or beeper. If the participant takes too long to respond, the answers are not thought to accurately represent the person's 'current' state as defined by the protocol. Time-stamped responses can be used to check adherence to the protocol and PDAs and smartphone can be programmed so the device doesn't allow data entry after suitable time period.

Smartphones also have the ability to capture more data than just the ESM questions, for example GPS (global positioning systems) can provide accurate information on participants' location and applications have been developed to monitor sleep activity (Andriod ; SleepCycle). Further developments in software for smartphones (ClinTouch ; Palmier-Claus, Ainsworth et al. 2012) only increases the potential for ESM as a viable data collection technique. Already used in conjunction with ESM is ambulatory assessment (AA). AA can be used to measure biological responses in the same momentary settings as ESM. Wilhelm and Grossman (2010) provide a comprehensive summary of biosignals and devices that can be used for AA with examples including heart rate, systolic/diastolic blood pressure, respiratory rate and cortisol measurements.

1.2 DATA STRUCTURE

Data collected using the ESM procedure are a series of repeated measures observed for each subject over a set period of time. This type of data is known as longitudinal data. Longitudinal data typically captures changes in variables over a long period of time – months, years or even decades. As several measurements are taken for each person the data are correlated, where measurements are likely to be more similar within-person rather than between-people. This correlation can be accounted for in the analysis of longitudinal data but is not often the focus. The ESM data presented here will be discussed as multilevel data, where measurements are considered nested within higher level clusters. ESM data can have a more complex structure than typical longitudinal data, where measurements can be nested within more than one unit.

Throughout this work ESM will be considered to have a three-level data structure, with measurements at the lowest level, level 1, referred to as the moment or ‘beep’ level (Delespaul 1995) with reference to the alarm used in signal contingent designs, nested within the level 2 ‘day’ level, nested within subject at level 3. Including this day level is uncommon in practice, as will be demonstrated in Chapter 3. However, measurements taken within a day may be more highly correlated than those taken the next day. Identifying this additional level of data allows for not only within- and between-person analysis but also within- and between-day analysis. That is, one can study how symptoms and behaviours vary moment to moment, day to day and person to person. Higher levels of data are also possible, such as participants nested within therapist or centre.

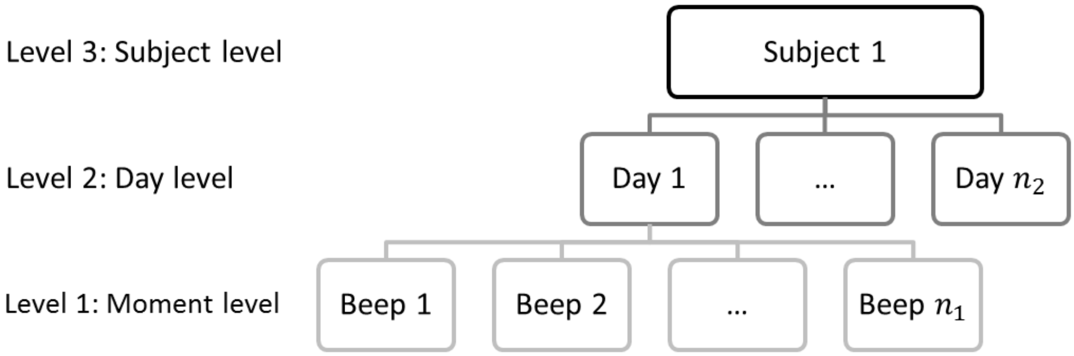


Figure 1:1 Three-level data structure for ESM

The following notation will be used throughout, with additional notation defined where necessary. To allow for various higher levels of ESM data, the moment level will always be defined as level 1, the lowest level. Measurements at level 1 will be recorded at moments $i = 1, \dots, n_{1jk}$ where the subscripts 1 denotes the level of measurement, j the day number to allow for a different number of measurements to be taken each day, and k to allow the number of measurements to vary per person. Day number is denoted $j = 1, \dots, n_{2k}$ where similarly the subscripts 2 refer to the level of measurement and k to the participant number, allowing each participant to be observed for a different number of days. Participants are numbered $k = 1, \dots, n_3$. For simplicity it will often be assumed that the number of moments per day and the number of days of measurement will be the same for each participant. In these circumstances the above notation can be simplified to n_1, n_2 and n_3.

Analysis of ESM data is dependent on the research question and at which level the interest of the researcher lies: the moment-level, day-level or subject-level. Schwartz and Stone (1998) categorise this into 3 areas: participant-level variation, within-subject variation, and whether participant-level characteristics predict changes in within-subject variation. Defining the level of interest will define the outcome of interest and lead to an appropriate form for the analysis model.

Investigating the between-subject variation is possible by aggregating the lower levels and performing statistical tests appropriate for single level data. One of the main issues with this approach is the problem of heteroskedasticity when one must aggregate over different numbers of assessments per participant. In this scenario, Schwartz and Stone (1998) describe that the aggregated value, or participant mean, is subject to sampling variability, where fewer available observations contribute to a greater amount of variation. Regressing on this aggregated value then leads to residual variation from both the usual unexplained variance and this sampling variability. As unbalanced and missing data are common within ESM, this is a significant problem as it violates the assumption of linear regression that variance of the residuals is constant. A second drawback of this approach is that restricting the analysis to the subject level ignores the potentially rich information held in the between-subject variation, one of the main benefits of ESM style data.

To analyse within-subject variation, a model is required that can accommodate repeated measurements for each participant and clustering at higher levels. Simple regression models are not suitable for analysing this type of data as the assumption of independent residuals is not met. This thesis will explore the use of multilevel models to analyse ESM data, where momentary-level, day-level and subject-level variation can be accommodated.

1.3 AIMS AND OBJECTIVES OF THIS THESIS

The aim of this work is to examine the optimal use of multilevel models for the design, reporting, analysis and exploration of ESM studies.

This work will be motivated by an example study presented in Chapter 2 which highlighted a methodological issue when using multilevel models to study change. The current ESM literature will be reviewed in Chapter 3 to identify the types of research questions being investigated with ESM, and to determine whether the statistical methods used to address these questions are appropriate. Chapter 4 will introduce missing data and how it can be explored in ESM data, and Chapters 5 and 6 will investigate methods for exploring momentary variation and change, presenting statistical challenges and discussing potential solutions. Chapter 7 will examine power and sample size calculations for ESM data. Finally, Chapter 8 will discuss the benefits and limitations of these methods in an ESM setting.

Objectives:

1. Reporting – Understanding missing data in ESM research will be thoroughly explored and the implications for analysis will be investigated
2. Exploration – alternative methods to aggregation will be sought to study momentary level variation or fluctuation of outcome that fully utilise the multilevel nature of the data
3. Analysis – models for predicting momentary change will be explored to investigate the unusual parameter estimates observed in the recovery data of Chapter 2, with an examination of methods and their appropriateness for ESM data.
4. Design – closed form expressions and empirical power estimates will be investigated for three-level data, appropriate for two example research questions identified in the systematic review.

Each of these four objectives will be developed in Chapters 4 - 7. Each chapter will begin with an introduction outlining how each concept is currently reported or analysed in ESM with a more detailed examination of relevant papers identified in the systematic review of Chapter 3. The statistical methodology for each concept will then be discussed with extensions developed for three-level data, followed by an application of the method to the example dataset of Chapter 2.

2 Introduction to multilevel models and motivating examples

The first half of this chapter will provide an introduction to multilevel modelling from which subsequent chapters will expand. To illustrate the data structure and analysis procedure of ESM data, two example data sets will be referred to throughout. The second half of this chapter presents the details of these two ESM studies. The first assessed self-reported recovery in schizophrenic patients, the research conducted by Richard Bentall, Tony Morrison and their team at the University of Manchester, and is currently being prepared for publication (Bentall et al in preparation). The second study examined the relationship between cannabis use and bipolar disorder. This research was undertaken by Elizabeth Tyler and colleagues at the University of Manchester (Tyler, Jones et al. 2015). For both studies, myself and Richard Emsley advised on and conducted the statistical analysis.

2.1 An introduction to multilevel models

ESM data can be analysed using multilevel models, also known as random effects, mixed effects or hierarchical models. These models allow for multiple levels of data to be considered without the need for aggregation, and can be used to examine variation at each level. They can accommodate the nested structure of ESM data, are valid for unbalanced data sets and can be extended to fit complex covariance structures arising in the data.

This section will provide a general overview of multilevel models and the data requirements necessary for valid parameter estimates. A more detailed examination of the statistical methodology for specific research questions will be provided in the introduction to the relevant chapters of this thesis.

2.1.1 Random intercept models

The parameterisation of random effect models can be first considered with a simple two level example. Equation (1) represents a two-level random intercept model for data in which measurements at level 1 are nested within participant at level 2,

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + u_k + e_{ik}. \tag{1}$$

In this model y_{ik} is the response value for the ith measurement of individual k and x_{ik} is a moment level explanatory variable. The model can be split onto two parts: the fixed part and the random part. The fixed part of the model, $\beta_0 + \beta_1 x_{ik}$, represents the population average effects. The intercept β_0 is the mean value of y_{ik} when $x_{ik} = 0$ and β_1 is the average effect of covariate x_{ik} on y_{ik}. The random part of the model represents the variation in outcome at each level: u_k is the level 2 random intercept and e_{ik} the level 1 residuals. The level 2 random intercept represent unobserved subject-level heterogeneity and can be thought of as the kth participant's deviation from the overall mean, or $\beta_0 + u_k$ as the subject-specific intercept for participant k. The random effects are assumed to be normally distributed such that $u_k \sim N(0, \sigma_u^2)$ and similarly the residuals $e_{ik} \sim N(0, \sigma_e^2)$, with u_k independent across participants k and covariates x_{ik}, and e_{ik} independent over both participants k and occasions i. This model can be used to investigate variation at both levels by providing estimates of σ_u^2, the between participant variation, and σ_e^2, the within participant variation, as well as examining the effect of level 1 and 2 covariates on the outcome.

Unbiased parameter estimates require the random effect models comply with several assumptions. Letting $\boldsymbol{x_{ik}} = (x_{1ik}, \dots, x_{pik})'$ be a set of covariates, the exogeneity assumptions from single level models still apply and are extended to this two-level scenario, where

$$E(u_k | \boldsymbol{x_{ik}}) = 0$$

and

$$E(e_{ik} | \boldsymbol{x_{ik}}, u_k) = 0$$

from which we have $E(e_{ik} | \boldsymbol{x_{ik}}) = 0$. These assumptions mean that the random effects and and level 1 residuals are uncorrelated with the covariates $\boldsymbol{x_{ik}}$. If this assumption is violated the parameter estimates can be substantially biased (Ebbes, Böckenholt et al. 2004).

Further assumptions of random intercept models include the independence of random effects between subjects, that is the random effects for two subjects k and k' are uncorrelated

$$cov(u_k, u_{k'}) = 0$$

the residuals for two observations are uncorrelated within-subject ($k = k'$) and for two different individuals ($k \neq k'$)

$$cov(e_{ik}, e_{i'k'}) = 0$$

and the random effects at levels 1 and 2 are uncorrelated within- and between-individuals

$$cov(u_k, e_{ik'}) = 0.$$

The two-level model can be extended to fit a three-level data structure where measurements i are now nested within days j within participants k, as in Figure 1:1, by including a random intercept for day v_{jk}:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_k + v_{jk} + e_{ijk.}$$

Here, $u_k \sim N(0, \sigma_u^2)$ represents the random intercept for each subject, $v_{jk} \sim N(0, \sigma_v^2)$ the random intercept for each day j within each subject k and $e_{ijk} \sim N(0, \sigma_e^2)$ the moment level residuals. From this model, as well as the participant level variation σ_u^2, the day level variation σ_v^2 can be estimated. .

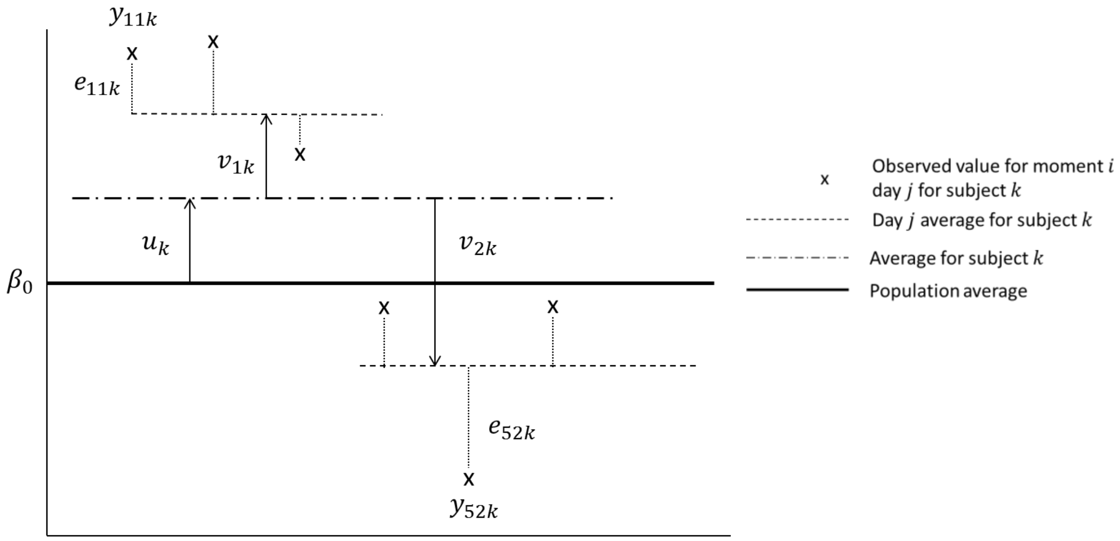


Figure 2:1 Visual representation of the three-level random intercept model

In addition to variation at each level of the data, multilevel models can be used to examine within- and between-cluster effects (Snijders and Bosker 1999). In this three-level structure of ESM data one can investigate both within- and between-subject effects as well as within- and between-day effects. The separation of these effects is achieved by including the group mean of a level one variable in the model, for example in the model

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \beta_2 \bar{x}_{..k} + u_k + v_{jk} + e_{ijk} \tag{2}$$

$\bar{x}_{..k}$ is the subject-specific mean of the level 1 variable x_{ijk}. In contrast to equation (1) this model estimates the within-subject effect of x, β_1, and, aggregating to the subject-level on both sides of the equation

$$\begin{aligned}\bar{y}_{..k} &= \beta_0 + \beta_1\bar{x}_{..k} + \beta_2\bar{x}_{..k} + u_k + \bar{v}_{.k} + \bar{e}_{..k} \\ &= \beta_0 + (\beta_1 + \beta_2)\bar{x}_{..k} + u_k + \bar{v}_{.k} + \bar{e}_{..k}\end{aligned}$$

the between-subject effect is shown to be $\beta_1 + \beta_2$. Similarly, within- and between-day effects can be studied by including the day-specific mean of x

$$\bar{y}_{.jk} = \beta_0 + \beta_1 x_{ijk} + \beta_2\bar{x}_{.jk} + u_k + v_{jk} + e_{ijk}.$$

2.1.2 Random coefficient models

Unlike a random intercept model where the relationship between x and y is fixed, a random coefficient model allows this relationship to vary between participants. A two-level random coefficient model has the form

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + u_{0k} + u_{1k}x_{ik} + e_{ik}$$

or, rearranging,

$$y_{ik} = (\beta_0 + u_{0k}) + (\beta_1 + u_{1k})x_{ik} + e_{ik}$$

where u_{0k} is the random intercept for individual k and u_{1k} is the random slope for x_{ik}. These random effects come from a multivariate normal distribution with

$$\begin{bmatrix} u_{0k} \\ u_{1k} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{bmatrix}\right).$$

As in the random intercept model, there is a subject-specific intercept, $\beta_0 + u_{0k}$, but now there is also a subject-specific effect of x, $\beta_1 + u_{1k}$. For example, using ESM to study psychosis one might hypothesise that increased anxiety leads to an increase in paranoia. Fitting a random intercept model one would assume that although subjects may start out with different levels of paranoia, accounted for by the random intercept u_{0k}, the effect of anxiety is equal for each individual, resulting in parallel subject-specific slopes with fixed gradient β_1. In a random coefficient model, this relationship is allowed to vary between people, such that a unit increase in anxiety might have a stronger or weaker effect on paranoia for each subject k; their individual slope being $\beta_1 + u_{1k}$.

The random intercept and slope covariance, σ_{u0u1}, is a measure of how the value of the intercept influences the slope for each individual. For a positive covariance, $\sigma_{u0u1} > 0$, a subject-specific line with a larger intercept value will have a steeper than average slope, while a subject-specific line with a smaller intercept will have a shallower than average slope – see Figure 2:2. Conversely, for a negative covariance, larger intercept values will lead to shallower slopes whereas a smaller intercept will lead to a steeper slope - see Figure 2:3. In terms of the psychosis example, a positive covariance would mean that for individuals who exhibit higher levels of paranoia for low levels of anxiety, anxiety will have a greater effect, where as those with minimal symptoms for low levels of anxiety would see a much smaller increase in paranoia as anxiety increases.

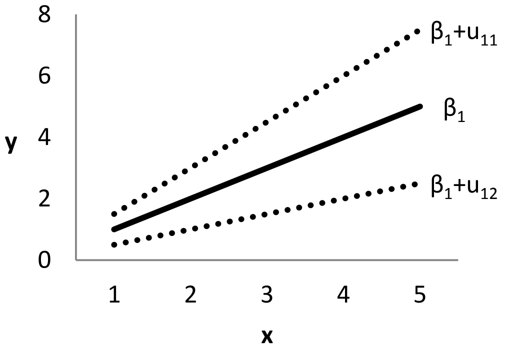


Figure 2:2 Random coefficient model with positive covariance

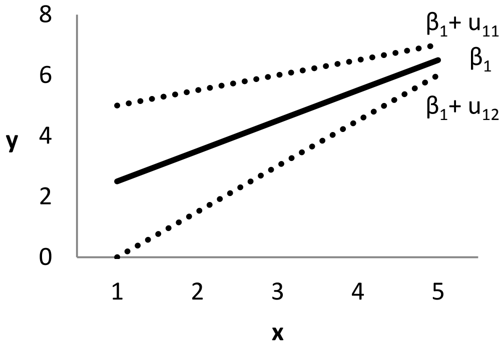


Figure 2:3 Random coefficient model with negative covariance

For three-level data it is possible to have random slopes at level 3 and level 2. For ESM data this corresponds to variation in subject-specific slopes when a random coefficient is included at level 3,

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_{0k} + u_{1k} x_{ijk} + v_{jk} + e_{ijk}$$

or variation in day-specific slopes when included at level 2,

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_k + v_{0jk} + v_{1jk} x_{ijk} + e_{ijk}.$$

The interpretation of three-level random slope models will be discussed in much greater detail in Chapter 5.

The multilevel models described in this thesis will be fitted using maximum likelihood (ML) estimation. It is argued that ML can produce biased estimates of the random effect variances when the number of clusters is small (Rabe-Hesketh and Skrondal 2012). Restricted maximum likelihood (REML) can be used as an alternative, however the difference between the two methods is considered trivial for large numbers of clusters (Snijders and Bosker 1999). A drawback of using REML, the likelihood function only includes parameters from the random part of the model and so likelihood ratio tests to compare nested models with different fixed effect specifications are not possible (Snijders and Bosker 1999; Fitzmaurice, Laird et al. 2012). As such, estimation using ML will be the preferred method. One exception to this will be when random effect variance and covariance estimates are the primary focus, i.e. when fitting random slope models, where REML will be used. In this case, likelihood ratio tests will only be used to compare models with nested random effects where the fixed effect specification is the same in both models.

2.1.3 Centring

For the models described above, interpretation of the regression coefficients is dependent on the scaling of the model variables. For the fixed effects, the intercept β_0 represents the population average value of y when all covariates are equal to zero. Inference on β_0 thus relies on an interpretable meaning of zero for x. For ESM research many of the questionnaire items are measured on a 1-7 Likert scale with the values representing an ordinal style measure of agreement, for example 1 = not at all, through to 7 = very much so. For these measures a value of zero makes no conceptual sense and as such in a model using these scales as covariates the fixed intercept would be uninterpretable. It is thus recommended that variables be rescaled to give meaningful zero values, known as centring.

There are several ways to choose the value with which to centre a variable, each providing a different context for interpreting both the fixed and random effects. On a 1-7 scale, centring around $x = 3.5$, β_0 represents the population average value of y for a mid-level x score of 3.5. If the scale is labelled as above, this would translate to the average value of y for a ‘neutral’ x score. Alternatively, one can center by the cluster mean, known as group-mean centring. In a two level model, the mean of a level 1 variable, $\bar{x}_{.j}$, taken over all units i within cluster k is deduced from x_{ik},

$$y_{ik} = \beta_0 + \beta_1(x_{ik} - \bar{x}_{.k}) + \beta_2\bar{x}_{.k} + u_k + e_{ik} \tag{3}$$

This is statistically equivalent to the two-level version of equation (2) above. It is recommended that level 1 variables be cluster-mean centred to investigate within- and between-group inference (Snijders and Bosker 1999) referred to in the longitudinal literature as longitudinal and cross sectional effects. In an ESM context, β_0 represents the average response for the mean x score within-subject, β_1 the effect of a unit increase in x within-subject and β_2 the effect of a unit increase x between-subject.

In the three-level data structure for ESM, level 1 variables can be centred at levels 2 or 3. Centring at level 2 $(x_{ijk} - \bar{x}_{.jk})$, β_1 represents an effect of an increase in x within-day, within-person, where as centering at level 3 $(x_{ijk} - \bar{x}_{..k})$ provides an estimate for the effect of an increase in x within-person, averaged across days.

Interpretation of random effects is also effected by the choice of centring. Random intercept and random slope variances are estimated at $x = 0$. In random intercept model subject-specific slopes stay constant so the choice of this zero value does not affect the intercept variance (Figure 2:4), however, in a random slope model subject-specific slopes are allowed to vary, and so the random slope variance will depend on the choice of centring for x (Figure 2:5).

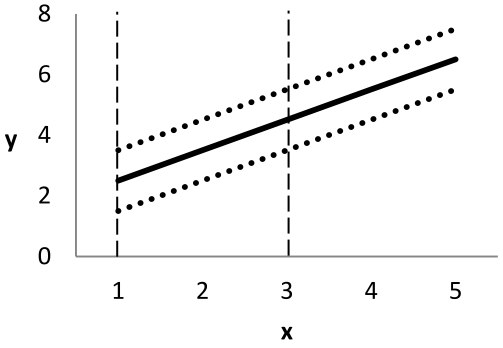


Figure 2:4 Choice of centring in a random intercept model - x centered at 1 or 3

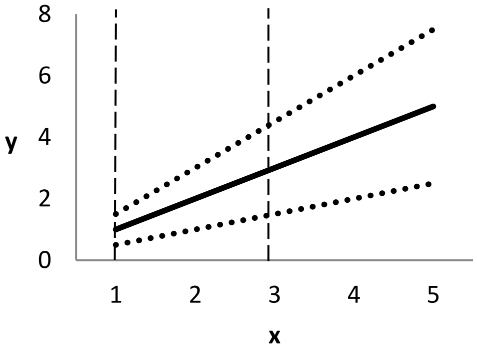


Figure 2:5 Choice of centring in a random slope models - x centred at 1 and 3

2.1.4 VARIATION IN A THREE-LEVEL MODEL

2.1.4.1 PARTITIONING VARIATION

Variation in measures can be studied at each level of the data. One method for describing this variation is the variance partitioning coefficient (VPC), which calculates the proportion of variance at each level. In a random intercept model the variance of measure y_{ijk}can be defined as

$$var(y_{ijk}) = var(\beta_0 + u_k + v_{jk} + e_{ijk}) = var(u_k) + var(v_{jk}) + var(e_{ijk})$$
$$= \sigma_u^2 + \ \sigma_v^2 + \sigma_e^2$$

where σ_u^2 represent the between-subject variation, σ_v^2 the between-day, within-subject variation and σ_e^2 the residual variation.

The proportion of variance at level 3 cab then be defined as the level 3 variance divided by the total variation

$$\rho_3 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2},$$

with the proportion of variance at level 2

$$\rho_2 = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}$$

and the proportion of variation at level 1

$$\rho_1 = \frac{\sigma_e^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}$$

similarly. An alternative way of describing variation is the intraclass correlation coefficient (ICC). Hox (2002) writes that the ICC can also be thought of as the “expected correlation between two randomly chosen elements in the same group” so for two observations y_{ijk} and $y_{i'j'k'}$

$$corr(y_{ijk}, y_{i'j'k'}) = \frac{cov(y_{ijk}, y_{i'j'k'})}{\sqrt{var(y_{ijk})}\sqrt{var(\ y_{i'j'k'})}}$$

where

$$cov(y_{ijk}, y_{i'j'k'}) = cov(\beta_0 u_k + v_{jk} + e_{ijk}, \beta_0 + u_{k'} + v_{j'k'} + e_{i'j'k'})$$
$$= cov(u_k, u_{k'}) + cov(v_{jk}, v_{j'k'}) + cov(e_{ijk}, e_{i'j'k'}).$$

For $i \neq i'\ j \neq j'\ k = k'$ the correlation between two observations from the same person measured on different days

$$corr(y_{ijk}, y_{i'j'k}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}$$

as $cov(v_{jk}, v_{j'k'}) = 0$ and $cov(e_{ijk}, e_{i'j'k'}) = 0$ when $i \neq i'$ and $j \neq j'$.

For the correlation between two moments within the same day, within the same person, i.e. $i \neq i'$ $j = j'$ $k = k'$, the ICC can be expressed as

$$corr(y_{ijk}, y_{i'jk}) = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}.$$

For random intercept models this interclass correlation coefficient is equal to the variance partitioning coefficient. However, for random slope models this is not the case. Taking a two-level random coefficient model

$$y_{ik} = \beta_0 + \ \beta_1 x_{ik} + u_{0k} + u_{1k} x_{ik} + e_{ik}$$

the variance at level 2 is now

$$\begin{aligned} var(u_{0k} + u_{1k}x_{ik}) &= var(u_{0k}) + 2cov(u_{0k}, u_{1k}x_{ik}) + var(u_{1k}x_{ik}) \\ &= \ \sigma_{u_0}^2 + 2\sigma_{u_{01}} x_{ik} + \sigma_{u_1}^2 x_{ik}^2 \end{aligned}$$

a quadratic function of covariate x_{ik}. The VPC at level 2 is thus

$$VPC = \frac{\text{level 2 variance}}{\text{total variance}} = \frac{\sigma_{u_0}^2 + 2\sigma_{u_{01}} x_{ik} + \sigma_{u_1}^2 x_{ik}^2}{\sigma_{u_0}^2 + 2\sigma_{u_{01}} x_{ik} + \sigma_{u_1}^2 x_{ik}^2 + \sigma_e^2}.$$

This extends to three level models where both the level 3 and level 2 variances may be affected by a random coefficient. For random slope models the VPC rather than the ICC should be used when discussing the proportion of variance at each level of the model.

2.1.4.2 Complex level 1 variation

It is possible to model patterns in variance at level 1 in a similar manner to modelling variation at the higher levels. In a standard multilevel model it is assumed that the level 1 variance, σ_e^2, is homoscedastic: that the variance of the residuals around the fitted line is constant.

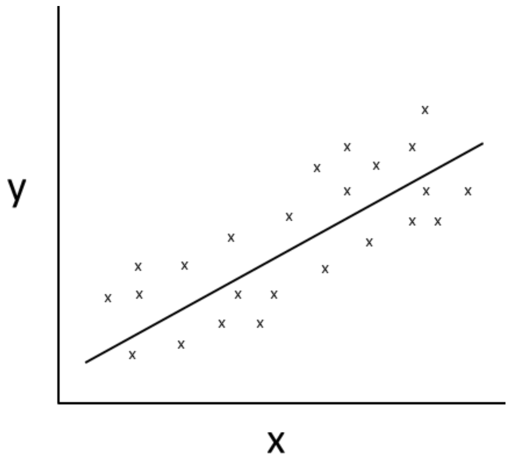


Figure 2:6 Homoskedastic errors around line of best fit

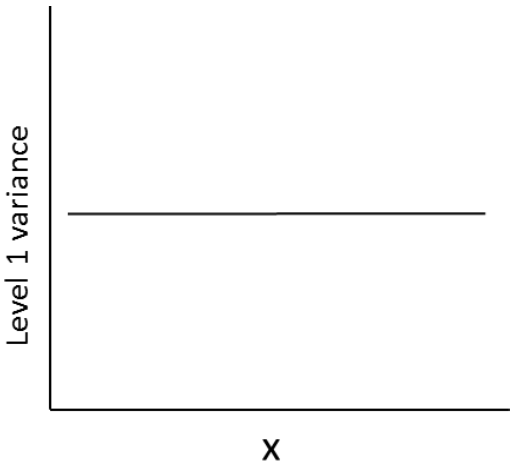


Figure 2:7 Constant level 1 variance with homoskedastic errors

This might not always be the case, and it may be of interest to model the level 1 variance in terms of some variable x, described by Steele (2008) as a "complex level 1 variance" model. If x is binary, heteroskedasticity can be measured by estimating the residuals separately for the two groups. For example, if the question of interest is how the outcome varies in each of two treatment groups then the following model can be used,

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_k + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_{ijk}(1)$$

where x_k is a dummy variable for treatment group, equal to 1 for patients receiving treatment and 0 for the control group. The residuals now follow a joint normal distribution with means zero and variance matrix

$$\Sigma_e = \begin{bmatrix} \sigma_{e0}^2 & 0 \\ 0 & \sigma_{e1}^2 \end{bmatrix}.$$

Including the group variable as a fixed effect will estimate the average effect of treatment on a moment level outcome y_{ijk}, while allowing for complex level 1 variance will give

$$var\left(e_{0ijk} x_k(0) + e_{1ijk} x_k(1)\right) = \sigma_{e0}^2 x_k^2(0) + \sigma_{e1}^2 x_k^2(1)$$

as σ_{e01} is assumed to be zero, and will provide separate variance estimates for treatment and control which can be interpreted as the level of variation in y at the moment level for the two groups.

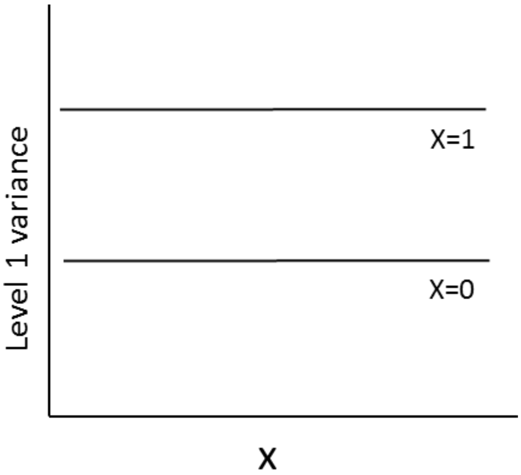
If x is continuous, the complex variation model becomes

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_k + v_{jk} + e_{0ijk} + e_{1ijk} x_k$$

where the level 1 variance should now be interpreted in terms of the variance function

$$var(e_{0ijk} + e_{1ijk} x_k) = \sigma_{e0}^2 + 2\sigma_{e01} x_k + \sigma_{e1}^2 x_k^2$$

providing a different estimate for each value of x, see Figure 2:9.



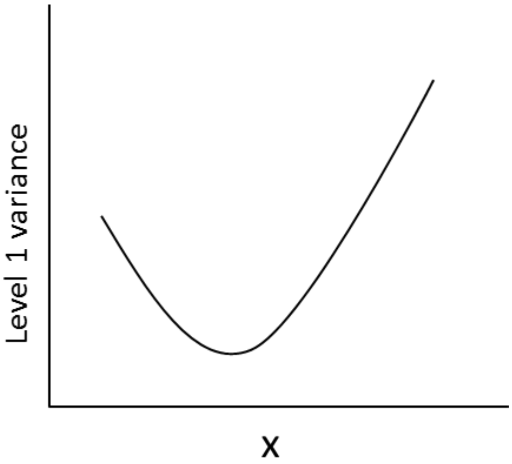


Figure 2:8 Complex level 1 variance for binary x

Figure 2:9 Complex level 1 variance for continuous x

Hedeker, Mermelstein et al. (2009) describe methods for examining heterogeneity as a function of subject-level covariates. Extending a two level random coefficient model

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \beta_2 \bar{x}_k + u_{0k} + u_{1k} x_{ik} + e_{ik}$$

where x_{ik} is a moment level variable, $\bar{x}_k$ is the subject level mean of x_{ik} and the random effects share a bivariate normal distribution $N(0, \Sigma_u)$ where

$$\Sigma_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{bmatrix},$$

they allow random effect variances to be modelled in terms of a subject-level covariate w_k such that

$$\sigma_{u_{0k}}^2 = \exp(\alpha_{00} + \alpha_{01} w_k)$$

$$\sigma_{u_{1k}}^2 = \exp(\alpha_{10} + \alpha_{11} w_k).$$

The variances now have subscript k to indicate their values depend on w_k. This subject-level covariate could also be included as a fixed effect to model its average effect on outcome y_{ik}. The covariate is modelled as an exponential function in the random variance to ensure the variance parameter is greater than zero.

When $w_k = 0$ the intercept and slope variances equal $\exp(\alpha_{00})$ and $\exp(\alpha_{10})$ respectively. When $w_k \neq 0$ the variances change by a function of w_k and its coefficient. For the random intercept, when $\alpha_{00} > 0$ the variation in subject-specific intercepts increases as w_k increases, or decreases when $\alpha_{00} < 0$. Similarly, for random slopes, when $\alpha_{11} > 0$ the variation in subject-specific slopes increases as w_k increases.

In their example, the authors are interested in how mood (y_{ik}) changes in adolescents during smoking and non-smoking events (x_{ik}), and how this relationship varies in those classed as frequent and non-frequent smokers (w_k). As the covariate x_{ik} is binary, the standard model with random coefficient for smoking event estimates σ^2_{u0} and σ^2_{u1}, the variation in mood at smoking and non-smoking events respectively. The addition of the binary smoking frequency w_k in the random effects then moderates this relationship: α_{00} and α_{01} estimate the variation in mood for frequent and non-frequent smokers during non-smoking events, and α_{10} and α_{11} estimate the variation in mood for frequent and non-frequent smokers during smoking events. Thus, for example, the negative values for α_{11} in their results indicate that "smoking related mood response … is significantly decreased for more frequent smokers, relative to less frequent smokers".

2.1.5 Level 1 covariance structures

2.1.5.1 Autocorrelation

As previously described, one of the assumptions of multilevel models is that residuals within-subject are uncorrelated. This assumption may be relaxed and correlation at level 1 can be estimated.

Serial autocorrelation occurs when successive observations are correlated, with observations taken closer more highly correlated than observations further apart. This is a likely feature of ESM data where many observations are taken over a short period of time.

In a commonly cited paper for ESM studies, Schwartz and Stone (1998) discuss the need to model autocorrelation in ESM data, however in the papers reviewed for this thesis (Chapter 3) only 16% were found to model autocorrelation. Though not often adopted in

ESM studies, modelling autocorrelation is common in similar data structures such as longitudinal analysis and time series analysis where methods include autoregressive models or dynamic models. Both model the outcome at time t as a function of the lagged outcome, i.e. outcome at $t-1$, and a set of covariates. This will be discussed in detail in Chapter 6.

A different approach widely adopted in the longitudinal literature models the autocorrelation within the level 1 errors. In a two-level random intercept model,

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + u_k + e_{ik}$$

the level 1 errors are considered to be normally distributed with mean zero and variance σ_e^2. To allow for correlation at level 1, an autoregressive covariance structure can instead be applied. Here the adjacent observations are correlated with a value ρ, those one step away correlated by ρ^2, two steps away by ρ^3 and so on. This structure assumes equally spaced time points and allows correlations to diminish as the time gap increases, albeit as a function of the original ρ. Diggle (1988) proposed a covariance structure for continuous time which might be more appropriate for ESM studies when observations are randomly spaced. Goldstein, Healy et al. (1994) generalized Diggle's work, expressing the covariance between level 1 errors at time point t and $t-s$ in the form

$$cov(e_t, e_{t-s}) = \sigma_e^2 f(\alpha, s) = \sigma_e^2 \exp\big(-g(\alpha, s)\big)$$

where "$g(\alpha, s)$ is any positive increasing function of s, not necessarily linear, and α is a vector of p parameters",

So, for a first order autocorrelation (AR(1)) model,

$$g(\alpha, s) = \alpha s$$

and so

$$cov(e_t, e_{t-s}) = \sigma_e^2 \exp(-\alpha s)$$

where s is the time interval between observations and α is to be estimated.

Other covariance structures which might suitably model ESM data's potential autocorrelation include the qth order model AR(q), for $q > 1$, as Skrondal and Rabe-Hesketh (2007) point out that the AR(1) structure is "often unrealistic" as "correlations fall off too rapidly with increasing time-lags". Alternatively, there exists a heterogeneous autoregressive structure (HAR) in which the diagonal elements, the variances, are allowed to be heterogeneous. The Toeplitz covariance structure gives further freedom, not restricting the correlations between observations to strictly diminish as a proportion of the previous correlation. Examples of each of these covariance structures are given below for 4 time points.

$$AR(1)\colon \Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix} HAR\colon \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_2\sigma_1\rho & \sigma_3\sigma_1\rho^2 & \sigma_4\sigma_1\rho^3 \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_3\sigma_2\rho & \sigma_4\sigma_2\rho^2 \\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2 & \sigma_4\sigma_3\rho \\ \sigma_1\sigma_4\rho^3 & \sigma_2\sigma_4\rho^2 & \sigma_3\sigma_4\rho & \sigma_4^2 \end{bmatrix}$$

$$Toplitz\colon \Sigma = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

2.1.5.2 Covariance misspecification

The importance of a correctly specified covariance structure has been widely discussed. Kwok, West et al. (2007) examine the impact of misspecification in terms of under-specification, over-specification and general misspecification. These definitions are based on the principle of nested covariance structures and they use the identity (ID), AR(1), first order autoregression with first order moving average (ARMA(1,1)) and second banded Toeplitz (TOEP(2)) covariance structures as examples.

$$ID\ \colon \Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad AR(1)\colon \Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

$$ARMA(1,1)\colon \Sigma = \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\,\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix} \quad TOEP(2)\colon \Sigma = \begin{bmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{bmatrix}$$

The authors define nested covariance structures by “whether one can obtain a specific Σ matrix by imposing constraint(s) on another Σ matrix”. In the covariance structures listed above it can be seen that ID is nested within AR(1) (setting $\rho = 0$), which is nested within ARMA(1,1) (setting $\gamma = \rho$). ID is also nested within TOEP(2) by setting $\sigma_1 = 0$. Under-specification is defined within nested covariance matrices as when a more simplistic covariance matrix is used than the true structure, for example when the true structure is AR(1) but an ID matrix is used. In this situation standard errors of the fixed effects have been found to be positively biased (Ferron, Dailey et al. 2002) as well as the variance of the random effects (Sivo, Fan et al. 2005). Over-specification occurs within nested models when a more complex covariance matrix is used and general misspecification occurs when the true covariance matrix and the assumed covariance matrix are not nested, for example when the true structure is TOEP(2) and AR(1) is used. In their simulation study, Kwok, West et al. (2007) found that under-specification and general misspecification lead to overestimation in the variance of the random effects and the standard errors of the fixed effects. However, they found no significant bias in over-specified covariance structures.

2.2 Motivating example: Recovery in psychosis (Bentall et al.)

The following sections describe the data sets that will be used to illustrate statistical methods proposed in the thesis. The recovery study will be the main example referred to throughout and so will be described in detail. Following an outline of the study design and description of diary items, the research hypotheses and corresponding statistical analyses will be presented, demonstrating how the multilevel models introduced above can be applied to answer specific ESM related questions.

In this study ESM was used to monitor self-reported feelings of 'recovery' in participants with psychosis. The team opted for a paper booklet method of ESM with a pseudo-random design. The participants were signalled to complete a diary 10 times a day for 6 days, the resulting data was of a 3 level structure with moments nested with days within participants. Data were gathered on 134 participants, 40 of which were control participants with no diagnosis of schizophrenia. The control participants were not included in this study as the recovery items were not relevant to this group. Participants were required to complete at least 20 diary entries over the 6 day period to be considered valid.

This study aimed to investigate self-reported 'recovery' as defined subjectively rather than by the specific reduction of symptoms. As such, three diary items were included to capture the construct: "Since the last beep…"

- I felt limited by psychological problems
- I have worries about psychiatric problems
- I have felt mentally well.

The diary also contained momentary measures of mood and psychological symptoms. These items were phrased "Right now…" to capture feelings at the present moment, rather than "Since the last beep…". Measures included current feelings of self-esteem (four items) and paranoia (three items) as well as visual and auditory hallucinations. These items were adapted from previous work by the authors (Thewissen, Bentall et al. 2008; Oorschot, Lataster et al. 2012; Udachina, Varese et al. 2012). Following each paranoia item was an item quantifying how deserving of the paranoid thoughts they felt. These three items were combined to create one measure of deservedness of paranoia. Following the hallucination items was a measure of how pleasant the subject felt these hallucinations were, graded from unpleasant to pleasant.

The items were all measured on 7 point Likert scales. Multiple items relating to each construct were combined using prorated means to create a single score for each measure. Items were first transformed such that higher scores relate to stronger symptoms. For example, higher scores relate to a greater sense of self-esteem or stronger feelings of paranoia. The full diary questionnaire can be seen in Appendix 1.

Finally, recovery was also defined at baseline using the Process of Recovery from Psychosis Questionnaire (QPR). Participants rated each of the 15 items on a 5 point Likert scale “strongly disagree” to “strongly agree”. They were also asked whether they considered themselves recovered as a simple binary yes or no.

There were three primary research questions:

1) How do fluctuations in momentary recovery differ between baseline defined recovery groups?
2) Are there any associations between momentary recovery and the variables self-esteem, hopelessness, paranoia, deservedness of hallucinations, either concurrently or over time?
3) Do any ESM measured items predict a subsequent change in recovery?

2.2.1 STATISTICAL MODELS

At baseline, recovery was measured as both a binary variable classifying the participants as either recovered or not recovered and using a continuous scale. Research question 1 was addressed first for the binary baseline recovery, using a 3 level multilevel model with baseline recovery as a fixed effect and allowing a different residual variance for the two baseline recovery groups. This model was used to estimate separate variance components for the two baseline groups at level 1, representing the fluctuation, or variation, in momentary recovery scores. This model can be expressed by the following equation

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_k + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1)$$

where y_{ijk} is the ESM recovery variable measured at moment i on day j for participant k, x_k is the binary baseline recovery variable measured only once for participant k with $x_k(0)$ representing the non-recovered group and $x_k(1)$ the recovered group at baseline. β_0 and β_1 are the fixed effects to be estimated, v_{jk} the random day effects and e_{0ijk} and e_{1ijk} the split residuals for the two baseline recovery groups. The estimates of the variances of e_{0ijk} and e_{1ijk} were of primary interest, to compare the variation in momentary recovery between the two groups.

For the continuous baseline recovery measure, the model was as above with continuous baseline recovery as the fixed effect, which was split into tertiles for the random part of the model to allow the residuals to vary by group. Group 1 included total baseline scores of 30-49, Group 2 scores of 50-56 and Group 3 scores of 57-73.

For research question 2, concurrent associations were tested using separate three-level models with the variables of interest included as fixed effects, controlling for binary baseline recovery x_k. For this analysis the covariates and the outcome were both measured at the same time point.

To evaluate temporal associations the same model was used but with lagged covariates,

$$y_{ijk} = \beta_0 + \beta_1 x_{i-1,jk} + \beta_2 x_k + u_k + v_{jk} + e_{ijk},$$

where the time lag was restricted to be within each day. Thus, in this model covariates measured at moment $i - 1$ are predicting outcome at the following moment, i.

For question 3 the change in outcome was calculated as recovery at the current moment i minus recovery at moment $i - 1$,

$$y_{ch} = y_{ijk} - y_{i-1,jk}$$

for each participant, restricting the lag to be within day as recovery at the first moment of the next day is not expected to be predicted by a measurement from the day before. The predictors were also lagged so that the results could be interpreted as the effect of the predictor at moment $i - 1$ on the change from the recovery score at moment $i - 1$ to moment i, with the analysis again controlling for baseline recovery x_k. Thus the model was of the form

$$y_{ch} = \beta_0 + \beta_1 x_{i-1,jk} + \beta_2 x_k + u_k + v_{jk} + e_{ijk}.$$

2.2.2 Results of recovery data analysis

2.2.2.1 Is there a relationship between baseline recovery and fluctuations in recovery?

Momentary variation or 'fluctuation' in recovery was compared in the baseline recovered and non-recovered groups using a complex level 1 variation model. The results are presented in Table 2:1.

	Fixed Effects			Random Effects			
	Coeff.	Std. Error	P value			Variance	Std. Error
Baseline Recovery	0.925	0.343	0.007	Subject		1.522	0.290
				Day		0.225	0.027
				Beep	Not recovered	0.530	0.020
					Recovered	0.281	0.016

Table 2:1 Complex variation model comparing momentary recovery between baseline groups

The fixed effects estimate of $\beta_1 = 0.925$ implies that, on average, the baseline recovered group had significantly higher momentary recovery scores than the baseline non-recovered group. The residual variance estimates suggest that there is more variation, or fluctuation, in momentary recovery in the non-recovered group than the baseline defined recovered group.

	Fixed Effects			Random Effects			
	Coeff.	Std. Error	P value			Variance	Std. Error
Baseline Recovery QPR	0.086	0.015	<0.001	Subject		1.043	0.199
				Day		0.217	0.028
				Beep	Group 1	0.539	0.027
					Group 2	0.491	0.025
					Group 3	0.368	0.022

Table 2:2 Complex variation model comparing momentary recovery for continuous QPR scale

In a similar pattern to the binary baseline recovery variable, there is greater variation in momentary recovery for those with lower QPR scores than higher QPR scores.

2.2.2.2 Are the 'Right Now' variables correlated with recovery?

Table 2:3 presents the results of the univariate concurrent association models between the diary recorded 'Right now' variables and recovery.

	Fixed Effects			Random Effects			
Covariate x_{ijk}	Coeff.	Std. Error	P value		Variance	Std. Error	N
Self Esteem	0.325	0.022	<0.001	Subject	1.009	0.195	2333
				Day	0.183	0.023	
				Beep	0.420	0.013	
Hopelessness	-0.303	0.020	<0.001	Subject	1.009	0.194	2298
				Day	0.175	0.022	
				Beep	0.19	0.013	
Visual hallucinations	-0.122	0.024	<0.001	Subject	1.300	0.248	2236
				Day	0.177	0.023	
				Beep	0.444	0.014	
These are pleasant	0.165	0.033	<0.001	Subject	1.268	0.398	347
				Day	0.068	0.028	
				Beep	0.327	0.028	
Auditory hallucinations	-0.406	0.019	<0.001	Subject	1.258	0.243	2203
				Day	0.176	0.023	
				Beep	0.443	0.014	
These are pleasant	0.128	0.031	<0.001	Subject	1.346	0.368	689
				Day	0.196	0.041	
				Beep	0.427	0.026	
Paranoia	-0.378	0.023	<0.001	Subject	0.709	0.138	2335
				Day	0.155	0.020	
				Beep	0.414	0.013	
Deservedness	-0.168	0.031	<0.001	Subject	1.269	0.339	913
				Day	0.324	0.053	
				Beep	0.466	0.024	

Table 2:3 Concurrent associations of Right now variables and recovery

There are significant associations between each of the 'Right now' variables and momentary recovery at the 1% level. Higher levels of momentary self-esteem are associated with higher recovery while higher levels of hopelessness, hallucinations, paranoia and deservedness are associated with lower recovery scores. However, the more pleasant the hallucinations are reported to be the more recovered the participant feels.

The variance estimates for the random effect for each model suggest that the majority of variation in recovery scores is between-person.

There were significantly fewer observations used for the appraisal of hallucination items (N=323 and N=691). These items were only required to be completed if the subject was currently experiencing visual or auditory hallucinations, thus there was a large amount of missing data. Observations were further reclassified as missing unless the hallucination items were scored as 2 or higher (indicating a current hallucination). Similarly, feelings of deservedness were only valid if the subject was currently experiencing paranoia, defined as scoring a 2 or higher on the corresponding paranoia item, resulting in fewer observations for this variable.

2.2.2.3 Is there a temporal association between the Right now variables and recovery?

The results of the univariate lagged analysis between the Right now variables and recovery are presented in Table 2:4.

Comparing the results of Table 2:4 to Table 2:3, the lagged relationships between the Right now variables and recovery were weaker than the concurrent association. The associations were largely significant at the 5% level, however the appraisal items of hallucinations and paranoia were not significantly associated with recovery.

	Fixed Effects			Random Effects			
Covariate($x_{i-1,jk}$)	Coeff.	Std. Error	P value		Variance	Std. Error	N
Self Esteem	0.144	0.026	<0.001	Subject	1.164	0.226	1824
				Day	0.187	0.025	
				Beep	0.418	0.015	
Hopelessness	-0.124	0.022	<0.001	Subject	1.190	0.230	1793
				Day	0.180	0.024	
				Beep	0.418	0.015	
Visual hallucinations	-0.056	0.028	0.042	Subject	1.288	0.248	1748
				Day	0.164	0.023	
				Beep	0.423	0.016	
These are pleasant	-0.030	0.037	0.408	Subject	1.020	0.344	272
				Day	0.130	0.049	
				Beep	0.293	0.029	
Auditory hallucinations	-0.072	0.021	0.001	Subject	1.227	0.239	1730
				Day	0.165	0.023	
				Beep	0.422	0.016	
These are pleasant	0.023	0.034	0.506	Subject	1.375	0.390	542
				Day	0.180	0.042	
				Beep	0.422	0.029	
Paranoia	-0.176	0.026	<0.001	Subject	0.983	0.196	1826
				Day	0.175	0.024	
				Beep	0.421	0.015	
Deservedness	-0.008	0.040	0.831	Subject	1.500	0.411	691
				Day	0.324	0.060	
				Beep	0.475	0.029	

Table 2:4 Lagged analysis between the Right now variables and recovery

2.2.2.4 What predicts change in recovery?

Change in recovery was calculated as the difference between recovery at moment $i-1$ and recovery at moment i. This score was restricted to changes within each day, i.e. change in recovery is not calculated between beep 10 of one day and beep 1 of the next day. Binary baseline recovery is again included as a covariate in each model.

The results of the univariate change model are presented in Table 2:5.

	Fixed Effects			Random Effects			
Covariate ($x_{i-1,jk}$)	Coeff.	SE	P value	Level	Variance	SE	N
Self Esteem	-0.039	0.015	0.011	Person	0.000	0.000	1910
				Day	0.000	0.000	
				Beep	0.752	0.026	
Hopelessness	0.045	0.014	0.001	Person	0.000	0.000	1879
				Day	0.000	0.000	
				Beep	0.751	0.025	
Visual Hallucinations	0.024	0.019	0.204	Person	0.000	0.000	1834
				Day	0.000	0.000	
				Beep	0.763	0.026	
These are pleasant	-0.121	0.032	<0.001	Person	0.000	0.000	276
				Day	0.000	0.000	
				Beep	0.636	0.054	
Auditory Hallucinations	0.000	0.011	0.998	Person	0.000	0.000	1805
				Day	0.000	0.000	
				Beep	0.759	0.025	
These are pleasant	-0.038	0.022	0.086	Person	0.000	0.000	540
				Day	0.000	0.000	
				Beep	0.725	0.050	
Paranoia	0.031	0.012	0.008	Person	0.000	0.000	1912
				Day	0.000	0.000	
				Beep	0.752	0.024	
Deservedness	0.011	0.020	0.588	Person	0.000	0.000	715
				Day	0.000	0.000	
				Beep	0.805	0.043	

Table 2:5 Change models for Right now variables and recovery

The fixed effect estimates for these models present unusual directions in the results. Increases of self-esteem, in which higher scores relate to higher levels of self-esteem, were associated with a decrease in recovery at the following time point. Conversely, increases in negatively oriented measures hopelessness and paranoia were associated with an increase in recovery, that is, a greater feeling of hopelessness and feeling more paranoid was followed by a greater feeling of recovery. These associations are in the opposite direction of that anticipated. Furthermore, the subject-level and day-level random effect variances were reduced to zero in these models. These two issues will be discussed further in Chapter 6.

2.3 Motivating example: Bipolar disorder and cannabis use (Tyler et al. 2015)

The primary objective of this study was to investigate how cannabis use affects symptoms and mood in individuals with bipolar disorder. Specifically, the authors identified 2 main hypotheses:

1. " frequency of cannabis use will increase as a function of affect and BD [bipolar disorder] symptom change (i.e. self-medication effects)
2. cannabis use will be associated with subsequent changes in affect and BD [bipolar disorder] symptoms"

Twenty-nine participants were recruited into the study, however analysis was conducted on a sample for 24: two participants dropped out due to personal reasons and three were excluded as they had completed fewer than 20 reports. This exclusion criterion was adopted as the validity of assessment may be compromised when less than a third of responses are returned (Palmier-Claus, Myin-Germeys et al. 2011). Participants completed 10 paper diaries a day at "unpredictable times" (pseudo-random intervals) when alerted by a digital wristwatch. A response was required within 15 minutes of the signal to be considered a reliable assessment of the current moment.

Each diary contained 10 items relating to affect, such as 'Right now I feel cheerful' and 'Right now I feel lonely'. Items were rated on a 7-point Likert scales with 1 = 'not at all' and 7 = 'very much so'. Five items were combined with prorated means to create a positive affect score and five to create a negative affect score. Bipolar symptoms were also measured in the diary using Likert scales with items including "Right now I feel full of energy" and "Right now I feel bad about myself". Three items were averaged to create a mania score, and four items were used to measure depression. Cannabis use was a binary variable recorded as "since the last beep I've used cannabis" 1=Yes or 0=No.

This data set will be used in addition to the recovery data set to illustrate methods presented throughout the thesis. The recovery data set, however, will feature as the primary motivating example and as such further particulars of the cannabis and bipolar analysis will not be presented here. Details of the models and results can be found in our published paper Tyler, Jones et al. (2015).

3 SYSTEMATIC REVIEW

3.1 AIMS OF REVIEW

A systematic review was conducted to determine the methodology being reported in studies using ESM. Unlike a traditional systematic review which might focus on the results of subset of papers in a particular field, the intentions of this review were to investigate which fields of research use ESM and examine their statistical methods and quality of their statistical reporting. The primary objective was to record what type of research questions were being investigated and which statistical methods were being used to analyse the data. Furthermore, the review aimed to establish the consistency of reporting in ESM studies in terms of design and analysis. Also of interest was whether any justification for sample size was provided. The data extraction form can be found in Appendix 2: Systematic review.

3.2 DESIGN OF REVIEW

The databases PubMed and Medline were used to search for published articles using the keywords “experience sampling” or “ecological momentary” in the title or abstract. ‘Ecological momentary’ was used as a search term as ESM is also referred to Ecological Momentary Assessment (EMA) (Shiffman and Stone 1998). The search was refined to exclude ‘review’ publication types and, due to the limited timeframe to conduct the review, to only include English language papers. These search criteria returned a total of 573 references from Medline and 641 references from PubMed (search carried out on 11/2/13). The two lists of references were combined in Endnote and duplicates were automatically removed. A manual screening of the references identified duplicates missed by Endnote and resulted in final sample of 659 papers.

Titles and abstracts were then screened and refined by a series of exclusion criteria:

1) Articles must be published journal articles, rather than published abstracts.
2) Exclude reviews or meta-analyses of ESM studies
3) Articles should describe a study using ESM, i.e. exclude papers only discussing the concepts or benefits of ESM, or discussing it as a method to use in future work.
4) Of the studies using ESM, exclude:
    a. pilot studies
    b. studies where ESM is not the primary focus of the study

c. studies where ESM is used to validate a measure or comparing ESM items against a traditional measure

d. those assessing feasibility of a future study or investigating compliance to the procedure

e. those used to assess recall bias.

The number excluded due to each of these criteria can be seen in Figure 3:1.

The exclusion criteria were chosen to ensure the final sample of papers would be comparable in terms of the information they contain regarding the ESM procedure and the models used for analysis. Pilot studies, though containing details on ESM procedure, may not analyse the data gathered. Similarly feasibility studies where the focus would more likely be on compliance. Studies where ESM is not the primary focus may not give sufficient details about the ESM procedure or analyse the ESM results. Where ESM was used to validate measures or to be compared to a singularly administered questionnaire, the analysis would likely aggregate the ESM data to compare to the other questionnaire – this type of analysis is not of interest for this work. Likewise, studies addressing recall bias were excluded as these studies would also be expected to aggregate ESM data to compare with a questionnaire.

3.3 Results of systematic review

Once the abstracts had been screened and studies excluded under these conditions, there were 459 eligible studies. Within the four month timeframe allocated for this review, reading and extracting data from all these papers would have not been possible so the search was further refined to only include papers published in 2012. This would capture papers ideally presenting the most up to date methods for analysing data of this structure. Although this refinement would likely limit the results to studies using multilevel models, and thus not pick up on the change in statistical methodology applied to this type of data, it would not detract too much from the original intentions of this review. Instead of investigating the range of methods used for analysis, knowing that multilevel models are better suited to ESM data than regular linear regression, say, the focus will be on the studies not using this methodology and the details provided by those that do. This left 91 eligible studies

There were 49 studies due to be screened for a second time to determine eligibility before the full text screening. This group does not include any 2012 papers and so does not affect the results of this review.

After reviewing the full papers, two were excluded as ineligible as defined by points (1-4) above and 16 were excluded as they were not print publications in 2012; some papers published online in 2012 were not initially excluded. The final sample of papers included in the review was 74. This final sample contains 74 studies rather than 73 as one paper (LePage, Price et al. 2012) contains the results of two different ESM studies and so these studies have been recorded separately.

Screening and data extraction was only carried out by one reviewer. While it is advisable that a proportion of papers are double entered to identify inconsistencies in classification or data extraction, this was not feasible in the current study. It is argued that repeatability of this process is not a major concern in the present research as this review was conducted in order to aid the investigation of contemporary methods of ESM, rather than to provide a comprehensive summary of all published ESM studies.

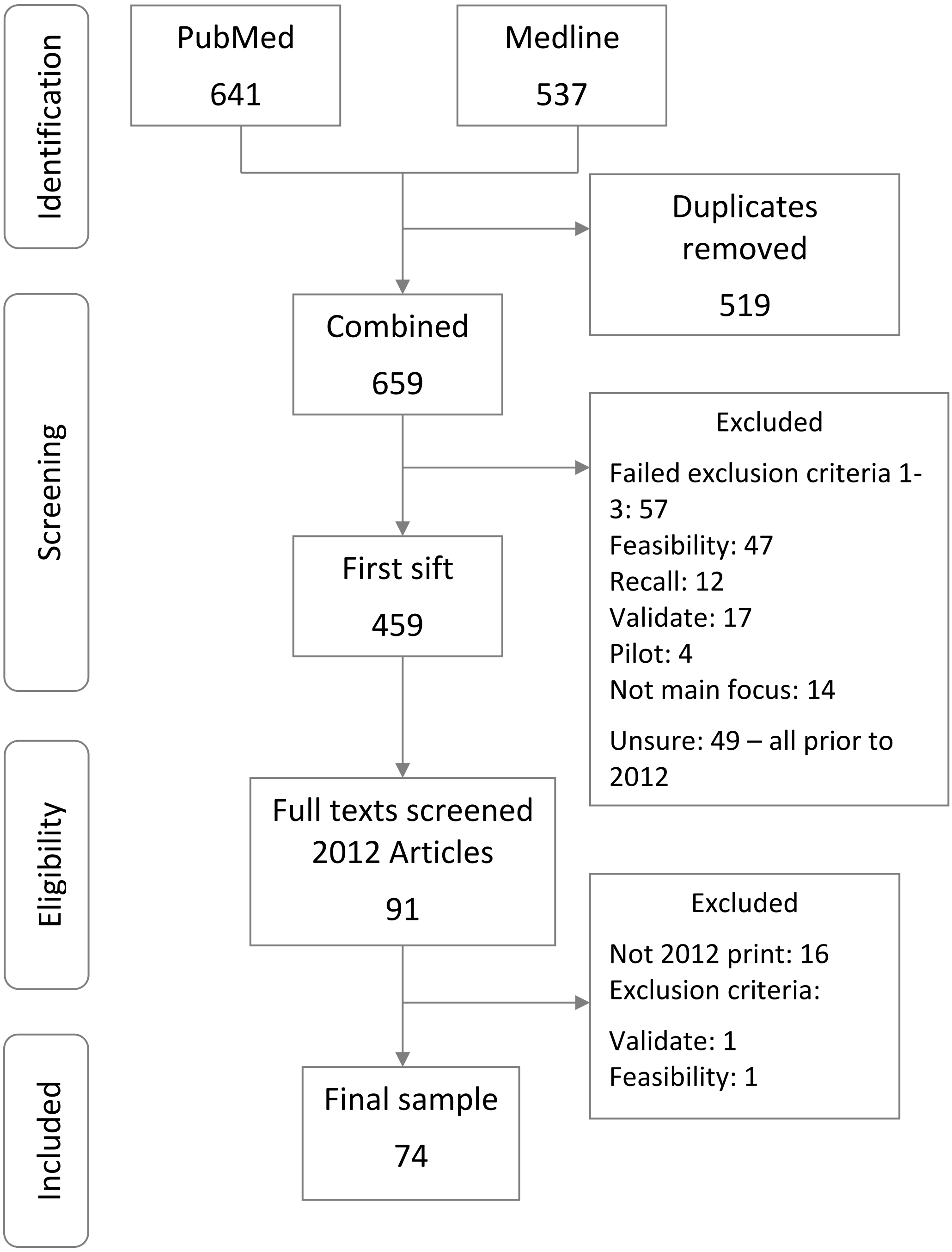


Figure 3:1 Systematic review flow of studies diagram

3.3.1 OVERVIEW OF STUDIES

The final papers were from a broad range of research areas and covered all ESM designs. Though mostly within the field of psychology other areas were represented, for example two papers used ESM to monitor biological mechanisms and three used the method to study flow, or optimal experience. A full list of research areas is given in Appendix 2: Systematic review.

Only four studies (5.48%) were classed as a randomised trial, though 22 (29.73%) were described as having used a control or a comparison group, for example drug users and non-drug users.

The majority of studies used either a paper diary (29.73%) or a type of handheld computer (or PDA) (50%). Seven studies used a phone to deliver questions: either by text message (1.35%), calling participants (5.41%) or using a programmed application on smart phone (2.70%). Four studies (5.41%) asked participants to complete an online questionnaire as the ESM method. Only one paper (Bolt, Piper et al. 2012) gave no description of method, referring only to “ESM reports”.

The ESM design was classified as either random, interval or event contingent. Many studies used a combination design: six as either event plus random (8.11%), two interval plus random (2.70%), five event plus interval (6.76%) and five using all three prompts (6.76%). Of those employing a single contingent design, the majority chose a random design (50%), which was defined as those specifying a random or stratified block random structure. Twelve used an interval design (12.16%) which also included those specifying a random structure around fixed time points, for example up to 10 minutes either side fixed points, on guidance of Delespaul (1995). Only four studies used an event only design (5.41%).

ESM can be used in conjunction with ambulatory assessment to record physical symptoms alongside psychological symptoms and behaviours. Only seven studies from the review recorded using a form of ambulatory assessment: three using an accelerometer (4.05%), one measuring ambulatory blood pressure (1.35%), one using an electrocardiogram to measure heart rate variability (1.35%) and two collecting saliva samples to monitor cortisol levels (2.70%).

	N	%
Randomised trial		
Yes	4	5.41
No	70	94.59
Control group used		
Yes	22	29.73
No	52	70.27
Data collection method		
Paper booklet	22	29.73
PDA	37	50.00
Phone – text	1	1.35
Phone – call	4	5.41
Phone – app	2	2.70
Internet	4	5.41
Other	3	4.05
No method reported	1	1.35
ESM design		
Event (E)	4	5.41
Interval (I)	9	12.16
Random (R)	37	50.00
Mix - E&R	6	8.11
Mix - I&R	2	2.70
Mix - E&I	5	6.76
Mix - I, R & E	5	6.76
No design reported	6	8.11
Ambulatory assessment (AA) used		
Yes	7	9.46
No	76	90.45
AA type		
Accelerometer	3	4.05
bloody pressure monitor	1	1.35
Electrocardiogram	1	1.35
Saliva sample	2	2.70

Table 3:1 Overview of published 2012 studies

3.3.2 Sample size and adherence to protocol

The number of participants ranged between 13 to 1504, with a mean of 169.8 (SD 255.03), median 85. Participants were monitored for an average of 13.4 days (range 1-175), though the majority of studies used a 7 day or ’1 week’ measurement period (22.97 %). The number of measurement taken per day was harder to define in event contingent or mixed designs and when the number of measurements varied per day or were dependent on previous responses. 47 studies gave a clear description of the number of measurements taken per day, with all but one study using between 1 and 12 measurements; only Koval and Kuppens (2012) used more, requesting participants complete 60 measures a day for two days.

Only one study (Myers, Ridolfi et al. 2012) mentioned any type of sample size calculation. Though not a formal calculation formula, Myers provided a justification for the sample size and number of observations chosen and conducted a post hoc power analysis using the software PINT (Power in two-level designs) (Snijders and Bosker 1993) to confirm sufficient power.

Compliance and missing data were unclearly defined in studies. For mixed designs adherence was gathered on either the random or interval prompts. Data was either not available or not clear for 33 studies, however the remaining studies had an average (median) compliance of 80.2%, ranging from 30% to ‘over 99%’ (Bruehl, Liu et al. 2012). Of those acknowledging missing data (63.51%) only five studies indicated the method used to address this issue: Giesbrecht, Campbell et al. (2012) estimates missing data using “full information maximum likelihood”, Forbes, Stepp et al. (2012) note that the expectation maximization (EM) algorithm used to estimate their linear growth curve model can accommodate missing data, Mak, Prynne et al. (2012) use a “complete case analysis”, while Elavsky, Molenaar et al. (2012) states “the standard time series technique” for missing data was used.

Compensation was recorded to provide a measure of incentive to comply with protocol. Compensation was defined as either monetary payment or course credit for student populations. It was also recorded if additional compensation was given for complying with protocol and returning a predetermined number of questionnaires. 51.35% of studies offered some form of compensation with 16 studies (21.62%) offering further rewards for high compliance. Comparing this with the adherence figures, mean percentage compliance for those receiving and not receiving compensation for participating in the study was 73.3% and 78.5% respectively. A t-test determined no significant difference between the two groups (p value = 0.367). Similarly, compliance rates for those receiving additional rewards for high levels of completion were not statistically different than those not receiving extra (mean difference -1.94, p value=0.740). These results indicate that it might not be necessary to offer compensation to secure high compliance, an interesting contrast to results of other studies (for a review see Morren, van Dulmen et al. (2009)).

3.3.3 Factor analysis and Cronbach's alpha

Eleven studies (14.9%) carried out a factor analysis on ESM items, only one of which (Yeh, McCarthy et al. 2012) accounted for the multilevel structure. However, in most cases this was purely confirmatory, with only Menne-Lothmann, Jacobs et al. (2012) weighting the items according to their factor loadings, potentially biasing their results. A further six studies used a principal components analysis to group ESM items and twenty six (35.1%) studies used Cronbach's alpha to justify subscales created from ESM items, none commenting on how the repeated measures might affect estimates of alpha.

3.3.4 Types of research questions

Research questions were recorded and the questions categorised into the following classifications: association; temporal association; group differences; response to treatment; predictors or risk factors; mediation and moderation. As studies often define multiple research questions per paper, the two most prominent ques6tions were recorded for each study if more than one were outlined. These have all been grouped in Table 3:2 to give an indication of all research questions. Frequencies are given but without percentages as only 46 studies defined two questions.

Research Question	Frequency
Association	46
Temporal association	9
Group difference	11
Response to treatment	3
Predictors or risk factors	4
Mediation	7
Moderation	24
Change in outcome	3
Variation in outcome	5
Other	5

Table 3:2 Research questions identified in the review

In the ‘other’ category questions include using ESM to monitor or count events (3), to measure time to relapse (1) and using instability in the explanatory variable to predict outcome (1).

Eight studies investigated temporal associations. Of these, four examined how mood changed in response to an event. Wichers, Peeters et al. (2012) defined this event as the largest within-day increase in positive affect and Wichers, Lothmann et al. (2012) as an increase in physical activity between two moments $(t-1, t)$, while Munsch, Meyer et al. (2012) and Muller, Mitchell et al. (2012) investigated how mood changed in the time preceding and following binge eating and compulsive buying events. These studies used multilevel models to examine how mood changes as a function of time, measured as hours or moments from the event. Buckner, Crosby et al. (2012) uses a simpler method, lagging the covariate of interest to see how x at time $t-1$ effects outcome at the following moment t. Similarly, Elavsky, Molenaar et al. (2012) uses lagged covariates within a cross-lagged dynamic model, treating the ESM data as time series to study within-subject variation with separate analyses for each individual. Oorschot, Lataster et al. (2012) also treated the data as time series using a vector autoregressive model, a multivariate time series technique used to study bidirectional associations. Finally, Shiyko, Lanza et al. (2012) uses a time-varying effects model.

Variation or instability of outcome was analysed in five studies, each defining variation in different ways. Udachina, Varese et al. (2012) and Peters, Lataster et al. (2012) both calculated the difference between successive observations, Peters using the absolute difference. An average of these differences was then taken across all time points, providing a single variability score for each participant. This method was also used by Palmier-Claus, Taylor et al. (2012), where the difference was squared, when calculating instability in a covariate used to predict outcome. Selby, Doyle et al. (2012) and Demiralp, Thompson et al. (2012) use more standard measures of variability: Selby et al calculating the standard deviation for the daily average of ESM reports, resulting in 2 level data which was analysed using MLM, while Demiralp et al. used the variance of the average outcome measured across all time points and a 2-way ANOVA to test for group differences. Finally, McCabe and Fleeson (2012) discussed the 'within-person and between-person variation' using the ICC and although referred to predicting variation in outcome appeared to predict an unadjusted ESM variable as outcome using a multilevel model.

Change in outcome was also addressed differently by each study. Mata, Thompson et al. (2012) used a change score ($y_{ij(t+1)} - y_{ij(t)}$) as outcome where as Kuppens, Champagne et al. (2012) used a lagged outcome variable as a covariate, referring to the model as an 'autocorrelation-crosscorrelation regression model'. Giesbrecht, Letourneau et al. (2012) were interested in how positive and negative affect changed over the course of pregnancy, which was modelled using a quadratic function of gestational age in a multilevel model.

3.3.5 Analysis method

The method of analysis was recorded from each paper along with any details of the model specified. As expected, the majority of papers used some form of multilevel or random effects model. For simplicity, variation of 'multilevel models', 'linear mixed models' and 'random effects models' will all be categorised as 'multilevel models'. As with research questions, multiple analysis models were used in each paper. The primary analysis model determined from each paper is listed in Table 3:3.

Though the primary analysis for over two thirds of studies (70.27%) uses multilevel models, some studies employed methods which underutilize the multilevel structure of ESM data. Seven studies chose to aggregate their data to the person level and use methods suitable for single level data.

Method	N	%
Multilevel model	52	70.27
2-way ANOVA	1	1.35
T-test	1	1.35
Correlation	2	2.70
Regression	4	5.41
Regression controlling for clusters	1	1.35
Generalized estimating equations	3	4.05
Latent growth curve model	1	1.35
Latent variable model	1	1.35
Structural equation model	2	2.70
Mixed design ANOVA	1	1.35
Experience fluctuation model	2	2.70
Time Varying Effect Model	1	1.35
Vector autoregressive (VAR) modelling	1	1.35
Times series model	1	1.35

Table 3:3 Primary analysis method identified in the review

Of those fitting multilevel models, the majority used random intercept models (67.3%). Only ten of these studies specifically stated the use of a random intercept model, the other 25 were presumed to be a random intercept model as no other random effects were mentioned. Of the remaining, twelve (23.1%) were random coefficient models and three (5.8%) were random slope models (where time, rather than a measured covariate was allowed to vary). One study (Giesbrecht, Letourneau et al. 2012) specified a multilevel model but that "results are based on estimation of fixed effects with robust standard errors" and one study was unclear (Cook, Calcagno et al. 2012), stating they "allowed all effects to vary randomly" but later referring only to tests of fixed effects. Most studies using multilevel models analysed their data in a two level structure (86.5%), though a minority specified three levels (13.5%).

3.3.5.1 Lagged Outcome

Six studies using multilevel models include the lagged outcome variable a covariate. Kuppens et al (2012) (Kuppens, Champagne et al. 2012) used the lagged outcome to model change between time $t-1$ and t, whilst Elavsky et al. (2012) (Elavsky, Molenaar et al.

2012) used the lag outcome within a time series analysis to “reflect the stability of [the outcome] across days”. Goldschmidt, Engel et al. (2012) and Ben-Zeev, Young et al. (2012) provide no justification for including the lagged outcome other than to adjust for outcome at the previous time point, while Udachina, Varese et al. (2012) included it as “possible confounder” when examining the relationship between paranoia at time t and self-esteem at time $t+1$. To, Fisher et al. (2012) state that the lagged outcome is included to account for the residual autocorrelation at level 1 and similarly Koval and Kuppens (2012) refer to the random slope of the lagged outcome as an autocorrelation parameter.

Bias due to the violation of the exogeneity assumption may be present when including a lagged outcome as a covariate (as will be discussed in Chapter 6), however, none of studies discussed any potential methodological issues to this effect.

3.3.5.2 Autocorrelation

Only twelve studies (16.2%) attempted to model autocorrelated residuals at level 1. Two studies (Koval and Kuppens 2012; To, Fisher et al. 2012) used lagged outcome variables, Oorschot, Lataster et al. (2012) using a vector autoregressive model and the remainder specifying an alternative level 1 covariance structure. Each of these studies used a first order autoregressive structure, with Schwerdtfeger and Scheel (2012) stipulating a continuous AR(1) structure and Goldschmidt, Engel et al. (2012) a heterogeneous AR(1) structure.

3.4 Discussion of systematic review

The reporting of ESM specific details were fairly consistent; only minimal missing data in terms of prompting and recording (1.4% missing number moments, 2.7% missing number days, 1.4% missing data collection method and 8% missing sampling method), though details on compliance and missing data were low. Whilst all studies reported at least one form of analysis method, those applying multilevel models gave varying levels of information on model specifics, with many not specifying the random effects of the model and even fewer reporting these in the results. Finally, there was a clear lack of sample size justifications or power calculations.

Guidance for conducting and reporting ESM studies is available within the psychological literature (Stone and Shiffman 2002; Palmier-Claus, Myin-Germeys et al. 2011), but perhaps more specific guidelines are needed on power and sample size calculations for ESM studies and how to thoroughly report statistical models used in ESM analysis.

Although the majority of studies used multilevel models, several studies still chose to analyse their data using non-multilevel methods. The limitations of these methods are as follows. Firstly, though appropriate for multilevel data, generalized estimating equations (GEE) provide population average effects, i.e. the effect of x for the average person, rather than within-subject effects as in some multilevel models. Gardiner, Luo et al. (2009) argues that these interpretations are equivalent for linear models, including those specifying an autoregressive covariance structure. However, when using a GEE model to predict a binary outcome the results are not equal to those using a multilevel model and must be interpreted as population specific effects. A benefit of GEE models is they are robust to misspecified covariance structures, however they do not provide estimates of the variation at each level and so can't be used to interpret within- and between-person variability or to investigate complex covariance structures at level 1.

Aggregating ESM data to the subject-level ignored potentially interesting information captured within-subject. Moreover, conducting ANOVA or regression analysis on aggregated data poses problems as the heterogeneity assumption of these models is violated when there is missing data at the lowest level (Schwartz and Stone 1998), as there is sampling variation when data with missing values are averaged at a higher level. This is also a problem for using a mixed design ANOVA, as aggregating incomplete data to higher levels will violate the heterogeneity assumption of this method. Mixed design ANOVA also assumes a compound symmetry covariance structure which may not be appropriate for this data.

This review has revealed that although multilevel models are being used to analyse ESM data, models are not consistent for similar research questions and have the potential to be more statistically sophisticated. Furthermore, statistical issues such as the use of lagged outcomes and misspecified covariance structures are not being considered.

4 MISSING DATA IN ESM RESEARCH

Missing data can occur in all types of research but it can be particularly prevalent in longitudinal studies where participants are subject to multiple follow ups over time. ESM studies are especially vulnerable: participants are required to complete questionnaires unsupervised, multiple times a day over several days, all while continuing with their usual daily routine. Absolute compliance, where all the data is collected as intended, is unlikely, with prompts missed due to the demands of everyday life or as a result of the intensive sampling procedure becoming too burdensome. Although ESM study design typically includes an element of participant training in the data collection method and consent regarding the intensive sampling procedure, the self-reported design means that data quality is entirely subject to the participant's adherence to the study protocol.

This chapter will discuss missing data in ESM research and how to investigate nonresponse in this type of data structure. Firstly, the reporting of missing data in ESM research identified in the systematic review will be discussed, followed by an exploration of missing data in the recovery study data. Missing data methodology will then be introduced, with applications to an ESM setting.

4.1 SYSTEMATIC REVIEW: ADHERENCE TO PROTOCOL AND MISSING DATA

One aim of the systematic review of ESM studies presented in Chapter 3 was to identify how missing data was being reported and to establish the extent of missingness present in this intensive longitudinal data structure.

Overall, missing data was underreported in the ESM studies reviewed, with 28 of the 74 studies (38%) failing to comment on the completeness of the data at all. A further five papers reported that the data were incomplete but did not quantify the amount of missing data. The 41 studies providing data on nonresponse typically reported in terms of adherence to protocol or 'compliance' rather than nonresponse, presenting the percentage of planned prompts which were completed. These studies had a wide range of adherence rates: from 30% to "over 99%" (Bruehl et al, 2012) – this is illustrated in Figure 4:1. It should be noted, however, that these figures include a mixture of reporting styles: some papers presenting total compliance figures, others an average of responses across subjects. When the number of diaries administered per subject is equal these rates are interchangeable, however if the sampling scheme varies for each subject the total and average rates will differ. With a heavy negative skew, the median compliance rate was

80.2%. The observed cut off of 30% is likely due to the common inclusion criteria of requiring each subject to return approximately one third of the prompts to be considered valid; of the studies contributing to these compliance figures, 13 studies (32%) required a minimum number of items or moments completed to be included.

Determining the amount of missing data in an ESM study may be problematic due to the study design. Event based sampling requires a diary to be completed only after the subject experiences a specific event, for example after smoking a cigarette. With this design there is no expected number of diaries to be completed and often no record of any missed events. Of the 33 studies not reporting adherence figures, three used event based designs. A further seven reported using a combination of event plus interval or signalled prompts. Where adherence rates were reported in a mixed design, these were taken to be the percentage of non-event, expected prompts completed.

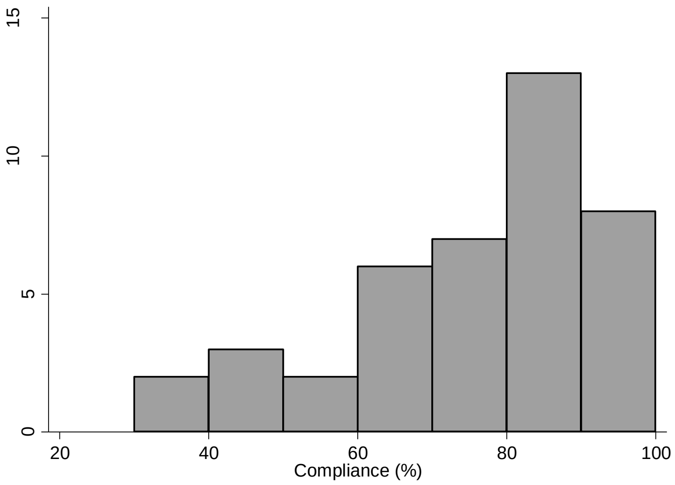


Figure 4:1 Compliance rates in systematic review (n=41)

The definition of compliance or adherence in these ESM papers was ambiguous. Authors refer to the number of 'valid' (e.g. Wichers et al 2012) or 'usable' (e.g. Walsh et al 2012) questionnaires completed without defining these terms. Moreover, the extent of missing items within each diary was not reported in any paper, and so it remains unclear as to whether a 'completed' questionnaire refers to those with at least one item completed, certain specific items completed, or all items completed.

Approaches to missing data were very rarely discussed by the papers in the review. Only five studies expanded on basic compliance figures: Giesbrecht, Campbell et al. (2012)

reported estimating missing data using full information maximum likelihood; Forbes, Stepp et al. (2012) noted that the expectation maximization (EM) algorithm used to estimate their linear growth curve model "handled" the missing data; Mak, Prynne et al. (2012) used a complete case analysis; while Elavsky, Molenaar et al. (2012) stated "the standard time series technique" for missing data was used. Though providing comment on how missing data was accommodated, none of these papers presented details or assumptions beyond the descriptions above, nor commented on any assumptions these methods require. The remaining studies provided no details on how missing data was addressed. No studies reported imputing missing data or investigating missing data mechanisms.

4.1.1 Examples of nonresponse in ESM data

Investigating the cause of missingness in ESM studies does not appear to be widely practiced, however two papers have been published explicitly examined missing data in an ESM context. Though not methodology papers, they both detail procedures for exploring the missing data mechanism. Silvia, Kwapil et al. (2013) focused on how momentarily measured symptom dimensions affect nonresponse in their study of 450 university students, where data were pooled from several smaller studies. Their participants completed eight diaries a day via a PDA for seven days. After each randomly timed alert, participants has five minutes to complete the questionnaire before it was declared a nonresponse. Participants were incentivised with a draw for a $100 gift card for those who responded to at least 70% of signals. The authors generated a binary variable denoting momentary missingness, and regressed this on a series of within- and between-subject factors using two-level logistic models. Within-subject models investigated whether time of day or day of study predicted missingness; in separate models, moment number and day number were entered as linear and quadratic terms to study time trends in missingness. All measures of time were found to be statistically significant, suggesting that both the within- and between-day trends in missing data follow "an inverted-U" pattern with fewer missed signals in the morning and evening, and fewer missed beeps at the start and end of the week. To investigate how emotional states might affect compliance, the authors fitted a two-level model with eight emotions and experiences (such as "Right now I feel happy", "I am currently alone") at the previous beep predicting missingness at the current beep. In this model only one variable was statistically significant – I feel enthusiastic - with an odds ratio $OR = 1.05$ suggesting that higher feelings of enthusiasm were followed with a greater odds of subsequent missing data. Several subject-level variables also predicted nonresponse, with males significantly less likely to respond to prompts than women and subjects with higher scores of positive schizotypy, depression and hypomania measured at baseline less likely to respond to prompts.

Messiah, Grondin et al. (2011) also explored predictors of nonresponse in their study of psychoactive substance use. Five diaries per day were completed on handheld computers by 224 university students at fixed time intervals for seven days. Each questionnaire was available for 45 minutes before being classified as missing. A total of 13.8% of diaries were uncompleted. Two-level logistic regression models were used to investigate predictors of missingness which included subject-level variables such as gender, age, temperament and character scores, as well as within-subject time variables: calendar day, day number and within-day time windows. As in Sylvia et al.'s paper, the outcome of interest was a missed moment, both studies using electronic data collection devices where nonresponse could be defined as an unanswered diary after the prompt. In this study, time was only entered as a linear term and each variable analysed as a categorical variable. A variety of subject level variables were found to significantly predict the increased odds of missing data, including gender (male vs female), degree course and drug consumption. Chronological day number (with day 1 as reference category) indicated a significant decreased odds of nonresponse at days 2, 3, 4 and 6 while within-day fewer diaries were completed in the morning (8am – 11am) compared to the evening (8pm – 11pm).

Both papers emphasise that steps should be taken during data collection to minimise nonresponse, for example meeting with or contacting participants during the study and using convenient data collection devices. Where predictors of nonresponse are known, particular attention can be paid to subgroups of participants to encourage adherence to protocol. A second conclusion of each paper was regarding appropriate statistical analysis in the presence of missing data. Silvia et al argued that future studies should control for known predictors of missingness to satisfy the missing at random assumption when using maximum likelihood estimation. Although finding several predictors of nonresponse in their population, the authors suggest that at the very least future studies should consider controlling for time of day in their analysis. Messiah et al, on the other hand, naively contended that their results were evidence of non-ignorable missingness after finding nonresponse was related to the topic of interest: a greater amount of missing data was present for poly-substance drug users, similar to the findings of Litt, Cooney et al. (1998) in their study of alcoholics. However, the fact that they are able to model this relationship with the observed data makes it necessarily ignorable. Including the variable measuring drug use in their model would satisfy the missing at random assumption required for maximum likelihood estimation to be valid, and would not require an alternative modelling procedure as suggested.

4.2 Nonresponse in the recovery example

This section will use the recovery study data to provide an example of missing data in an ESM setting. Suggestions will be given on possible ways to summarize missing ESM data, both numerically and graphically, as well as methods for understanding how missing data varies and how it can be investigated further.

With its complex data structure, ESM has the potential for missing data at several levels; to succinctly summarize this multilevel nonresponse, new terms for categorising missing data in ESM studies will be defined as follows.

Item nonresponse will refer to missing data at the item level, i.e. the proportion of uncompleted questions within a diary. This may be expressed as total item nonresponse – the proportion of missing items overall – or as average item nonresponse, describing the average proportion of missed items within a diary. Succinctly summarizing item nonresponse is challenging. Possible approaches will be presented in this chapter.

Moment nonresponse will refer to missing data at the moment level, that is, the proportion of uncompleted diaries. Again this may be expressed in terms of total moment nonresponse or as an average per subject. The definition of moment nonresponse will depend on what is considered a ‘completed’ diary, to be defined in terms of item nonresponse. An intuitive definition would be to class a diary as uncompleted if all items are missing. However, the reciprocal of this is to define a diary as ‘completed’ if at least one item has been answered. The argument of response validity holds here as when requiring a minimum overall response rate: when only one item has been completed, is the response actually representative of the current experience? If one argues for a minimum response rate overall for validity then an alternative moment nonresponse definition might be required. This does, however, then rely on the researcher’s discretion to define a somewhat arbitrary cut-off point in terms of within-diary item completion.

Day nonresponse will be the proportion of missing data at the day level either overall or by subject, occurring when all dairies are considered missing in a day.

Complete nonresponse is a subject-level definition of missingness, where no diaries are considered complete.

4.2.1 Complete nonresponse

Two subjects returned no data for the entire sampling period. These will be removed from the sample. The following summaries will therefore be the amount of missing data for the remaining 68 subjects who completed at least one item during the six day sampling period

4.2.2 Item nonresponse

Each ESM diary in this study consists of 50 questions designed to establish current behaviours and states of mind. A full summary of item nonresponse can be found in the table of Appendix 3: Missing data. However, while some of these items stand alone, the majority of items were combined to create specific measures. For example, the four items *"I feel cheerful/excited/relaxed/satisfied"* were grouped to create a measure of positive mood, or affect. As these total measures rather than the individual component items were of primary interest, item nonresponse will be defined as an uncomplete measure rather than its component questions. It should be noted, however, that as measures were computed as the pro-rated mean of the item scores, only half of the items per measure were required to produce a mean score. It was thus possible for a measure to be considered 'complete' while some of its component items were missing. Table 4:1 summarizes the proportion of missing data for the main ESM measures. Each column of the table presents the percentage of missing data for each measure at each time point, averaged across all days, provided that at least one item was completed in a diary. The percentages represent the proportion of missing data at each time point, conditional on the diary having being answered. That is, the percentage of items left incomplete for the subset of diaries where at least one item had been answered on any scale at that moment. Partitioning the data in this way helps to differentiate item nonresponse from moment nonresponse. Therefore the resulting data can be used to compare nonresponse across measures by reading down the table, to examine patterns of item nonresponse, as well as illustrating any trends in missing data by reading across the table.

	Beep Number										
	1	2	3	4	5	6	7	8	9	10	Total
	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=408)	(n=4080)
Moment nonresponse	48%	43%	36%	36%	33%	32%	30%	31%	38%	42%	37%
Item nonresponse	(n=213)	(n=232)	(n=263)	(n=261)	(n=275)	(n=279)	(n=285)	(n=280)	(n=254)	(n=236)	(n=2578)
Self-esteem	4%	5%	4%	4%	3%	2%	2%	2%	2%	1%	3%
Hopelessness	8%	6%	5%	5%	5%	5%	4%	3%	4%	1%	5%
Paranoia	4%	4%	3%	3%	3%	1%	1%	1%	1%	1%	2%
Deservedness	62%	60%	65%	62%	64%	60%	61%	64%	63%	60%	62%
(N/A removed)[1]	4%	5%	4%	5%	4%	3%	2%	4%	2%	2%	3%
Visual Hallucinations	7%	8%	8%	7%	8%	5%	6%	5%	4%	2%	6%
Pleasantness	86%	85%	87%	85%	86%	85%	85%	84%	86%	87%	86%
(N/A removed)	6%	6%	8%	7%	7%	5%	6%	5%	4%	2%	6%
Auditory Hallucinations	7%	9%	9%	8%	9%	7%	6%	8%	7%	8%	8%
Pleasantness	75%	75%	74%	70%	75%	71%	68%	73%	70%	73%	72%
(N/A removed)	7%	9%	10%	9%	9%	6%	6%	8%	6%	8%	8%
Recovery	8%	5%	5%	5%	3%	3%	2%	3%	2%	2%	4%

Table 4:1 Recovery data - Percentage missing data conditional on momentary response and that at least one item completed within the whole sampling period: IDs 37 and 51 dropped.

[1] Deservedness and Pleasantness items originally contained a not applicable missing value for when no corresponding paranoia or hallucinations were currently observed. The N/A removed percentages thus represent missing branched items where the original item was rated > 1.

The results indicate that item nonresponse conditional on momentary response is low: diaries are more likely to be missing completely than sparsely filled. The presence of some item level missingness, however, does demonstrate a degree of selective reporting in 'completed' diaries. Moreover, as better expressed through Figure 4:2 and Figure 4:3, certain items are more consistently left unanswered than others. Though the difference in magnitude is small, visual and auditory hallucination items are unanswered more than the other measures, both within-day and across the week. While it can only be speculated, it may be possible that these items are skipped when subjects are not currently experiencing a hallucinatory event rather than being given a null score. Selective item reporting is more prominent in the bipolar study data, where cannabis use is much more often recorded than items relating to mood or symptoms (Figure 4:4 and Figure 4:5).

The low level of missing data for each item is relatively consistent, with no clear indication of drop-off towards the end of the questionnaire (comparing the table rows). Recovery, the final three items of the questionnaire for example, has similar levels of nonresponse as self-esteem which is positioned towards the start, though the level of missing recovery is slightly elevated on Day 1. There are no striking time trends in nonresponse, though item completion seems to improve slightly for Hopelessness and Recovery towards the end of the day and for most items towards the end of the week (see Figure 4:2 and Figure 4:3), though this is likely due to the increase in moment nonresponse observed, resulting in fewer partially completed questionnaires.

Deservedness and the two Pleasantness items exhibited consistent, high levels of nonresponse across both the day and the week. However, these three items were the second stage of conditional branching providing additional information on their preceding items. *Deservedness* quantified how deserving the subjects felt of their current paranoid thoughts, while *pleasantness* appraised their current hallucinations (unpleasant to pleasant). As such, if the main item is given a null score or is missing the subsequent item is not applicable. If the non-applicable answers are recoded to remove them from those not reported, the proportion of missing data for these items reduces dramatically (row N/A removed in Table 4:1).

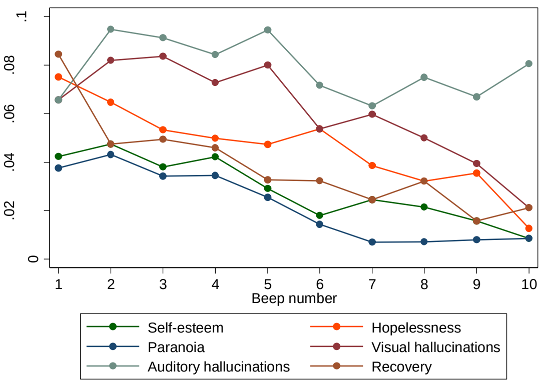


Figure 4:2 Proportion of item nonresponse at each beep, conditional on event response. Recovery data

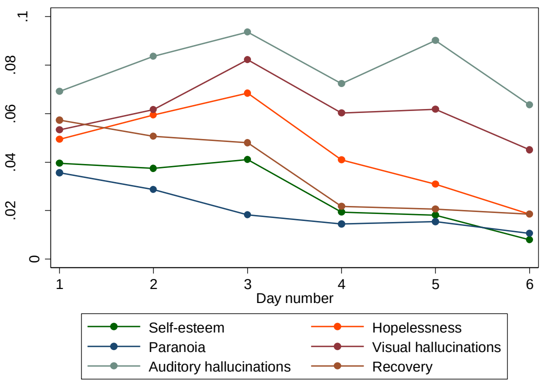


Figure 4:3 Proportion of item nonresponse on each day, conditional on event response. Recovery data

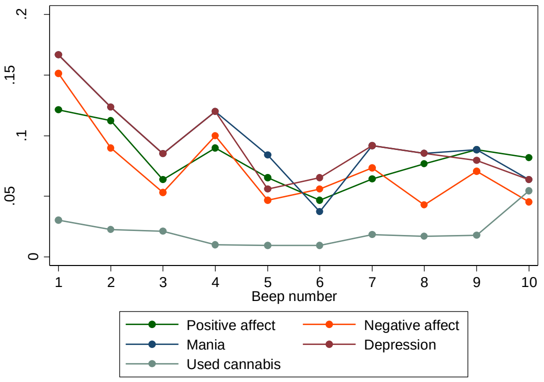


Figure 4:4 Proportion of item nonresponse at each beep, conditional on event response. Bipolar data

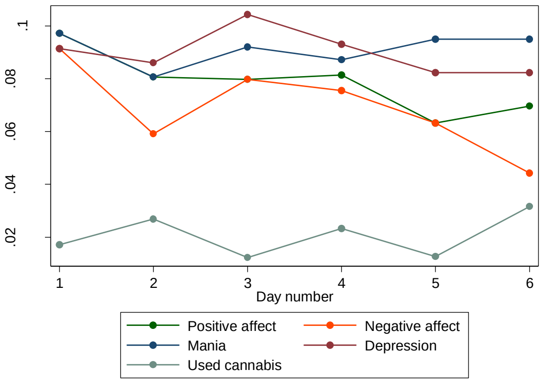


Figure 4:5 Proportion of item nonresponse on each day, conditional on event response. Bipolar data

In this study the phrasing of the deservedness items were particularly ambiguous and resulted in inconsistent reporting: these items often scored at strong positive or negative feelings of deservedness during moments when no paranoia was observed. These items perhaps captured a more global feeling of deservedness of paranoid thoughts, rather than momentary reflections on current feelings. Care should thus be taken when making inference on this measure as it may not be ecologically valid.

4.2.3 Moment nonresponse

For the recovery data moment nonresponse was classified as all 50 items of the diary being incomplete. An alternative was to define moment nonresponse when the primary outcome variable *recovery* was missing, however, as item nonresponse was low in this data there were very few occasions for which recovery was missing and at least one other item was complete.

		At least one item observed	All items missing
Outcome	**Observed**	2481	0
	Missing	97	1502

Table 4:2 Moment nonresponse definition comparison for recovery data

Out of the 4,080 diaries administered to participants, 1,502 (37%) were returned with full item nonresponse. Between-subject, variation in moment nonresponse was high, ranging from zero to 59 (98%) missed diaries from the total of 60 intended entries per subject. The distribution of the number of uncompleted diaries was skewed, with a median of 16 fully incomplete diaries per subject, as presented in Figure 4:6.

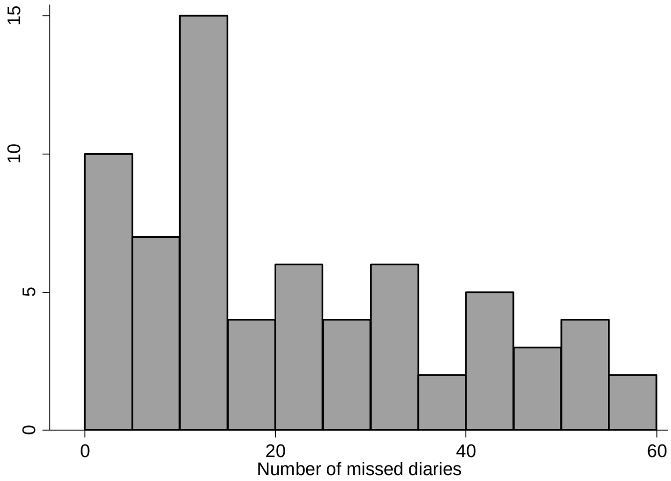
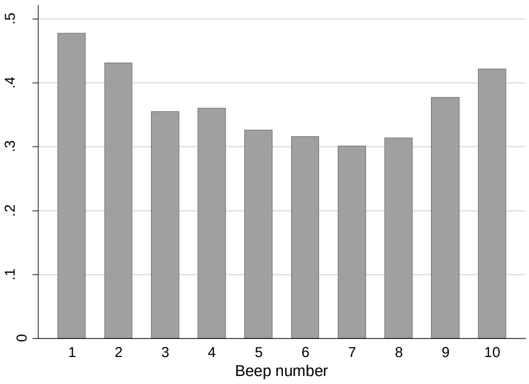


Figure 4:6 Distribution of missed diaries in recovery data

The pattern of nonresponse can be further described by when the missed diaries occur. Are signals ignored more often in the morning, for example, or does nonresponse increase across the week? The proportion of completely empty diaries for each moment, averaged across the six days, is presented in Table 4:1. Presented graphically, Figure 4:7 shows the average proportion of missed diaries at each moment across all days and Figure 4:8 shows the proportion of missed diaries on each day of the sampling period.



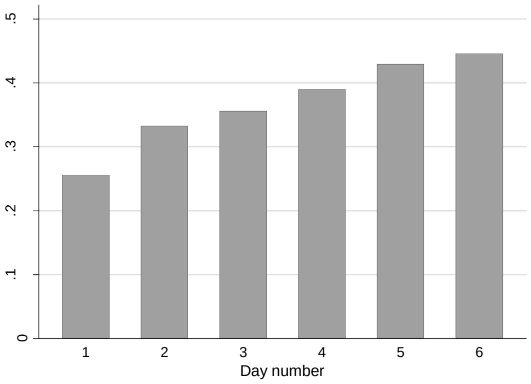


Figure 4:7 Average proportion of missed diaries throughout the day in the recovery data

Figure 4:8 Proportion of missed diaries over the week in the recovery data

The graphs indicate greater moment nonresponse at both the start and end of the day, and the proportion of missed diaries increases over the course of the week, with almost twice as many missing diaries on Day 6 as on Day 1.

4.3 STATISTICAL ANALYSIS WITH PARTIALLY OBSERVED DATA

Missing data can be problematic for longitudinal studies such as ESM. Fewer observations available for analysis leads to a reduction in power to detect effects and with subjects completing different numbers of observations, analysis methods which require balanced data cannot be used. Depending on the reason data are missing, known as the missing data mechanism, and how the data are subsequently analysed, there is the potential for substantial bias. Solutions to these problems will be discussed in this section.

4.3.1 MISSING DATA MECHANISMS

The classification of missing data can be described by the missing data mechanisms as defined by Little and Rubin (1987): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), also known as non-ignorable (NI) missingness. The standard definitions of these mechanisms state that all data intended to be collected Y can be partitioned into Y^O, the observed data, and Y^M the missing data and that a binary indicator R is equal to 1 when Y is observed and 0 when Y is missing. The three missing data mechanisms can then be defined as follows. MCAR occurs when the probability that data are missing depends on neither the observed data nor the missing data, that is

$$P(R = 1|Y^O, Y^M) = P(R = 1).$$

In this case the observed data are simply a random subset of the sample. When data are stated to be MAR the probability that the data are missing depends on the observed data but not the missing data:

$$P(R = 1|Y^O, Y^M) = P(R = 1|Y^O).$$

The observed data are now no longer a random subset of the sample. However, when stratified by Y^O, the observed data within strata are considered a random subset. Finally, the data are described to be MNAR when the probability of missingness depends on the values of the missing data

$$P(R = 1|Y^O, Y^M) \neq P(R = 1|Y^O).$$

These definitions have been applied to longitudinal settings ((Diggle 2002; Fitzmaurice, Laird et al. 2012) amongst others), however, there appears to be no extension to a three-level ESM setting currently defined in the literature. When n_1 measurements are taken on each of n_2 days for $k = 1, \dots, n_3$ subjects, we now define an $n_1 n_2 \times 1$ vector of all possible responses for subject k

$$Y_k = \left(Y_{11k}, Y_{21k}, \dots, Y_{n_1 1k}, Y_{12k}, Y_{22k}, \dots, Y_{n_1 2k}, \dots, Y_{n_1 n_2 k}\right)'$$

of which some elements are observed Y_{ijk}^O and others are missing Y_{ijk}^M. R_k is now also defined as an $n_1 n_2 \times 1$ vector

$$R_k = \left(R_{11k}, R_{21k}, \dots, R_{n_1 1k}, R_{12k}, R_{22k}, \dots, R_{n_1 2k}, \dots, R_{n_1 n_2 k}\right)'$$

where $R_{ijk} = 1$ if Y_{ijk} is missing and $R_{ijk} = 0$ if Y_{ijk} is observed. Unlike the standard notation above, for longitudinal data Y_k contains the data solely for the variable of interest. It is assumed that all other measured data are contained within the $n_1 n_2 \times p$ matrix of covariates X_k. The missing data mechanisms can then be redefined using this longitudinal notation. The definition of MCAR is largely unchanged: the probability the data are missing is again independent of both the observed data and the missing data

$$P\left(R_{ijk} = 1 \middle| Y_k^O, Y_k^M, X_k\right) = P\left(R_{ijk} = 1\right).$$

In an ESM study momentary level missingness might be MCAR if, for example, a booklet is lost after collection. Alternatively item nonresponse may be considered MCAR due to poor layout of the diary. An extension to this definition is presented as covariate dependent

MCAR (Little 1995; Fitzmaurice, Laird et al. 2012), where missingness is independent of both observed and missing Y but can depend on covariates X

$$P\left(R_{ijk} = 1 \middle| Y_k^O, Y_k^M, X_k\right) = P\left(R_{ijk} = 1 | X_k\right).$$

This is akin to the standard MAR definition where the observed data Y_O contained all recorded observations, not just the response. For longitudinal data the distinction is made between covariate dependent MCAR and MAR as for MCAR it is still assumed that the missing response does not depend on the observed response. The assumption of MCAR only holds in this case when conditioning on all covariates that are predictive of R_k.

MAR occurs when the probability that data are missing depends on the observed $\left(Y_k^O\right)$ but not missing $\left(Y_k^M\right)$ data,

$$P\left(R_{ijk} = 1 \middle| Y_k^O, Y_k^M, X_k\right) = P\left(R_{ijk} = 1 | Y_k^O, X_k\right).$$

Differentiating between covariate dependent MCAR and MAR may be problematic in an ESM context where momentary nonresponse is present, i.e. all diary items are missing. Although this longitudinal extension to the missing data mechanism allows data at other time points to be predictive of missingness, as all items are missing, instead of a vector of all measurements on one variable Y_k is now a $n_1 n_2 \times p$ matrix for all ESM diary items. Now, using a diary item X_1 at $t - 1$ to predict missingness at time point t the data can no longer be considered covariate dependent MCAR as X_1 is captured with Y_k and is thus defined as Y_k^O. Fortunately, a distinction between the two mechanisms is not necessary for analysis; valid inference can be obtained with a likelihood-based analysis provided that care is taken to ensure both the fixed and random parts of the model are correctly specified, as data which are MAR can be sensitive to model misspecification (Fitzmaurice, Laird et al. 2012).

MNAR applies when the probability the data are missing depends on the missing values themselves. In an ESM setting this might occur when data on hallucinations, for example, are missing when a subject is currently experiencing a strong hallucination. In contrast to MAR, the distribution of Y_k^M conditional on Y_k^O is not representative of the intended sample; in this case the distribution of the missing data is said to depend on both Y_k^O and $P(R_k = 1 | Y_k, X_k)$ and a joint model is required to model both the response and the missing data mechanism. Two general approaches to this joint modelling procedure include selection models (Heckman 1976; Little 1995) and pattern mixture-models (Little 1993;

Little 1995), though these methods are beyond the scope of this thesis. Where data are MNAR and the pattern of nonresponse is not sufficiently modelled, parameter estimates can be biased (Collins, Schafer et al. 2001).

Whilst distinguishing between missing data mechanisms is empirically untestable, one can test for predictors of nonresponse in an attempt to satisfy the covariate dependent MCAR or MAR assumptions. The variables predictive of missingness should be included in the analysis model to reduce bias when assuming these mechanisms. If the model is correctly specified, estimation using maximum likelihood where all available information is included should produce unbiased estimates. However, for large amounts of missing data the reduction in sample size could lead to inflated standard errors.

Finally, it is recommended that following the primary analysis assuming a particular missing data mechanism, a sensitivity analysis is carried out to evaluate how robust the results are to departures from this assumed mechanism (Carpenter, Kenward et al. 2007; Carpenter and Kenward 2012). Carpenter and Kenward (2007) describe how this can be achieved by either explicitly modelling the non-response mechanism or via imputation and comparing the conclusions of this to the primary analysis to see how they vary.

4.3.2 Approaches to analysis with missing data

There are several common approaches to missing data, each with benefits and drawbacks in a longitudinal setting. The first approach is to simply ignore the missing data and use a complete case analysis. With a complete case analysis only observations for which all variables are complete are included in the analysis; missing data in one covariate results in no data for this observation being using in the model. For only one variable with missing data this might not be too detrimental, but if data are missing on many variables this can result in a dramatically reduced dataset. Moreover, a complete case analysis is only valid when data are MCAR. If data are MAR a complete case analysis will result in biased fixed effect estimates (Fitzmaurice, Laird et al. 2012). So as not to lose all of this observed information, missing values are often filled in or 'imputed'. Several options for imputation are available and will be discussed in terms of their appropriateness for ESM research.

The most straightforward approach for imputing missing data is known as mean imputation which involves replacing the missing values with the mean value of each variable. This benefits from its simplicity to execute, however, this form of imputation will underestimate variances, and coefficients will likely be regressed towards the mean. However, for repeated measures data where one may expect consecutive measurements to be correlated this approach may not be appropriate. Furthermore, for three-level ESM data there are several 'mean' values to choose from: the grand mean, averaging over all measurements; the subject mean, averaging over all measurements within-subject; and the day mean, averaging over the measurements within each day for each subject. The day mean might be the most appropriate if one expects a large amount of between-day or between-subject variation, however, if there are large amounts of missing data on certain days this value may be biased, only based on a small amount of data.

The mean imputation method, even accounting for the nested structure of ESM data, is still flawed in that it is unrepresentative in the case of serially correlated data. Where data follow a pattern of autocorrelation or time trend, imputing with this method will bias this trend towards the mean. An alternative method from the longitudinal literature is to use the last observation carried forward (LOCF). Here the value for Y_{ijk} at the time point before the missing value ($Y_{i-1,jk}$) is used as a substitute, preserving any correlation between successive observations more than using a mean value would. This method is widely criticised as it produces biased estimates of the means and (co)variances even under the assumption of MCAR (Molenberghs, Thijs et al. 2004; Siddiqui, Hung et al. 2009) .

A more sophisticated method for imputing missing data is to use information from completed observations of other variables for each subject. Simple imputation involves regressing the variable with missing data on complete variables and predicting the unobserved values. This method benefits from using a greater amount of information to inform the selection of missing values and being able to take into account any within-subject correlation in its prediction by imputing using a multilevel model appropriate for the data.

A more sophisticated alternative to missing data, avoiding the bias produced by these more simple methods (Donders, van der Heijden et al. 2006), is using multiple imputation. Multiple imputation (Rubin 1978) imputes missing data based on complete data from other variables. Working under a Bayesian framework, the missing data are sampled from their predicted distribution given the observed data. This process is repeated on m copies of the original dataset resulting in m different, complete datasets. The original model of interest is then fitted to each of the complete datasets and the parameter estimates averaged to give the final results (Sterne, White et al. 2009). Standard errors and confidence intervals for the parameter estimates are calculated using Rubin’s rules (Rubin 1987). By imputing the missing data on multiple data sets and combining the results, multiple imputation aims to model the uncertainty of the missing values and thus reduce bias in the final result.

Evidence suggests that the imputation model should have the same structural form as the intended analysis model (Rubin 1996; Schafer and Olsen 1998), and as such for an intended multilevel analysis model a multilevel imputation model should be used. In a simulation study, Black, Harel et al. (2011) investigated changes to parameter estimates in multilevel models when imputing using normal (single level) and linear mixed models at varying levels of missing data. The authors found that while the fixed effects were “generally unbiased” for both imputation models, multiple imputation using a normal model resulted in substantial bias at level 1, significantly overestimating the residual variance (σ_e^2) for all levels of missing data. They also reported that although the linear mixed model produced a less bias estimate, it slightly overestimated σ_e^2 for higher levels of missing data. When estimating the variance of the random intercept, the normal model underestimated the parameter for all levels of missing data, the estimate biased to zero as random intercepts cannot be included in the normal imputation model.

As the level 1 and level 2 variance estimates may be of interest in an ESM analysis, a method of imputation which reduces any bias will be preferential. However, when the fixed effects are of primary importance these results suggest a multilevel imputation model may not be necessary.

A second argument against multiple imputation can be made when estimating using maximum likelihood. This procedure uses all available information to calculate parameter estimates rather than dropping rows as in a complete case analysis. For large samples, Schafer (1999) suggests little is gained from multiple imputation other than random noise produced by the random draws. Furthermore, in addition to the single-level and multilevel multiple imputation models, Black, Harel et al. (2011) also compared the results of simply using maximum likelihood estimation in their simulations. They found that the maximum likelihood model produced generally unbiased fixed effect estimates, and although bias was small across all models (“less than 33%, or one-third of one standard deviation”), for large amounts of missing data (50%) estimates were more biased using the multiple imputation models than for the maximum likelihood model. Bias in the random effect estimates using maximum likelihood was comparable to multilevel multiple imputation, and considerably better than imputation using a single level model. Although underperforming in some aspects, for example in efficiency and accuracy of random slope variances, maximum likelihood was found to be comparable to using multilevel multiple imputation in many areas, and superior to a single level imputation model in most.

Multiple imputation can be carried out in most software packages, including SAS, Stata, and R. However, to impute multilevel data the clustering must be accounted for when creating the multiply imputed data sets, which is not a clearly defined process in these packages. Multiple imputation is possible in MLwiN for multilevel data using a series of macros (missingdata.org.uk) but only for missing data at level 1. To impute missing data at both levels of a two-level dataset the newly developed software REALCOM-Impute (Carpenter, Goldstein et al. 2011) is available which can be run in conjunction with either MLwiN or Stata. Here the multilevel data is exported to REALCOM-Impute along with the variables selected for the imputation model. Multilevel imputation models are then specified, in which variables with missing data are jointly modelled with multivariate normal random effects and estimated by MCMC (Carpenter and Kenward 2012). The imputed data sets are subsequently exported back to the original program. The model of interest can then be fitted to each of the imputed data sets and combined for an estimate in the usual way. For three-level data Goldstein (2009) suggests using a dummy variable for the level 3 clustering variable as a third hierarchical level is not currently accommodated. However, considering the findings of Black, Harel et al. (2011), the inability to correctly specify the third level in the imputation model may result significantly biased results compared to simply fitting a multilevel model to the incomplete data and estimating with maximum likelihood.

4.4 PREDICTORS OF NONRESPONSE IN THE RECOVERY DATA

Likelihood based analysis methods require the assumption that data are MAR. While this is strictly untestable, predictors of nonresponse can be investigated in the observed data. Any significant variables can then be included as covariates in the analysis model to satisfy the MAR assumption. The following tables present the results of univariate three-level logistic regression models, with random intercepts for subject and day, where the outcome is a binary indicator variable equal to 1 if the recovery data are missing and 0 if the data are observed at each moment. The fixed effects estimates are presented as odds ratios, representing the odds of the response being missing for larger values of the predictor.

4.4.1.1 Item Nonresponse

As partial completion of diaries was possible in the recovery data, the recovery measure may be missing in a diary but other items completed. As such these observed items may be used to predict missingness in recovery at the same time point. As scales were to be used in the main analysis rather than individual items, nonresponse at the item level was investigated in terms of the grouped measure. Table 4:3 presents the results of the concurrent predictors of item-level nonresponse in recovery.

	Fixed effects			Random effects			
Covariate	OR	Std. Err	P Value		Variance	Std. Error	N
Self-esteem	1.023	0.199	0.906	Subject	4.601	2.621	2505
				Day	1.077	0.847	
Hopelessness	1.017	0.167	0.918	Subject	1.137	0.759	2460
				Day	0.000	0.000	
Paranoia	1.113	0.229	0.603	Subject	4.931	2.948	2523
				Day	2.549	1.315	
Auditory Hallucinations	1.099	0.195	0.593	Subject	8.808	4.432	2421
				Day	1.808	1.130	
Visual Hallucinations	1.223	0.195	0.208	Subject	12.681	6.469	2375
				Day	1.875	1.115	

Table 4:3 Concurrent predictors of missing recovery item

No variables were significantly associated with an increase in odds of missingness in recovery. This is likely due to the findings of Table 4:2, that when recovery is missing the whole booklet is also likely to be missing, thus there were very few occasions in the above models where predictor data are available when $R = 1$.

4.4.1.2 Moment nonresponse

Instead of a concurrent prediction of missingness, the diary items can instead be used to examine missingness at the following moment. This is achievable using a lagged covariate model, with items at time point $i - 1$ predicting momentary level missingness at occasion

i. Moment nonresponse was defined as a level 1 binary variable denoting whether a whole booklet had been left uncompleted or whether at least one item was answered.

	Fixed effects			Random effects			
Lagged covariate	OR	Std. Err	P Value		Variance	Std. Error	N
Recovery	0.946	0.063	0.403	Subject	1.563	0.394	2250
				Day	0.00	0.00	
Self-esteem	0.843	0.066	0.029	Subject	1.582	0.411	2271
				Day	0.00	0.00	
Hopelessness	1.050	0.076	0.499	Subject	1.481	0.379	2227
				Day	0.00	0.00	
Paranoia	1.176	0.078	0.015	Subject	1.467	0.377	2289
				Day	0.00	0.00	
Auditory Hallucinations	1.035	0.091	0.693	Subject	1.714	0.435	2190
				Day	0.00	0.00	
Visual Hallucinations	1.053	0.072	0.445	Subject	1.718	0.436	2158
				Day	0.00	0.00	

Table 4:4 Lagged predictor of missing diary

Both self-esteem and paranoia significantly predicted a missing diary at the following time point: higher self-esteem scores decreased the odds of subsequent momentary nonresponse (OR=0.84, SE=0.07, p=0.029) while higher paranoia scores significantly predicted greater odds of a diary being missed (OR=1.18, SE=0.08, p=0.015).

The within-day variation is estimated as negligible in these analyses, likely due to the lack of variability in momentary missingness after conditioning on the lagged covariate being observed. On further examination of the data, the proportion of missing moments where the previous moment was observed is far smaller than when consecutive diaries are both completed, as displayed in Table 4:5. Only the data from the top row of the table are included in the analysis, i.e. the outcomes for when the lagged predictor is observed, and so out of the 2342 available observations for analysis, $Y = 0$ in only 344. Such a low proportion of missing diaries following a completed diary suggests that missing diaries are not sporadic; completion of the current diary is dependent on completion of the previous diary.

		Diary at moment i	
		Observed ($\boldsymbol{Y = 1}$)	**Missing ($\boldsymbol{Y = 0}$)**
Predictor at	**Observed**	1,998 (85%)	344 (15%)
$i - 1$	**Missing**	367	963

Table 4:5 Table of observed and missing consecutive diaries displaying the relative proportions in response categories when the predictor is observed

4.4.1.3 Time trends in event nonresponse

Time trends both within- and between-day moment nonresponse in the recovery data was investigated to compare to the evidence found by both Silvia, Kwapil et al. (2013) and Messiah, Grondin et al. (2011). Daily trends were identified by entering beep number (centred at 1) as a predictor and weekly trends using day number (also centred at 1). In both models the response was the binary indicator variable moment nonresponse, equal to zero if at least one diary item was complete and one if the whole diary was missing. Nonlinear trends were studies by including each predictor as a linear and quadratic term.

	Fixed effects			Random effects		
	OR	Std. Err	P Value		Variance	Std. Error
Intercept	1.139	0.324	0.648	Level 3	4.596	0.945
Beep number (linear)	0.596	0.034	<0.001	Level 2	1.430	0.212
Beep number (quadratic)	1.052	0.006	<0.001			

Table 4:6 Within-day quadratic trend in event nonresponse

The results in Table 4:6 suggest that there is a quadratic trend in missing data within the day: data are more likely to be missing both at the start and end of each day as demonstrated in Figure 4:7.

	Fixed effects			Random effects		
	OR	Std. Err	P Value		Variance	Std. Error
Intercept	0.222	0.066	<0.001	Level 3	4.379	0.894
Day number (linear)	1.564	0.232	0.003	Level 2	1.060	0.169
Day number (quadratic)	0.965	0.027	0.209			

Table 4:7 Between-day quadratic trend in event nonresponse

The results of Table 4:7 suggest there is a significant linear weekly trend in missing data; as day number increases the odds of nonresponse increase by 56%. This implies that, as one might expect, missing data are more prevalent towards the end of the sampling period, indicative of typical longitudinal attrition.

These results suggest that the hours set for the sampling scheme might not have been appropriate for this study population (perhaps with subjects waking later than expected), resulting in more ignored or missed prompts. This could potentially be informative for the design of future ESM research focused on a population of people with psychosis. The drop off towards the end of the week echoes Silvia, Kwapil et al. (2013)'s message, that more focus needs to be directed at reducing nonresponse in the study design to prevent participant fatigue or loss of interest as the study progresses.

4.5 Summary

ESM research has the potential for large amounts of missing data: at the item-level within each questionnaire; at the moment-level, where all items are left uncomplete; and at the day-level, where no data have been recorded in any diary for the day. Current published research poorly describes missing data (where it is even discussed at all) as summarising all of this information into one value defined as 'compliance' or 'adherence' is not informative of the breadth of possible sources of nonresponse.

A detailed understanding of missing data will aid in both the analysis of current research and in the design of future studies. This chapter has described options for presenting nonresponse at each level so as best to fully describe any missing data patterns both within- and between-days. It has demonstrated how current and previous diary information can be used to predict missing data and how time trends can be explored.

Finally, under the assumption of MAR, it is suggested that multilevel multiple imputation may not be necessary for missing ESM data when analysis methods estimate using maximum likelihood.

5 Momentary level variation and time trends

One of the suggested benefits of ESM is its ability to capture subtle variation in trait-like symptoms that would be unobtainable using a more traditional procedure (Delespaul 1995; Csikszentmihalyi 2014). Avoiding the problem of recall bias when providing a retrospective account of symptoms and behaviours, ESM can capture moment to moment changes and allow researchers a greater understanding of within-subject variation in symptoms and behaviours.

The aim of this chapter is to explore methods for investigating momentary variation in ESM measures. Firstly, current practices in studying variation will be presented as identified in the review of Chapter 3. This will be followed with an alternative approach, utilizing the flexibility of thee-level random slope models.

5.1 Introduction from systematic review

Although novel to this methodology, studying moment to moment variation is underutilized in practice with only five of the 74 papers reviewed explicitly identifying variation in measures as an interest. Though briefly discussed in Chapter 3, these five papers will be described more thoroughly here. Each of the papers defined 'variation' in slightly different ways; these definitions will be presented with details on the study designs and analysis methods.

5.1.1 Papers identified in systematic review studying variation

Peters et al (2012) used ESM to investigate appraisals of symptoms in participants with psychosis, questioning whether symptom appraisals and delusional convictions were constant. Twelve participants were recruited from the Psychological Interventions Clinic for Outpatients with Psychosis in London and were asked to complete an ESM questionnaire 10 times a day for six days. Items within the questionnaire were designed to assess psychotic symptoms and appraisals of these symptoms, rated on 7-point Likert scales. Variation was defined in terms of 'constancy': the difference between successive observations, averaged to create one subject-level score, $\sum_i (x_{ik} - x_{i-1,k})$, for measurements i within subjects k. The mean constancy of appraisal for each symptom was presented with the results of a t-test to test whether each mean was significantly different than zero. Non-zero values were considered representative of significant changes in appraisal over time.

Similar to Peters, Udachina et al (2012) considered variation as an outcome and were interested in group differences in the instability of deservedness and self-esteem in patients with paranoia. The authors monitored 41 patients and 23 healthy controls for six days, requiring them to complete 10 paper diaries a day designed to capture deservedness of paranoid thoughts and self-esteem. ‘Instability’ was defined as the averaged absolute difference in successive scores for each participant, $\sum_i \left| x_{ik} - x_{i-1,k} \right|$, “indicating a mean change from moment to moment”. Group differences were then assessed using multilevel models with the subject-level instability measure as the outcome,

$$Instability_k = \beta_0 + \beta_1 Group_k + u_k + e_{ik} \qquad (4)$$

where it is presumed $Group_k = 0, 1, 2, 3$ representing the groups: controls, PM, BM and remitted; PM “Poor Me” and BM “Bad Me”, two paranoid states identified from the Persecution and Deservedness Scale (PaDS; Melo et al., 2006).

Palmier-Claus et al (2012) used a sample of 27 individuals at ultra-high risk of developing psychosis to investigate the association between instability of affect and suicidal ideation. In this study, the variable measuring instability was used as a predictor rather than outcome. As in the previous studies, patients completed questionnaires 10 times a day for six days, accumulating data on positive and negative mood items. ‘Instability’ of mood was calculated similarly to Peters and Udachina, using the mean successive squared difference (MSSD) (von Neumann et al. 1941) , $\sum_i \left(x_{ijk} - x_{i-1,jk} \right)^2$. Rather than averaging over all observations, the MSSD was calculated for each day j resulting in two-level data with a mean daily mood variability measure for each day per person. Suicidal ideation was only measured once as a subscale of the Comprehensive Assessment of At-Risk Mental State (CAARMS) semi-structured interview delivered in the debriefing at the end of the ESM period. A linear regression with robust standard errors was used to account for the clustering within subject to assess the association between the two variables.

Daily variability was also created in Selby et al’s (2012) study of affect in patients with bulimia nervosa (BN) and borderline personality disorder (BPD). The authors hypothesized subjects with both BN and BPD would experience greater daily variability in positive and negative affect than subjects with BN only. Twenty five individuals with BN and BPD were compared against 108 with BN alone. Both groups were measured for 14 days and were required to complete a questionnaire on a PalmPilot at six random times during the day plus at any time in which they “engaged in eating disorder behaviours (e.g. binge eating, laxative use, etc.) or self-destructive behaviours (e.g. self-injury, drug use, etc.)". A measure of 'daily variability' was calculated for each subject as the standard deviation of the mean ESM measured affect for each day. A two-level model was then used with daily variability as the dependent variable with predictors including group variable, an indicator for whether the day contained an bulimic event and a group-by-event interaction.

Finally, Demiralp et al (2012) examined emotional differentiation in patients with Major Depressive Disorder (MDD), and whether this was unique from other emotional constructs such as emotional intensity or variability. The 106 participants were measured eight times a day for "7 to 8 days", completing a questionnaire when prompted on a PalmPilot. At each point participants were asked to rate their current emotions on a 4-point scale. The questionnaire contained seven negative emotion items: *sad, anxious, angry, frustrated, ashamed, disgusted*, and *guilty* and four positive emotion items: *happy, excited, alert*, and *active*. Similarly to Selby's study, 'temporal variability' was measured as the variance in intensity of each emotion over the whole sampling period, where 'emotional intensity' was defined as the average of the emotion items at each prompt, which was then averaged across the whole sampling period for each subject. 'Emotional differentiation' was calculated by taking the correlations of all pairs of emotion items (e.g. $r_{sad,anxious}, r_{sad,angry}, r_{sad,frustrated}, \ldots$) to create separate measures for positive and negative emotions, "the average of the Fisher's z-transformed correlations" representing the positive and negative emotional differentiation score. High scores indicated higher differentiation in emotions. As the variability and differentiation measures were subject level variables, a simple linear regression model was used to predict the changes in emotional differentiation with depression, emotional intensity and emotional variability as predictors.

5.1.1.1 MSSD MEASURES VS. VARIANCE MEASURES

The various definitions of variation in the papers can be sectioned into two categories: those using a measure of difference in successive observations (MSSD measures; Peters et al, Udachina et al and Palmier-Claus et al) and those using a measure of dispersion about the mean (variance methods; Selby et al and Demiralp et al). Although these definitions are distinct, they both capture the spread of the data; large scores indicating a wide spread, small scores indicating observations are more similar.

The MSSD methods differ from the variance methods by measuring this spread moment to moment, where a large score implies that successive observations are widely spaced.

Variance methods cannot make this distinction, instead quantifying difference from the population mean. For example, when examining the variation in mood using ESM, if mood is consistently low for the first half of the sampling period then consistently high for the second half, a variance measure would be large but an MSSD measure much smaller. The MSSD measure is therefore superior for detecting momentary changes, satisfying the definitions of 'constancy' or 'instability', but is potentially less sensitive to variation as a whole. Specific interpretation of this measure, however, is not straightforward. Although generally speaking larger values correspond to greater momentary variation, the value itself has little meaning and will depend heavily on the way in which it was created. The three MSSD style measures used by Peters et al, Udachina et al and Palmier-Claus et al, for example, produce quite different results. Peters' formula simply sums the difference between successive moments, while Udachina's sums the absolute difference and Palmier-Claus' the squared difference. By definition, it is obvious that the three formulae will result in very different values. As successive observations may increase and decrease in value, summing without accounting of the potential difference in sign Peter's formula does not accurately represent the magnitude of change: negative momentary changes will cancel out positive changes, tending the average to zero. Taking the absolute difference counteracts this problem, as does squaring the difference. These two methods, however, can also result in substantially different totals when momentary changes are large.

To illustrate how these MSSD style measures differ, each was computed using the recovery data set for the variable self-esteem. Figure 5:1 and Figure 5:2 present box plots of each method. The first demonstrates the range of values obtained from using the simple mean difference compared to the absolute mean difference. The second shows how squaring the difference can result in much larger values (note the change in scale on the y axis). Although the three MSSD methods are not being compared within papers, it is important to note how different the results might be depending on which measure is chosen to represent variability.

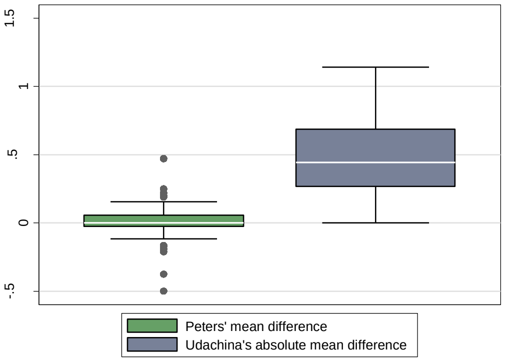


Figure 5:1 Box plot of Peters' vs Udachina's MSSD method for recovery data

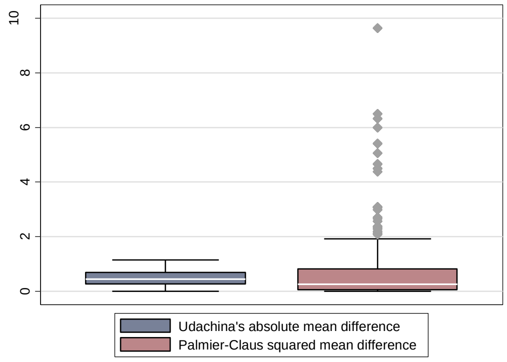


Figure 5:2 Box plot of Udachina's vs Palmier-Claus' MSSD method for recovery data

5.2 Alternative approaches to investigating variation: Time trends

One major drawback of the MSSD methods is how to interpret the value of variability obtained. Peters et al. presented the mean value for each of their symptom appraisals (subject-level differences averaged across all participants) and used a t-test to establish whether these appraisals, as they described it, were constant over time. A mean change of 0.62 (SD=0.5) in control of hallucinations, for example, was found to be significantly different than zero, this appraisal thus described as being “non-constant” over time. However, no description of how the mean difference varied across participants was given and the summary measure provides no further information on the mean change. As so much information is gathered using ESM, much more informative analyses can be conducted.

One study from the systematic review better utilized the intensive longitudinal data structure to investigate variation in their sample. Giesbrecht, Letourneau, Campbell, and Kaplan (2012) were interested in the trajectory of positive and negative affect over the course of pregnancy. They tracked momentary level affect using an electronic ESM diary to gather measurements on 76 women for two days in each trimester of their pregnancy. Measurements were taken five times a day, 30 minutes after waking and then at four semi-random intervals throughout the day. The resulting data was thus of a four level structure: measurement moments within days, within trimester, within subject. Instead of using a four-level model, the authors chose to aggregate scores to the trimester level “for simplicity”, resulting in three mean positive and negative affect scores for each subject. Two-level multilevel models were then used to investigate the change in affect over time, where ‘time’ was measured as gestational age, centered at six weeks, which was entered into the model as both a linear and quadratic term. The authors were also interested in whether depressive symptoms moderated the two affect trajectories. Depressive symptoms were measured using the Edinburgh Postnatal Depression Scale (EPDS) within each trimester, the three scores then aggregated to give a mean score for the whole pregnancy where higher values correspond to higher levels of depression.

Significant quadratic trends were present in both positive and negative affect. For positive affect, opposite trends were observed in women with high depressive symptoms (where positive affect increased in early gestation but decreased in later gestation) to those with low symptoms (where affect decreased during early and middle gestation and increased towards the end of pregnancy). For negative affect the same trend was observed in both groups, negative affect decreasing over time but increasing towards the end of pregnancy. However, women with higher depressive symptoms experienced significantly higher levels of negative affect throughout pregnancy.

The results presented were “based on estimation of fixed effects with robust standard errors”, apparently not using two-level models as originally stated. As such, it is assumed that no between-subject heterogeneity was modelled in the intercept or slopes, the authors choosing not to measure subject-level variation. Aggregating affect to the trimester level eliminates all within-trimester variation in these variables. Similarly, aggregating the EPDS to the subject level throws away between-trimester information, subtlety further lost in dichotomising this variable for analysis.

This type of model is often referred to as a growth model in the longitudinal literature, and is typically used to study trends in longitudinal models where ‘time’ may be a function of age, for example. The systematic review suggests the use of growth models in ESM is far less common. It may be the fact that researchers do not consider this form of research question for ESM studies or are unaware of the potential to use it to examine short time periods. The principals of growth modelling can be used to study trends in ESM data and can be adapted to accommodate three level data. This chapter will describe how ‘time’ can be defined in ESM data and how momentary level variables can be studied across time.

5.2.1 Two-level time trend models

To discuss time trends in the application of three level models for ESM data it is useful to begin with a simpler, two level description. Here, the models will describe the variation in an outcome y_{ik} for moments or 'beeps' $i = 1, \dots, n_{1k}$, nested within participants $k = 1, \dots, n_2$. Time will be denoted t_{ik} and refer to the time at which a questionnaire is completed. The subscripts i and k are used for the general scenario where the number of moments can differ for each participant or when, such as in a random prompt design, participants complete questionnaires at different time points to each other. This may be simplified to just t_i when each participant is required to complete the same number of questionnaires under a fixed time schedule or when the beep number is used as a proxy for time and the number of beeps per subject is balanced.

Figure 5:3 - Figure 5:6 represent the four possible trends one can observe within a two-level model. In each figure the black line represents the population average while the blue and green lines represent two subject-specific time trends.

Figure 5:3 represents the simplest model with no fixed or random effects of time on outcome

$$y_{ik} = \beta_0 + u_k + e_{ik}.$$

In this variance components model each subject-specific line runs parallel to the population average line with random intercepts $\beta_0 + u_k$ for each subject k. There is no variation in outcome over time.

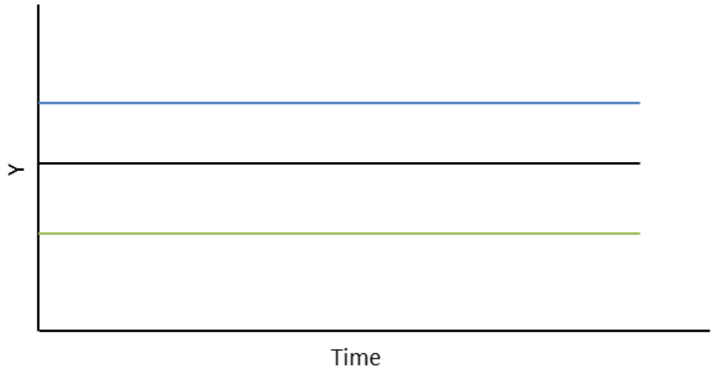


Figure 5:3 Two-level model – Subject level random intercept, no fixed or random effect of time

When a fixed effect for time is included

$$y_{ik} = \beta_0 + \beta_1 t_{ik} + u_k + e_{ik}$$

each subject sees a β_1 increase in y for each unit of time, with subject-specific lines again parallel to the average. Here there is a liner trend for time which is the same for each subject, Figure 5:4.

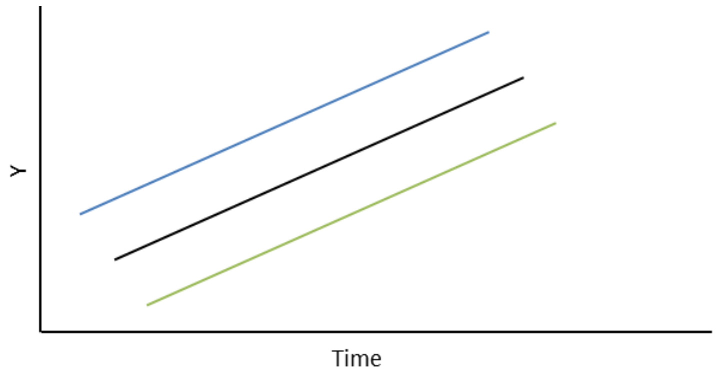


Figure 5:4 Two-level model – Subject level random intercept, fixed effect of time

In contrast, in a model with a random effect of time with no fixed effect

$$y_{ik} = \beta_0 + u_{0k} + u_{1k}t_{ik} + e_{ik}$$

each subject k can have a different slope u_{1k}. This allows the subject-specific effects of time on outcome to vary but only such that the overall effect of time is assumed to be zero, as depicted in Figure 5:5.

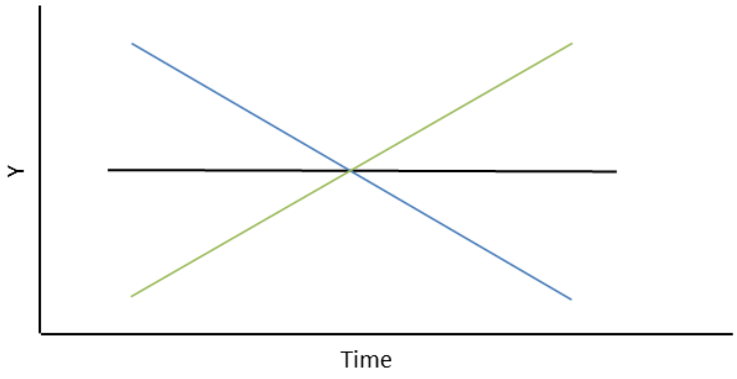


Figure 5:5 Two-level model – Subject level random intercept and random gradient for time

Finally, when both fixed and random effects for time are included in the model

$$y_{ik} = \beta_0 + \beta_1 t_{ik} + u_{0k} + u_{1k}t_{ik} + e_{ik}$$

each subject is allowed their own effect of time on outcome $(\beta_1 + u_{1k})$ which is allowed to differ from population average slope β_1.

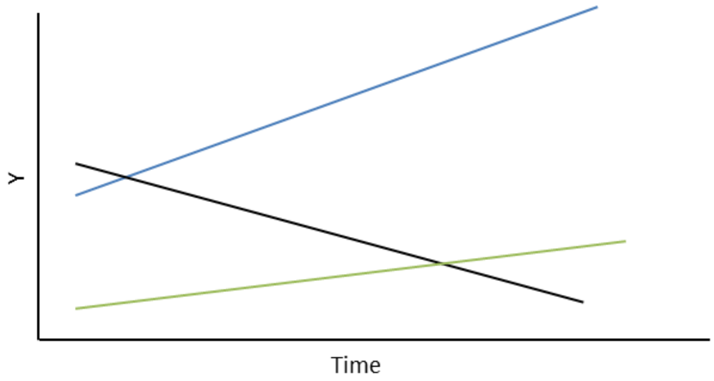


Figure 5:6 Two-level model – Subject level random intercept, both fixed and random effect of time

5.2.2 Three-level paradigm for ESM data

When expanding to three level models, where moments are nested within days, there are more options to consider. In the three-level ESM data structure time can be defined in two ways: as within-day time t_{ijk} (time of beep or beep number) and between-day time s_{jk} (day number). Defining time trends in three level models now requires a combination of a within- and between-day time variables in the fixed effects and at each level of the random effects.

In the two-level models, random effects were only required at the subject level; as a random intercept u_{0k} and as a random slope u_{1k} for time. For three level models, in which levels are now defined as moments $i = 1, \dots, n_{1jk}$, days $j = 1, \dots, n_{2k}$, and subjects $k = 1, \dots, n_3$, there is an additional random intercept for day, v_{0jk}, and the possibility for random slopes at the day level, v_{1jk}. The four two-level models based on the four permutations of the fixed and random effects can now be extended to 32 models as there are now two fixed terms, within-day time t_{ijk} and day number s_{jk}, and three possible random slopes: t_{ijk} and s_{jk} at level 3, and t_{ijk} at level 2. Each can be included or omitted giving $2^5 = 32$ model combinations. These models are defined in Table 5:1 with their corresponding graphical illustration provided in Figure 5:7. In these graphs the black line again represents the average slope and the blue and green lines depict two example subject-specific slopes. The dashed vertical lines separate the days of measurement, i.e. each graph depicts an example of two participants measured over three days. For ease of interpretation of time trends, day-level random intercepts are not depicted on the graphs. The reader is advised to assume random variation in daily intercept lines but to interpret departures from initial intercept in the graphs as a between-day time trend.

The equations and graphs will be labelled 1a) in the top left of the array to 8d) in the bottom right, numbers denoting rows and letters columns. To aid interpretation in this example it will be assumed that measurements of individuals' mood are collected over seven days. As such, within-day trends will be referred to as daily trends and between-day trends will be referred to as weekly trends.

Fixed effect interpretation

These three level models are comprised of fixed effects and random effects, the fixed effects representing the average trends and the random effects the subject and day variations in these trends. The 32 models fit a combination of daily and weekly time trends in the fixed part of the model. When modelling within-day time t_{ijk} (rows 2, 4, 6, 8), the corresponding coefficient β_1 is estimated as the average daily time trend. If $\beta_1 > 0$ a positive within-day trend is present, i.e. mood improves over the course of the day. If $\beta_1 < 0$, on average, mood deteriorates throughout the day, and if $\beta_1 = 0$ no trend is observed.

In the models fitting day number s_{jk} (columns b and d), the corresponding coefficient β_2 represents the average weekly trend. If $\beta_2 > 0$ the model suggests that, on average, mood improves over the week, if $\beta_2 < 0$ on average mood worsens and if $\beta_2 = 0$ there is no observed weekly trend.

The fixed intercept β_0 in each of the models represents the average mood score when all fixed effects are equal to zero. The interpretation of β_0 thus relies on meaningful zero values for all covariates. As there is no natural zero value for time of prompt or day number these variables will need to be centred. The choice of centring point, as discussed in Chapter 1, will influence both the interpretation of β_0 and the variance of the random effects. As we are interested in studying trends over the course of the day and week, both time variables will be centered such that the zero value relates to the start of the sampling period, i.e. day numbers_{jk} will be centred at 1 and within-day time t_{ijk} if moment number is used as a proxy for time this too will be centered at 1, otherwise this variable will be centered around the time of the first prompt of the day. The intercept β_0 can therefore be interpreted as the average mood score at the start of the day/week.

The inclusion of t_{ijk} to model daily trends restricts the trend to be the same each day, equal to β_1. This assumption can be relaxed to investigate whether daily trends differ each day. As will be discussed, allowing daily trends to vary across the study period can be achieved by modelling t_{ijk} as a level 2 random slope, however, this can also be modelled in the fixed effects. Adding an interaction between within-day and between-day time $(t_{ijk} * s_{jk})$ in the fixed part of the model,

$$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk} + \beta_3 t_{ijk} s_{jk} + (random\, part)$$

β_1 now represents the within-day time trend for Day 1 and $\beta_1 + \beta_3 s_{jk}$ the trends for each subsequent day $s_{jk} = 1, \dots, n_{2k} - 1$. Alternatively, s_{jk} can be modelled as a categorical variable to obtain separate estimates for each day. For simplicity, fixed effect interactions have been omitted from the equations in Table 5:1 but should be included when it is expected that population average daily trends vary across days. Adding random effects for within-day time (beep number) at levels 2 will then allow these trends to vary by subject each day. Adding random effects for within-day time at levels 3 will allow these trends to vary between subjects

In addition to linear trends, non-linear trends can be studied by fitting higher order polynomials in the fixed effects. Interpretation of the model is now the combination of the mean response over time and the rate of change in y dependent on time. For a p order polynomial,

$$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 t_{ijk}^2 + \beta_3 t_{ijk}^3 + \dots + \beta_p t_{ijk}^p + (random\, part)$$

relates to the mean response over time and

$$\frac{dy}{dt} = \beta_1 + 2\beta_2 t_{ijk} + 3\beta_3 t_{ijk}^2 + \dots + p\beta_p t_{ijk}^{p-1} + (random\, part)$$

the rate of change. The magnitude and sign of the coefficients can be used to draw inference on the shape and direction of the trend. For a quadratic trend, for example, the turning point can be found by solving $dy/dt = 0$, the sign of d^2y/dt^2 indicating whether the point is a maximum or minimum. Moreover, including higher order terms in the random effects can allow these non-linear trends to vary between- and within-subject.

Random effect interpretation

The random intercepts for subject u_{0k} and day v_{0jk} allow a different intercept for each subject and each day within-subject respectively, i.e. a specific mood score at the start of the study can be estimated for each subject, $\beta_0 + u_{0k}$, which can differ by v_{0jk} for each day within subject. For example, Subject 1 will have a mood score $\beta_0 + u_{01} + v_{011}$ at the start of Day 1 but $\beta_0 + u_{01} + v_{021}$ at the start of Day 2.

The random slopes u_{1k} and v_{1jk} allow different time trends for each individual in the same way as the random intercepts. The choice of time variable at each level of the random effects will determine which trend varies and how. For weekly trends, including day s_{jk} at level 3 will allow for a different trend per person, each subject k now having their own weekly trend $\beta_2 + u_{1k}$. These effects are present the models in column d. Where models include a random effect but no fixed effect for s_{jk}, as in models in column c (i.e. when $\beta_2 = 0$), they still allow for a different weekly trend per person and can be used either when the average trend is of no interest but for model fit the between-subject variation is still included, or when the average effect is expected to be zero, i.e. the subject-specific slopes average to zero. The variance in random slopes $\left(var(u_{1k}) = \sigma_{u_1}^2\right)$ estimates the degree of variation in weekly trends. If $\sigma_{u_1}^2$ is large, there is substantial heterogeneity in subject-specific weekly trends. If $\sigma_{u_1}^2$ is small, individuals experience much more similar weekly trends in mood.

For within-day trends, the models fitting a random slope for t_{ijk} at level 3 (models on rows 5-8) allow each subject a different daily trend, $\beta_1 + u_{1k}$, but the trend is expected to be the same each day. For example, if the population average within-day trend is positive ($\beta_1 > 0$), i.e. on average subjects see an improvement in mood over the course of the day, Subject 1 may also have a positive trend ($\beta_1 + u_{11} > 0$, where $u_{11} > 0$) indicating they see a greater improvement in mood over the day than average, whereas Subject 2 may have a negative within-day trend ($\beta_1 + u_{21} < 0$) suggesting their mood declines over the day. The variance in random slopes describes the degree to which individuals share similar daily time trends. A large variance suggests people are experiencing very different daily mood trends to one another, a small variance would imply a more homogeneous daily mood profile.

When a random slope for t_{ijk} is fitted at level 2 (models on rows 3, 4, 7 and 8) the within-day time trend is allowed to differ each day within-subject; the trend $\beta_1 + v_{1jk} + u_{1k}$ containing additional variation v_{1jk} for each day j for each subject k. The model estimates $var\left(v_{1jk}\right) = \sigma_{v_1}^2$, the variation in day-specific slopes for subject. Unlike $\sigma_{u_1}^2$, this variance is comparing daily trends within subject. If $\sigma_{v_1}^2$ is large, within-day trends vary greatly one day to the next. If $\sigma_{v_1}^2$ is small the daily trends can be assumed to be largely the same within each subject.

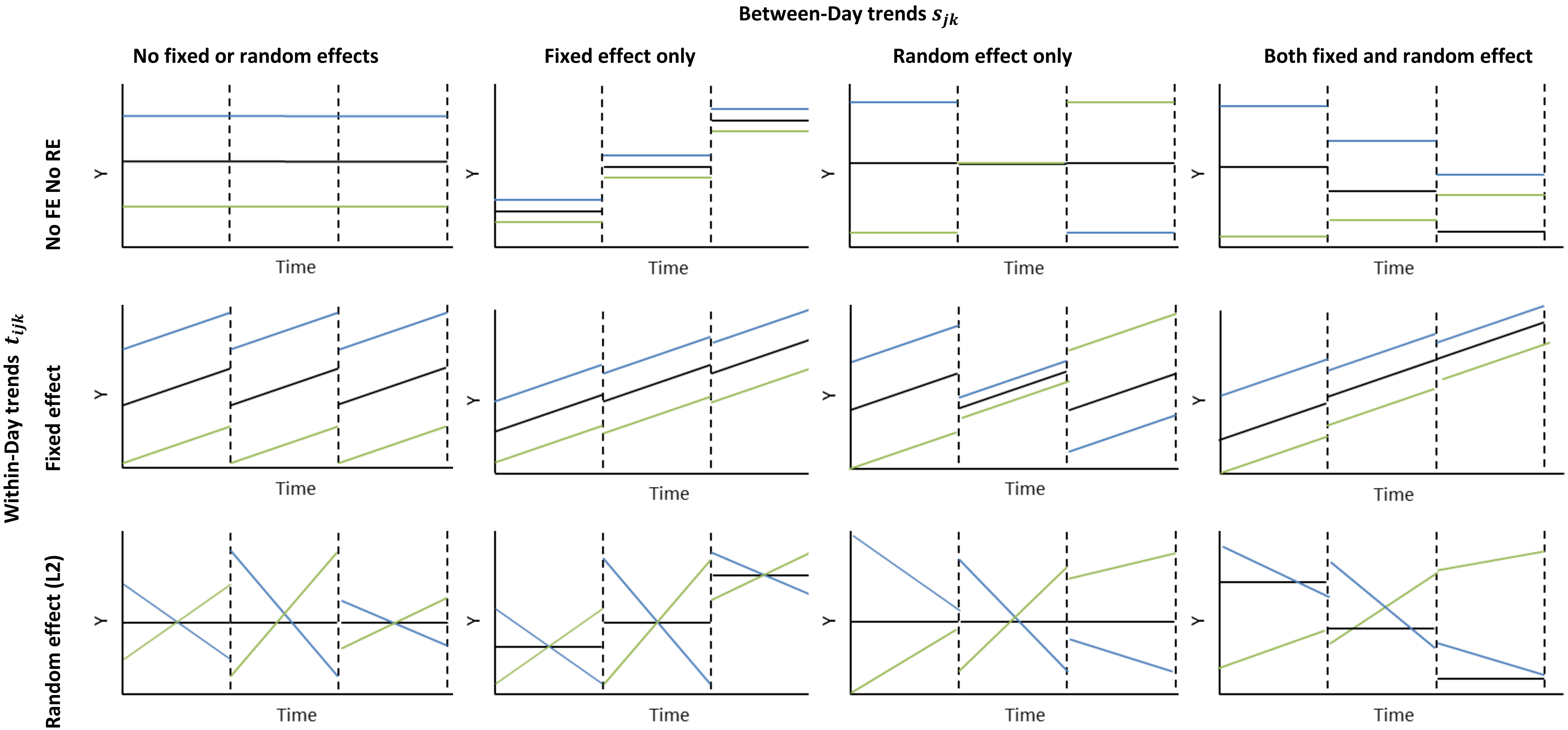
In addition to the covariance between random intercept and slope, models 5c-8c and 5d-8d estimate the covariance between the random slopes t_{ijk} and s_{jk}, σ_{u12}. In these models σ_{u12} represents the relationship between daily and weekly time trends. When $\sigma_{u12} > 0$ positive daily trends result in positive weekly trends, or negative daily trends in negative weekly trends. When $\sigma_{u12} < 0$, the results indicate that while daily trends may be positive over the week mood decreases (or vice versa).

Likelihood ratio tests can be used with nested models to determine the significance of the parameters. However, as variance estimates must be non-negative, likelihood ratio tests for examining the significance of random effects may not be valid as the null hypothesis (that the variance parameter is equal to zero) is on the boundary of the parameter space (Fitzmaurice, Laird et al. 2012; Rabe-Hesketh and Skrondal 2012). The distribution of the null hypothesis is usually chi-squared with the degrees of freedom equal to the difference in the number of parameters in the full model minus those in the reduced model. When on the boundary, the null distribution is now a combination of two chi-squared distributions. As a consequence, significance tests will be conservative, producing an over estimated p-value and resulting in overly parsimonious models. An ad hoc solution to this is to simply divide the naïve p-value by 2. Alternatively, the correct p-value can be calculated when the distribution mix is known. The analysis of the presented random slope models will be conducted in Stata 13, which Rabe-Hesketh and Skrondal (2012) state due to the default estimation metric for the covariance of the random effects, the asymptotic null distribution when testing the significance of the variance of the qth random effect is $0.5\chi_q^2 + 0.5\chi_{q+1}^2$, which can be computed.

		Between-Day Effects s_{jk}			
Within-Day Effects t_{ijk}	RE at Level 2	No fixed or radom effects	Fixed effects	Random effects	Both fixed and random effects
	No FE No RE	$y_{ijk} = \beta_0$ $+u_k + v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_k + v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{jk} + e_{ijk}$
	Fixed effect	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_k + v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_k + v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{ijk}$ $+u_{0k} + u_{1k} t_{jk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k} t_{jk}$ $+v_{jk} + e_{ijk}$
	Random effect	$y_{ijk} = \beta_0$ $+u_k + v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_k + v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$
	Both FE and RE	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_k + v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{ijk} + \beta_2 t_{jk}$ $+u_k + v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k} s_{jk}$ $+v_{0jk} + v_{1jk} t_{ijk}$ $+e_{ijk}$

Within-Day Effects t_{ijk}	**RE at level 3**	**Random effect**	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{jk} + e_{ijk}$
		Both FE and RE	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{jk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{jk} + e_{ijk}$
	RE at levels 2 and 3	**Random effect**	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$
		Both FE and RE	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$	$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 s_{jk}$ $+u_{0k} + u_{1k}t_{ijk} + u_{2k}s_{ijk}$ $+v_{0jk} + v_{1jk}t_{ijk} + e_{ijk}$

Table 5:1 Three-level fixed and random time effects array


Between-Day trends s_{jk}
No fixed or random effects
Fixed effect only
Random effect only
Both fixed and random effect
Within-Day trends t_{ijk}
No FE No RE
Fixed effect
Random effect (L2)
Y
Time

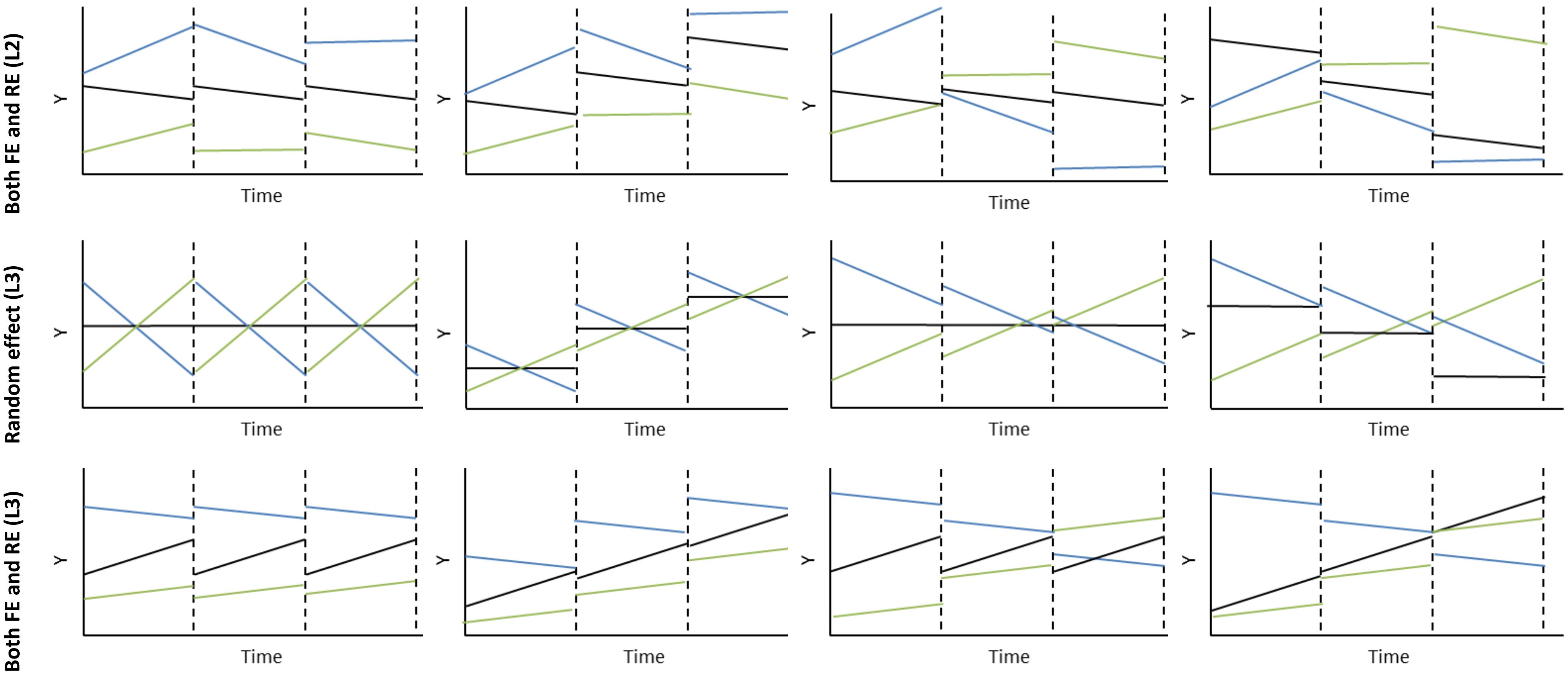


Figure 5:7 Three-level fixed and random time effect array

Graphs for when within-day trends are modelled as random effects at both level 2 and level 3 (rows 7 and 8 in Table 5:1) are omitted as they are analogous to rows 3 and 4 (within-day random effects at level 2).

5.3 Applications to data

The methods outlined for using random slope models to summarize variation in three-level data and to investigate time trends will be applied to the Recovery study data. These data are a subset of the data collected for an ESM study investigating participant reported feelings of recovery in a sample of individuals diagnosed with schizophrenia. The data are summarized in Chapter 2. Firstly, the proportion of variance at each level will be calculated to provide a general description of how recovery scores vary at the subject-level, day-level and moment-level. Random slope models will then be used to further investigate this variation and determine whether there are trends in feelings of recovery over time. These trends will be explored both linearly and non-linearly. Finally, it will be demonstrated how variation in a level 1 measure can be examined between groups, the example extending the original analysis of Chapter 2, and across a continuous variable.

5.3.1 Variance components model

A random intercept model with no covariates, otherwise known as a variance components model, can be used to examine the proportion of variance in outcome at each level of the data. As ESM are often analysed as two-level data, both a two- and three-level variance components models will be presented to demonstrate day-level variation. The two-level variance components model is defined as

$$y_{ik} = \beta_0 + u_k + e_{ik}$$

where measurements $i = 1, \dots, n_1$ are nested within subjects $k = 1, \dots, n_2$. As the recovery data are balanced, the subscript k can be dropped from n_1. The subject-level random intercepts are denoted $u_k \sim N(0, \sigma_u^2)$ and the level 1 residuals $e_{ik} \sim N(0, \sigma_e^2)$. The three-level variance components model is defined as

$$y_{ijk} = \beta_0 + u_k + v_{jk} + e_{ijk}$$

with measurements $i = 1, \dots, n_1$ nested within days $j = 1, \dots, n_2$ nested within subjects $k = 1, \dots, n_3$, where $u_k \sim N(0, \sigma_u^2)$ is the subject level random intercept, $v_{jk} \sim N(0, \sigma_v^2)$ is the day level random intercept and $e_{ij} \sim N(0, \sigma_e^2)$ are the level 1 residuals. The results of these two models are presented in Table 2.

Two level model			Three level model		
Random Effects	Variance	Std. error	Random Effects	Variance	Std. error
Subject level: σ_u^2	1.648	0.295	Subject level: σ_u^2	1.627	0.298
			Day level: σ_v^2	0.217	0.026
Residual: σ_e^2	0.667	0.019	Residual: σ_e^2	0.496	0.015

Table 5:2 Estimates of a two- and three-level variance components model for recovery

The significance of the day-level random effect can then be tested using a likelihood ratio test of the two models where the null hypothesis H_0: $\sigma_v^2 = 0$ is tested against the alternative hypothesis H_1: $\sigma_v^2 \neq 0$. The log likelihoods for the two level model $(\mathrm{LL} = -3161.013)$ and the three level model $(\mathrm{LL} = -2997.785)$ rejects the null hypothesis $(\chi_1^2 = 326.46, \text{adjusted } p < 0.0001)$ at the 1% level, indicating the day level random intercept is significant and a three-level model should be used.

The proportion of variance, as described in Section 2.1.4.1, can then be calculated for each level of the three-level model:

$$Level\ 3\text{:} \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2} = \frac{1.627}{1.627 + 0.22 + 0.496} = 0.694$$

$$Level\ 2\text{:}\ \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2} = \frac{0.217}{1.627 + 0.217 + 0.496} = 0.093$$

$$Level\ 1\text{:}\ \frac{\sigma_e^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2} = \frac{0.496}{1.627 + 0.217 + 0.496} = 0.214$$

These results show that approximately 70% of the variation in recovery scores lies between subjects, 9% is between days within subject and 21% is moment to moment variation within days. The high between-subject variation indicates that individuals in this sample report very different levels of recovery to one another and that in comparison, day to day variation in recovery within-subject is small. In addition, the correlation between two observations within the same day, within subject can be calculated as

$$corr(y_{ijk}, y_{i'jk}) = \frac{\sigma_v^2 + \sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2} = \frac{0.217 + 1.627}{1.627 + 0.217 + 0.496} = 0.788$$

which suggests that recovery scores within a subject are highly correlated moment to moment. This correlation can be investigated using time trends.

5.3.2 Time trends in recovery data

Although the models defined in Section 1.2.2 can be used to explore patterns in the data, model choice here will be driven by specific research questions. Model 1 will be used to assess the time course of recovery across the week – on average how does recovery change over the week? – and by allowing each subject to have a different slope, how similar are subject specific trends? Model 2a tests for a within-day time trend – how does recovery change within a day? – with each subject allowed to have a different within-day slope to assess how similar these trends are between subjects. Model 2b allows the within-day slope to vary each day, this model describing how different daily trends are both between and within subjects. The model equations are presented in Table 5:3.

A random intercept model with a fixed effect for both within-day and between-day time will be used as a base model to compare to the random slope models to test for time trends. A fixed effect of time t_i will describe how on average recovery changes over the course of a day and s_j how recovery changes over the week. The subscripts j and k are dropped from t_i as beep number will be used as a proxy for within-day time and each subject is instructed to complete 10 beeps each day. Similarly, day number s_j requires no subscript k as each subject is monitored for an equal number of days.

The results of each of these models can be seen in Table 5:4.

	Random Intercept model	$y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 s_j + u_k + v_{jk} + e_{ijk}$
Model 1	Between-day random slope model	$y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 s_j + u_{0k} + u_{1k} s_j + v_{jk} + e_{ijk}$
Model 2a	Within-day random slope model	$y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 s_j + u_{0k} + u_{1k} t_i + v_{jk} + e_{ijk}$
Model 2b	Within-day random slope model with across-day variation	$y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 s_j + \beta_3 t_i s_j$ $+u_{0k} + u_{1k} t_i + v_{0j} + v_{1jk} t_i + e_{ijk}$

Table 5:3 Time trend models to be applied to the recovery data

5.3.2.1 Results of time trend models

Between-day time trends – Model 1

There is small downward weekly trend $(\beta_2 = -0.04, p = 0.044)$ which suggests that on average participants feel less recovered over the course of the week. The variance in subject-specific slopes $(\sigma_{u1}^2 = 0.02)$ indicated that 95% of slopes within $\pm\sqrt{0.02} * 1.96 =$

± 0.28 units of the average, that is 95% of subjects have weekly trends between -0.32 and 0.24. The likelihood ratio test comparing this model with the random intercept model indicates that the random slope is significant ($\chi_1^2 = 20.06, \text{adjusted } p = 0.0001$). Variation in subject-specific intercepts is larger, with a range of 4.6 units for initial recovery scores. The positive covariance ($\sigma_{uo1} = 0.05$) indicates that participants with higher recovery scores at the start of the week have a stronger than average negative trend over the week.

Within-day time trends (models 2a and 2b)

There is no evidence of linear daily time trends in recovery ($\beta_1 = 0.003, p = 0.770$, model 2a) and no apparent variation in trend across days ($\beta_3 = -0.002, p = 0.606$, model 2b). However, a likelihood ratio test comparing model 2a with the random intercept model suggests the subject-level random slope is significant ($\chi_2^2 = 43.36$, adjusted $p < 0.0001$). Similarly, the day-level random slope was found to be significant when comparing model 2b with 2a (with the addition of the day by beep interaction term in the fixed effects) ($\chi_2^2 = 38.01$, adjusted $p < 0.0001$). Interpreting these random slope coefficients suggests there is a small amount of variation in daily trends at the subject level ($\sigma_{u1}^2 = 0.003$ model 2a , $\sigma_{u1}^2 = 0.002$ model 2b) with model 2b suggesting a greater amount of variation in slopes at the day level ($\sigma_{v1}^2 = 0.006$), i.e. there is more variation in time trends one day to the next than there is between subjects.

		Random intercept model			Model 1 Between-day RS model			Model 2a Within-day RS model			Model 2b Across-day RS model		
Fixed effects		β	Std. Error	P-value	β	Std. Error	P-value	β	Std. Error	P-value	β	Std. Error	P-value
Intercept		5.028	0.169	<0.001	5.031	0.154	<0.001	5.035	0.167	<0.001	5.010	0.172	<0.001
Beep (t_i)		0.003	0.005	0.601	0.002	0.005	0.645	0.003	0.009	0.770	0.008	0.013	0.520
Day (s_j)		-0.047	0.017	0.009	-0.045	0.022	0.044	-0.046	0.017	0.008	-0.036	0.025	0.141
Beep*Day		--	--	--	--	--	--	--	--	--	-0.002	0.004	0.606
Random effects			Variance	Std. Error		Variance	Std. Error		Variance	Std. Error		Variance	Std. Error
Level 3	Intercept		1.654	0.305		1.351	0.266		1.597	0.303		1.575	0.303
	Day (slope)		--	--		0.014	0.006		--	--		--	--
	Covariance (I, D)		--	--		0.049	0.027		--	--		--	--
	Beep (slope)		--	--		--	--		0.003	0.001		0.002	0.001
	Covariance (I, B)		--	--		--	--		0.0001	0.012		0.005	0.012
Level 2	Intercept		0.210	0.025		0.163	0.024		0.218	0.026		0.333	0.051
	Beep (slope)		--	--		--	--		--	--		0.006	0.001
	Covariance (I, B)		--	--		--	--		--	--		-0.025	0.007
Level 1	Residual		0.497	0.015		0.497	0.015		0.473	0.015		0.437	0.014
Log likelihood			-3002.683			-2992.663			-2981.001			-2966.573	

Table 5:4 Fixed and random effect estimates of time trend models fitted to the recovery data

5.3.2.2 Incorporating non-linear trends

Further exploration of the time trends in recovery can include investigating non-linear trends. The results of 5.3.2.1 show no linear trend in beep number, however, a non-linear trend may be present. The simplest way to incorporate non-linear trends is by including higher order time terms.

To begin, the three-level random intercept model was refitted with the inclusion of quadratic trends in both beep number and day number

$$y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 s_j + \beta_4 s_j^2 + u_k + v_{jk} + e_{ijk}$$

		Random intercept – Full			Random intercept - Reduced		
Fixed effects		β	Std. Error	P-value	β	Std. Error	P-value
Intercept		4.959	0.174	<0.001	4.968	0.170	<0.001
Beep		0.043	0.020	0.026	0.043	0.020	0.026
Beep^2		-0.004	0.002	0.030	-0.004	0.002	0.030
Day		-0.029	0.060	0.623	-0.044	0.017	0.010
Day^2		-0.003	0.012	0.791	--	--	--
Random effects		Variance	Std. Error		Variance	Std. Error	
Level 3	Intercept	1.628	0.298		1.628	0.298	
Level 2	Intercept	0.210	0.025		0.210	0.025	
Level 1	Residuals	0.495	0.015		0.495	0.015	
Log likelihood		-2991.930			-2991.965		

Table 5:5 Full and reduced quadratic random intercept models fitted to the recovery data

The results show there to be a significant negative quadratic trend in within-day time, indicating that recovery decreases non-linearly over the course of the day. The quadratic day trend is does not improve model fit and so will not be included in the subsequent models.

Models 2a and 2b were refit with both the linear and quadratic 'beep number' terms to investigate how the non-linear daily trends vary between subjects. These models are now expressed as models 3a and 3b:

$$Model\ 3a: y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 s_j + u_{0k} + u_{1k} t_i + u_{2k} t_i^2 + v_{jk} + e_{ijk}$$

$$Model\ 3b: y_{ijk} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 s_j + \beta_4 t_i s_j + \beta_5 t_i^2 s_j + u_{0k} + u_{1k} t_i + u_{2k} t_i^2 + v_{0j} + v_{1jk} t_i + v_{2k} t_i^2 + e_{ijk}$$

The quadratic function $\beta_1 t_i + \beta_2 t_i^2$ in the fixed effects will describe the population average trend, with turning point where $dy/dt = 0$. Both terms are also included in the random effects to allow these trends to vary between-subjects (and between-day in Model 3b). As solving the first derivative of the equation provides the turning point at time t, the random effect $u_{1k} t_i$ allows for variation in subject-specific (and day-specific) turning points while the term $u_{1k} t_i^2$ allows for subject-specific (and day-specific) gradients.

These full models were compared to a reduced form for each, only including the linear term in the random effects and in the interaction of model 3b. Whilst both converge, the quadratic variance estimates are very small ($Model\ 3a$: $\sigma_{u_2}^2 = 0.00009$, for example) and the variance standard errors could not be calculated whilst specifying unstructured covariance matrices at levels 2 and 3; this model is simply too complex for the data. The results of reduced forms of these models, with the linear time term in the random effects only, are presented in Table 5:6.

For each of the reduced models there is roughly the same within-day trend: recovery increases slightly as the day progresses followed by a turning point towards the middle of the day (5.5 in model 3a and 6 in model 3b) after which the recovery starts to decrease with time.

The significance of the between-day and within-day random slopes were tested using likelihood ratio tests. The linear random slope at level 3 were found to be significant $(\text{Model 3a}: \chi_2^2 = 43.64, \text{adjusted}\ p < 0.0001)$ as was the inclusion of the linear random slope at level 2 $(\text{Model 3b}: \chi_2^2 = 37.91, \text{adjusted}\ p < 0.0001)$; the within-day models demonstrating significant variation in daily trends between subjects and days.

		Random intercept model			Model 3a Within-day RS model			Model 3b Across-day RS model		
Fixed effects		β	SE	P-value	β	SE	P-value	β	SE	P-value
Intercept		4.968	0.170	<0.001	4.972	0.170	<0.001	4.949	0.174	<0.001
Beep		0.044	0.020	0.026	0.044	0.021	0.035	0.048	0.022	0.031
Beep^2		-0.004	0.002	0.030	-0.004	0.002	0.028	-0.004	0.002	0.029
Day		-0.044	0.017	0.010	-0.046	0.017	0.008	-0.036	0.025	0.141
Beep*Day								-0.002	0.004	0.614
Random Effects			Variance	SE		Variance	SE		Variance	SE
Level 3	Intercept		1.628	0.298		1.594	0.303		1.573	0.302
	Day (slope)		--	--		--	--		--	--
	Covariance (I, D)		--	--		--	--		--	--
	Beep (slope)		--	--		0.003	0.001		0.002	0.001
	Covariance (I, B)		--	--		0.001	0.012		0.005	0.012
Level 2	Intercept		0.210	0.025		0.218	0.026		0.332	0.051
	Beep (slope)		--	--		--	--		0.006	0.001
	Covariance (I, B)		--	--		--	--		-0.025	0.007
Level 1	Residual		0.495	0.012		0.472	0.015		0.436	0.014
Log likelihood			-2991.965			-2983.866			-2969.404	

Table 5:6 Fixed and random effect estimates of reduced non-linear time trend models fitted to the recovery data – linear terms only in the random effects

5.3.2.3 Expanding trends into groups

One of the main research questions attached to the recovery data was the degree to which participant reported recovery fluctuated and whether this differed between groups categorised at baseline as ‘recovered’ or ‘non-recovered’. To address this question a three level random intercept model with complex level 1 variation was fitted to the data:

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_k + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1)$$

where x_k is the binary group variable $x_k = 0$ the non-recovered group and $x_k = 1$ the recovered group. The residuals were estimated separately for each group, $e_{0ijk} \sim N\left(0, \sigma_{e_0}^2\right)$

the residuals for the non-recovered group and $e_{1ijk} \sim N\left(0, \sigma_{e_1}^2\right)$ for the recovered group. The results (presented in Chapter 1.3) indicated that the baseline ‘recovered’ group had significantly higher reported ESM recovery than the ‘non-recovered’ group, although the residual variance suggested there was more variation in the ‘non-recovered’ group’s scores. Combining this model with the random slope methods outlined in this chapter, we can further investigate the group differences in participant reported recovery. Firstly, weekly trends will be compared between groups followed by group differences in daily trends. For clarity, the non-recovered baseline group will be referred to as Group 1 and the recovered baseline group as Group 2.

GROUP DIFFERENCES IN WEEKLY TRENDS

A random intercept model was again used to test the significance of the random slopes, the nested structure of the models enabling likelihood ratio tests for model comparisons. Firstly, weekly trends were compared between baseline groups using a between-day random slope model with nonlinear weekly trends

$$\begin{aligned} y_{ijk} = {} & \beta_0 + \beta_1 x_k + \beta_2 s_j + \beta_3 s_j^2 + \beta_4 x_k s_j + \beta_5 x_k s_j^2 \\ & + u_{ok} + u_{1k} s_{jk} + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1) \end{aligned}$$

where average weekly trends for Group 1, $\beta_2 s_j + \beta_3 s_j^2$, and Group 2, $(\beta_2 + \beta_4)s_j + (\beta_3 + \beta_5)s_j^2$, are separately estimated with additional variation u_{1k} for subject specific trends. As in the random intercept model, separate residuals are estimated for the two baseline groups.

Although there is a greater amount of variation in the ESM reported recovery scores for Group 1, there appears to be no significant week trend. Group 2, however, display a significant quadratic trend, with ESM reported recovery decreasing at the start of the week and then increasing after Day 2, as displayed in Figure 5:8. Finally, there is significant variation in subject-specific weekly trends ($\sigma_{u1}^2 = 0.014$; $\chi_1^2 = 17.69$, adjusted $p = 0.0003$).

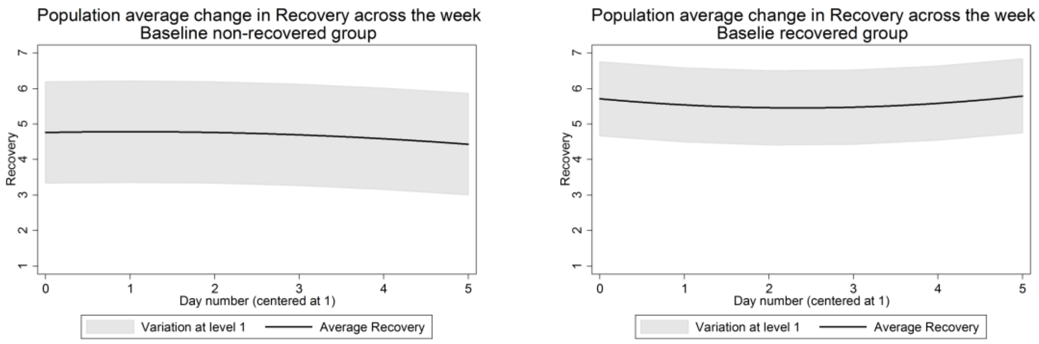


Figure 5:8 Variation in weekly recovery for the two baseline groups: population average trends plus variation at level 1

GROUP DIFFERENCES IN DAILY TRENDS

Quadratic within-day trends were also explored in the two groups, using the random slope model

$$y_{ijk} = \beta_0 + \beta_1 x_k + \beta_2 t_i + \beta_3 t_i^2 + \beta_4 x_k t_i + \beta_5 x_k t_i^2 \\ + u_{ok} + u_{1k} t_i + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1).$$

This model estimates the average daily time trend $\beta_2 t_i + \beta_3 t_i^2$ for Group 1 and $(\beta_2 + \beta_4)t_i + (\beta_3 + \beta_5)t_i^2$ for the 'recovered' group. In the random part of the model, separate residual variances are estimated for the two baseline groups in order to measure the degree of variability at the momentary level, while the level-3 random slope accounts for the subject level variation in the daily trends.

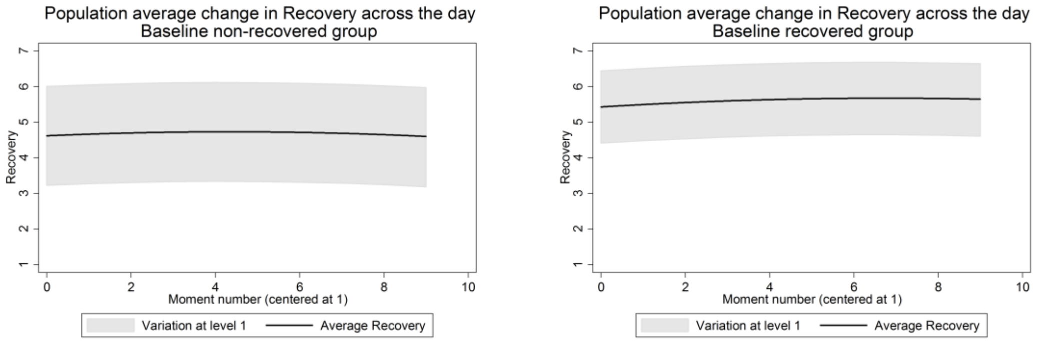


Figure 5:9 Variation in daily recovery for the two baseline groups: population average trends plus variation at level 1

Group 1 exhibit a significant quadratic daily trend, with ESM reported recovery increasing over the morning and beginning to decline around Beep 5. Group 2 display similar daily trends.

		Weekly trends (quadratic dayXgroup interaction)			**Daily trends** (quadratic beepXgroup interaction)		
Fixed effects		β	SE	P-value	β	SE	P-value
Intercept		4.760	0.186	<0.001	4.617	0.203	<0.001
Group		0.949	0.328	0.004	0.810	0.356	0.023
Beep (Not recovered)		--	--	--	0.052	0.026	0.049
Beep^2 (Not recovered)		--	--	--	-0.006	0.003	0.018
Beep (Recovered)		--	--	--	0.021	0.040	0.592
Beep^2 (Recovered)		--	--	--	0.001	0.004	0.884
Day (Not recovered)		0.044	0.070	0.535	--	--	--
Day^2 (Not recovered)		-0.022	0.013	0.100	--	--	--
Day (Recovered)		-0.264	0.122	0.030	--	--	--
Day^2 (Recovered)		0.069	0.023	0.003	--	--	--
Random effects			Variance	SE		Variance	SE
Level 3	Intercept		1.251	0.254		1.557	0.304
	Day (slope)		0.014	0.006		--	--
	Covariance (I, D)		0.042	0.028		--	--
	Beep (slope)		--	--		0.002	0.001
	Covariance (I, B)		--	--		-0.009	0.011
	Covariance (B, D)		--	--		--	--
Level 2	Intercept		0.166	0.024		0.233	0.027
Level 1	Residual (Not recovered)		0.529	0.020		0.503	0.019
	Residual (Recovered)		0.282	0.016		0.268	0.016
Log likelihood			-2703.862			-2702.454	

Table 5:7 Within- and between-day random slope group difference models fitted to the recovery data with quadratic time trends and complex level 1 variance

Time trends in residual variance between groups

The variation in recovery at level 1 was largely similar in the two models. For both models the recovery scores for the 'non-recovered' group fluctuated more greatly than the 'recovered' group, this degree of fluctuation expressed as the shaded areas in Figure 5:8 and Figure 5:9. To examine this question of 'fluctuation' more closely, a complex variation model allowing for differences in group residuals to depend on time can be fitted.

For weekly trends the model becomes

$$\begin{aligned} y_{ijk} = {} & \beta_0 + \beta_1 x_k + \beta_2 s_{jk} + \beta_3 s_j^2 + \beta_4 x_k s_{jk} + \beta_5 x_k s_j^2 + u_{0k} + u_{1k} s_{jk} \\ & + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1) + e_{2ijk} s_j \end{aligned}$$

where

$$\begin{bmatrix} e_{0ijk} \\ e_{1ijk} \\ e_{2ijk} \end{bmatrix} \sim N(0, \Sigma_e):\ \Sigma_e = \begin{bmatrix} \sigma_{e0}^2 & & \\ 0 & \sigma_{e1}^2 & \\ \sigma_{e02} & \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix}$$

and the level 1 variance function is

$$\begin{aligned} & var\left(e_{0ijk} x_k(0) + e_{1ijk} x_k(1) + e_{2ijk} s_j\right) \\ & \qquad = \sigma_{e0}^2 x_k(0) + \sigma_{e1}^2 x_k(1) + 2\sigma_{e02} x_k(0) s_j + 2\sigma_{e12} x_k(1) s_j + \sigma_{e2}^2 s_j^2 \end{aligned}$$

The covariance of $x_k(0)$ and $x_k(1)$ is constrained to be zero as observations cannot be in both groups. The results of this model are presented in Table 5:8. The fixed effect estimates and levels 3 and 2 variance parameters are much the same as the previous complex variation model. The addition of day number as a level 1 variance parameter is statistically significant ($\chi_1^2 = 61.3, \text{adjusted } p < 0.0001$), indicating there is significant heteroskedasticity with respect to time. The week trends in level 1 variance for the two baseline recovery groups can be seen in Figure 5:11 where the variance function is plotted over days, the solid line representing Group 1 and the dashed line Group 2. Greater fluctuations in recovery can be observed in Group 1 at the start of the week, reducing as the week progresses. This reduction in variability, together with the fixed effect results plotted in Figure 5:10, imply that ESM recovery is higher but more variable at the start of the week and lower, but more consistently so at the end of the week for Group 1, the non-recovered group. Group 2, classified as 'recovered' at baseline, report higher ESM recovery scores than the 'non-recovered' group, with far less variability that stays constant across the week.

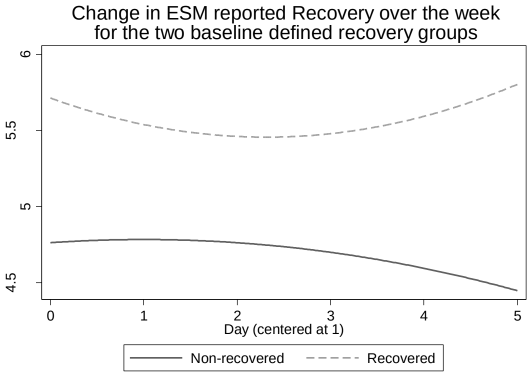


Figure 5:10 Non-linear population average weekly trends in recovery for two groups

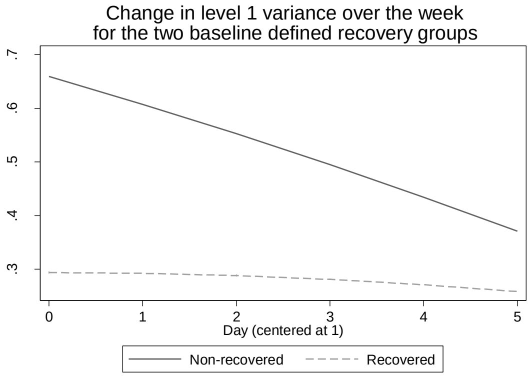


Figure 5:11 Time dependent heteroskedasticity between two groups: weekly trend

A similar extension can be applied to the within-day trend model to investigate whether group differences in outcome fluctuation change over time

$$\begin{aligned} y_{ijk} = {} & \beta_0 + \beta_1 x_k + \beta_2 t_i + \beta_3 t_i^2 + \beta_4 x_k t_i + \beta_5 x_k t_i^2 + u_{0k} + u_{1k} t_i \\ & + v_{jk} + e_{0ijk} x_k(0) + e_{1ijk} x_k(1) + e_{2ijk} t_i \end{aligned}$$

The fixed effect estimated (presented in Table 5:8), as in the weekly trend model, were largely similar to the model with constant level 1 variance. The parameters of interest here are the variance and covariance estimates at level 1 making up the variance function

$$\begin{aligned} & var\big(e_{0ijk} x_k(0) + e_{1ijk} x_k(1) + e_{2ijk} t_i\big) \\ & \qquad = \sigma_{e0}^2 x_k(0) + \sigma_{e1}^2 x_k(1) + 2\sigma_{e02} x_k(0) t_i + 2\sigma_{e12} x_k(1) t_i + \sigma_{e2}^2 t_i^2. \end{aligned}$$

Testing the significance of this complex variance against the constant variance model ($\chi_3^2 = 104.146, \text{adjusted } p < 0.0001$), it can be concluded that momentary level fluctuations in the two groups depends on time. This is illustrated in Figure 5:13 where the variance function is plotted for the two groups, the solid line representing level 1 variance over the day for Group 1 and the dashed line variation over the day for Group 2.

		Between-day random slope model (quadratic day group interaction)		Within-day random slope model (quadratic beep group interaction)	
Fixed effects		β	Std. Error	β	Std. Error
Intercept		4.761	0.184	4.615	0.201
Group		0.949	0.323	0.840	0.354
Beep (Not recovered)		--	--	0.053	0.027
Beep^2 (Not recovered)		--	--	-0.006	0.003
Beep (Recovered)		--	--	0.010	0.041
Beep^2 (Recovered)		--	--	0.002	0.004
Day (Not recovered)		0.043	0.070	--	--
Day^2 (Not recovered)		-0.022	0.013	--	--
Day (Recovered)		-0.263	0.121	--	--
Day^2 (Recovered)		0.069	0.023	--	--
Random effects		Variance	Std. Error	Variance	Std. Error
Level 3	Intercept	1.207	0.242	1.521	0.294
	Day (slope)	0.014	0.005	--	--
	Covariance (I, D)	0.041	0.026	--	--
	Beep (slope)	--	--	0.002	0.001
	Covariance (I, B)	--	--	-0.009	0.011
Level 2	Intercept	0.160	0.022	0.242	0.027
Level 1	Residual (Not recovered)	0.660	0.039	0.575	0.048
	Residual (Recovered)	0.294	0.023	0.480	0.053
	Residual (Day)	-0.001	0.002	--	--
	Covariance (NR, D)	-0.025	0.007	--	--
	Covariance (R, D)	0.000	0.000	--	--
	Residual (Beep)	--	--	0.010	0.002
	Covariance (NR, B)	--	--	-0.040	0.010
	Covariance (R, B)	--	--	-0.055	0.010
Log likelihood		-2675.6515		2684.403	

Table 5:8 Within- and between-day random slope group difference models fitted to the recovery data with quadratic time trends and complex, time dependent level 1 variance

Both groups exhibit similar patterns within-day, with ESM reported recovery improving over the course of the day, Group 1 taking a downward turn later in the day, but with the Group 2 exhibiting significantly higher levels of recovery throughout the day. Fluctuations in recovery also follow a similar pattern within each group, with more consistent scoring towards the middle of the day and higher variability in the morning and evening, Group 1 (the ‘non-recovered’ group), however, having higher levels of variability across the whole day.

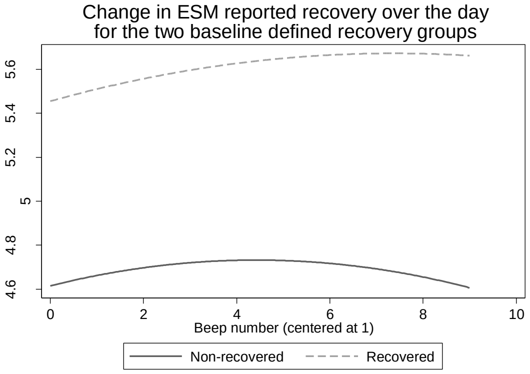


Figure 5:12 Non-linear population average daily recovery trends in two groups

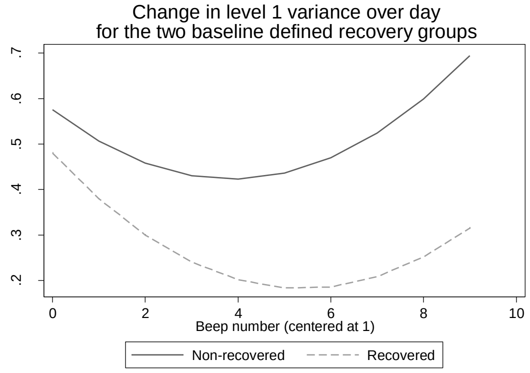


Figure 5:13 Time dependent heteroskedasticity between two groups: daily trend

The within-day trends of the current models have been restricted to be the same each day. However, in Section 5.3.2.1 within-day trends were extended to allow for across-day variability. Studying across-day variation between groups in the fixed effects and in terms of level 1 heteroskedasticity is conceptually possible, however it requires several two-way and three-way interactions. The fixed part of the within-day complex variation model above is extended to

$$\begin{aligned} y_{ijk} = {} & \beta_0 + \beta_1 x_k + \beta_2 t_i + \beta_3 t_i^2 + \beta_4 x_k t_i + \beta_5 x_k t_i^2 \\ & + \beta_6 s_j + \beta_7 t_i s_j + \beta_8 t_i^2 s_j + \beta_9 x_k t_i s_j + \beta_{10} x_k t_i^2 s_j \end{aligned}$$

where $\beta_2 t_i + \beta_3 t_i^2$ provides the trend for Group 1 on Day 1, $(\beta_2 + \beta_4)t_i + (\beta_3 + \beta_5)t_i^2$ the trend for Group 2 on Day 1, and $(\beta_2 + \beta_7 s_j)t_i + (\beta_3 + \beta_8 s_j)t_i^2$ the trends for Group 1 on days $s_j > 1$ and $(\beta_2 + (\beta_7 + \beta_9)s_j)t_i + (\beta_3 + (\beta_8 + \beta_{10})s_j)t_i^2$ the trends for Group 2 on days $s_j > 1$.

In the random part of the model the inclusion of the level 2 random slope t_i, in addition to the level 3 random slope, allows for variation in fixed trends within-subject.

Currently the variance function estimates the momentary variation in recovery for the two groups, allowing the variation to depend on time. A quadratic dependency is observed in both groups. To allow this dependency to be different each day, s_j can also be modelled in the variance function by specifying the random part of the model as

$$u_{0k} + u_{1k}t_i + v_{0jk} + v_{1jk}t_i + e_{0ijk}x_k(0) + e_1 x_k(1) + e_2 t_i + e_3 s_j$$

The level 1 variance function is then

$$\begin{aligned} var\left(e_{0ijk}x_k(0) + e_1 x_k(1) + e_2 t_i + e_3 s_j\right) \\ &= \sigma_{e0}^2 + \sigma_{e1}^2 + 2\sigma_{e02}x_k(0)t_i + 2\sigma_{e12}x_k(1)t_i + 2\sigma_{e03}x_k(0)s_j \\ &+ 2\sigma_{e13}x_k(1)s_j + \sigma_{e2}^2 t_i^2 + 2\sigma_{e23}t_i s_j + \sigma_{e3}^2 s_j^2 \end{aligned}$$

which can be plotted against beep number for each day to compare within-day fluctuation in recovery between the two groups, across days. These functions are presented in Figure 5:14 where each graph depicts the Group 1 and Group 2 variance functions for each day.

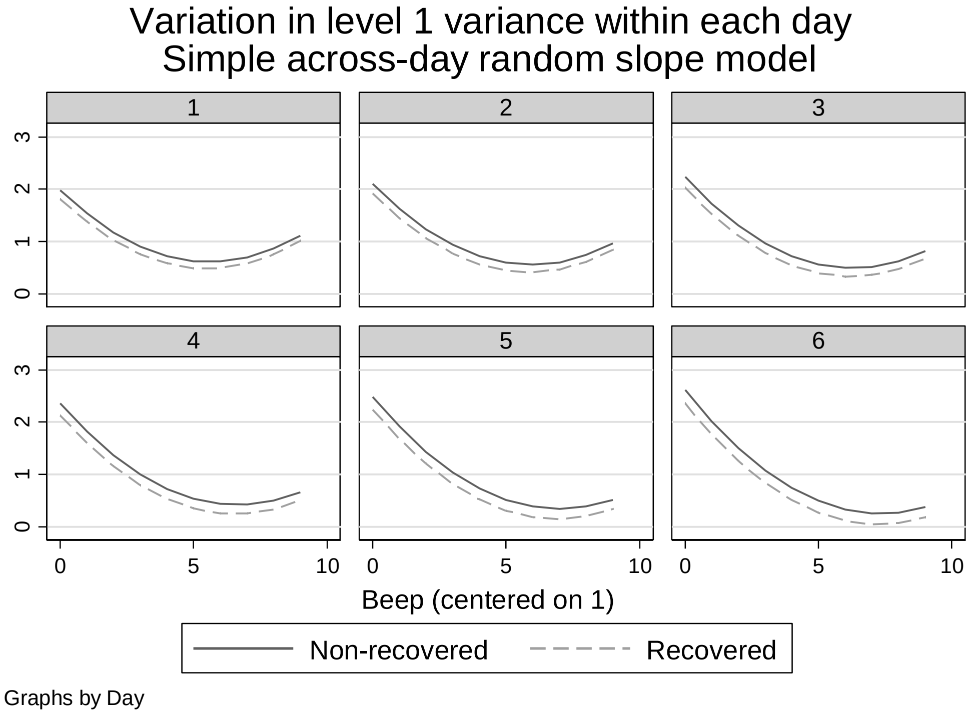


Figure 5:14 Time dependent heteroskedasticity between two groups: across-day variation in within-day trend

5.4 Summary

The three-level structure of ESM data allows for variation in measures both within-day and across the whole sampling period. Currently used methods for examining moment to moment variation were found to underutilize this rich data, condensing all the variability into summary measures.

As an alternative to summary measures, a three-level paradigm for growth models was introduced which allows variation to be studied in terms of time trends, where time can be measured at both level 1 (beep number) and level 2 (day number). Random slope models including all permutations of these time variables as fixed and random effects were presented, with an explanation provided for how these can be used to study daily and weekly trends. Combined with a beep by day interaction term, it was shown how to random slopes can also be specified to model daily trends that vary across days.   Further extensions to these base models were then explored, including modelling nonlinear time trends and how complex level 1 variance models can be used to investigate group differences in momentary fluctuation in terms of time dependent heteroskedasticity.

In summary, this chapter has demonstrated how variation in ESM data can be explored, and how complex random slope models can be used to gain a detailed understanding of the momentary variation ESM is designed to capture.

6 Predicting momentary change

Investigating predictors of momentary change was one of the main research questions of the Bentall et al study. As presented in Chapter 2.2, using a simple change score in an ESM context is not appropriate. This chapter will present how change models are currently applied in the ESM literature, the drawbacks these particular models face when used with multilevel data and potential methodological solutions. The suggested alternatives will then be applied to the Recovery data set.

6.1 Introduction

The term 'change models' in the context of ESM research will refer to models used to investigate predictors of a change in outcome from one moment to the next. This can be defined in two ways:

1. using a change score as the dependent variable, calculated as the difference in consecutive moments i and $i-1$, $y_{ijk} - y_{i-1,jk}$
2. using the dependent variable y_{ijk} while fitting the lagged outcome $y_{i-1,jk}$ as a covariate.

These are analogous to the change score and ANCOVA approaches in single-level models, but both methods present challenges when applied to ESM data due to the nesting of moments within days. For clarity, definition 1 will be referred to as a change score model, while definition 2 will be referred to as a lagged outcome model.

Inference on change will also depend on the how covariates are entered into the model. To investigate predictors of change, covariates x should be time lagged such that, in the context of definition 2 above,

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + \beta_2 x_{i-1,jk} + u_k + v_{jk} + e_{ijk}. \tag{5}$$

Here both covariates x and y are included in the model at moment $i-1$, such that the coefficient β_2 is interpreted as the effect of a unit increase in x at moment $i-1$ on a change in y from moment $i-1$ to i. The inference of this model is subtly different from the model where x is entered at moment i

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + \beta_2 x_{ijk} + u_k + v_{jk} + e_{ijk} \tag{6}$$

where x and the dependent variable y are measured at the same time point and the lagged outcome, y at moment $i - 1$, is included as a covariate. In this model β_2 is the concurrent effect of a unit increase in x on y, controlling for the previous value of y.

These two models are referred to by many different terms in the literature, for example autoregressive models, dynamic panel models (predominantly in economics literature) and lagged-response models (Rabe-Hesketh and Skrondal 2012), but in this chapter they will be defined as lagged outcome models, with only model (5) referred to as a change model.

The remainder of this chapter will be sectioned into three parts. Firstly, ESM papers discussing change models identified in the systematic review of Chapter 3 will be presented to determine how these models are being defined and used in practice. Following this, the issues involved in using change models with intensive longitudinal data will be discussed with a motivating example from the recovery data set. Potential solutions to these issues will then be investigated and applied to this data.

6.1.1 CHANGE MODELS IDENTIFIED IN THE SYSTEMATIC REVIEW

Three papers from the 2012 review investigated associations with change in outcome at the momentary level, that is, what predicts a change in outcome from moment to moment. A further three papers included the lagged outcome as a predictor. These papers were identified either from hypotheses presented on the subject of change or where statistical models were presented defining the outcome at the previous time point was "controlled for".

Mata et al (2012) conducted an ESM study to investigate the relationship between self-initiated physical activity and affect in subjects diagnosed with major depressive disorder (MDD) and healthy controls. A total of 106 subjects (53 with MDD and 53 controls) were monitored for "seven to eight days". Participants were given a palmpilot which emitted eight semi-random beeps within 90 minute blocks during waking hours (10am – 10pm). At each beep participants were required to complete items relating to their current positive and negative affect and questions regarding any physical activity they had undertaken since the previous beep. The authors hypothesised that positive affect would increase and negative affect decrease following physical activity. Two-level random intercept models were used to test for this association, with physical activity at moment $i + 1$ predicting a change in affect (from moment i to $i + 1$)

$$Affect_{i+1,k} - Affect_{ik} = \beta_0 + \beta_1 Group_k + \beta_2 Activity_{i+1,k} + \beta_3 G * A + \beta_4 Time_{ik}$$

$$+ \beta_5 T * G + \beta_6 Time_{ik}^2 + \beta_7 T^2 * G + u_k + e_{ik.}$$

where $Group_k$ is an indicator variable for control or MDD participants and $Time_{ik}$ represents time of day, measured as the number of minutes since the first moment of the day, centred across participants. Physical activity was entered into the model at moment $i+1$ rather than i as it was assumed that activity occurred in the interval between prompts and was recorded at the next, thus $Activity_{i+1,k}$ represents activity during the interval i to $i+1$. The change score was restricted to occur within the same day, that is measures at $i+1$ were excluded if they occurred on the next day to measurements at moment i.

The authors acknowledge the day level within their data but argue that as physical activity was not reported on a substantial number of days (44%) these days would be excluded from analysis, and as such chose a two level model with moments nested within participants. It was unclear from the paper whether "days without reported physical activity" referred to days where the lack of physical activity was recorded, i.e. activity coded as zero and as such there would be no variation in this measure, or whether instead reports of physical activity were missing. It is presumed the latter, as the software used for the analysis (HLM 6.08; (Raudenbush, Bryk et al. 2004)) uses listwise deletion of missing data.

The results showed that there was no significant change in affect when there was no reported physical activity, and although positive affect increased with physical activity, there was no significant difference in this change between the depressed and control groups.

Kuppens, Champagne and Tuerlinckx (2012) used ESM to assess the relationship between appraisals of events and core affect. A total of 79 students were monitored for two weeks, completing 10 ESM questionnaires each day on a Palmpilot using a stratified random sampling procedure. Core affect was measured in terms of valence (attractiveness of an event) on a scale of displeasure to pleasure, and arousal, recorded on a scale of active to passive, and was scored on a 99x99 square grid – a modified version of the 9x9 Affect Grid (Russell, Weiss et al. 1989). Appraisal items related to the question "What is causing your feelings right now?" and were scored on a continuous slider scale of 0-100. The authors state that separate two-level "autocorrelation-crosscorrelation regression models" were used to evaluate the relationship between each appraisal item and the change in valence and arousal. This term appears to have been coined by Kuppens who explains more fully in his 2011 paper (Jackson, Kuppens et al. 2011) the ‘autocorrelation’ relating to the use of the lagged outcome as a covariate and the ‘cross correlation’ relating to the use of a lagged predictor. The change in core affect was modelled by fitting the lagged outcome model, for example,

$$Valence_{ik} = \beta_{0k} + \beta_{1k} Valence_{i-1,k} + \beta_{2k} Appraisal_{i-1,k} + e_{ik} \tag{7}$$

where β_{2k} represents the effect of appraisal at time $t-1$ on the change in valence from $t-1$ to t. The authors argue that this “is the closest one can come to examining directional and causal relationships on the basis of time series data” referencing Granger (1969) and Gottman (1990). This model was presented in the paper with the additional information that both “intercept and slope coefficients were allowed to vary across persons”. Although no further details were provided as to which covariates were given a random slope, it is assumed that only a random slope for appraisal was included at level 2, allowing for separate effects of appraisal on change in valence for each subject. The authors also state the lagged variables were restricted to be within the same day but did not allow for separate day random intercepts. This suggests that the following model may have been used

$$y_{ik} = \beta_0 + \beta_1 y_{i-1,k} + \beta_2 x_{i-1,k} + u_{0k} + u_{1k} x_{i-1,k} + e_{ik}.$$

However, the notation of their sample model (equation (7) above) with the double subscript on the β coefficients is indicative of alternative multilevel notation where models for different levels are defined separately, where it can be assumed model (7) is the level 1 equation of

$$Level\ 1\colon y_{ik} = \pi_{0k} + \pi_{1k} y_{i-1,k} + \pi_{2k} x_{i-1,k} + e_{ik}$$

where

$$\begin{aligned} Level\ 2\colon \pi_{0k} &= \beta_{00} + u_{0k} \\ \pi_{1k} &= \beta_{10} + u_{1k} \\ \pi_{2k} &= \beta_{20} + u_{2k} \end{aligned}$$

If this is the case, both covariates may have random slopes as becomes clearer in the full model

$$y_{ik} = \beta_0 + \beta_1 y_{i-1,k} + \beta_2 x_{i-1,k} + u_{0k} + u_{1k} y_{i-1,k} + u_{2k} x_{i-1,k} + e_{ik}.$$

The fixed effects with accompanying standard errors were reported for each appraisal item along with estimates of the standard deviation of the random slope for each item. No coefficient of the lagged outcome was presented and so it remains unclear whether this variable was allowed to vary at level 2.

Finally Koval and Kuppens (2012) presented a lagged outcome model to study changes in emotional intertia when anticipating a stressful event. A sample of 71 university students and recent graduates were monitored using palmtop computers for two days, each day completing 60 questionnaires with items relating to their current feelings. The authors define ‘emotional inertia’ as the “autocorrelation of an emotion” here specifically studying the ‘threat emotion’ comprised as the average of two ESM items measuring current levels of anxiety and stress by way of a 1 – 100 VAS slider. Rather than the typical statistical definition of autocorrelation estimated in the residuals, the authors refer to autocorrelation in emotional inertia as how feelings “carry over from one moment to the next”, what might better be described as ‘change’. The first day of the study period was used as a baseline or control measure of the subjects’ emotional experience of normal daily life. On the second day they were informed they would completing the Trier Social Stress Test (TSST) designed to induce social stress later that day. The second day of sampling therefore measured the variation in anxiety and stress in anticipation of this task. The authors wished to compare the change in threat emotion during corresponding time periods on day 1 as day 2 (after being briefed and up to the TSST) and whether this effect was moderated by individual differences in depression, self-esteem and fear of negative emotion. The specific models were described in the paper using alternative notation of separately specifying the level 1 and level 2 models:

$$Level\ 1\text{:}\ Threat_{ik} = \pi_{0k} + \pi_{1k} SA_{ik} + \pi_{2k} Threat_{i-1,k} + \pi_{3k} SA_{ik} * Threat_{i-1,k} + e_{ik}$$

$$Level\ 2\text{:}\ \pi_{pk} = \beta_{p0} + \beta_{p1} RSE_k + u_{pk} \ \text{ for } p = 0,1,2,3.$$

where RSE_k is the subject level Rosenberg Self-Esteem scale, SA_{ik} is an indicator variable for whether the measurement is on the control or Stressor Anticipated time block and Threat is the anxiety/stress composite scale measured at the momentary level. The random coefficient π_{2k} is described as representing the “autocorrelation (inertia) of person k’s threat emotion during the Baseline block” and π_{3k} “represents the change in that autocorrelation slope from Baseline to Stressor Anticipation". Presented in the as one model this expands to

$$\begin{aligned}Threat_{ik} = {} & \beta_{00} + \beta_{01}RSE_k + \beta_{10}SA_{ik} + \beta_{11}SA_{ik} * RSE_k + \beta_{20}Threat_{i-1,k} \\ & + \beta_{21}Threat_{i-1,k} * RSE_k + \ \beta_{30}SA_{ik} * Threat_{i-1,k} \\ & + \beta_{31}SA_{ik} * Threat_{i-1,k} * RSE_k \\ & + u_{ok} + u_{1k}SA_{ik} + u_{2k}Threat_{i-1,k} + u_{3k}SA_{ik} * Threat_{i-1,k} + e_{ik}\end{aligned}$$

where the interpretation of individual coefficients is not as intuitive. A table of results were presented for the mean scores (β_{00}, β_{01}, β_{10} and β_{11}) and what the authors refer to as the autocorrelation parameters ($\beta_{20}, \beta_{21}, \beta_{30}$ and β_{31}) corresponding to the coefficients of the lagged outcome covariate and any of its interactions. The effect of anticipatory stress on emotional inertia was found to be moderated by self-esteem: subjects with low self-esteem exhibiting a decrease in threat emotion when anticipating a stressor.

Three other studies also included the lagged outcome as a covariate in their multilevel model, though their hypotheses were not related to change. In addition to their models investigating instability of deservedness of paranoid thoughts and self-esteem, Udachina et al (2012) also examined whether paranoia was a predictor of subsequent self-esteem, while controlling for current self-esteem "as a possible confounder", and whether this effect differed between baseline measured paranoia groups and control subjects. Ben-Zeev, Young et al. (2012) also reported controlling for the outcome at the previous time point in their study examining lagged predictors of suicidal ideation in 31 in-patients with Major Depressive Disorder, stating that

"Predicting suicidal ideation at time t from affect and symptom ratings at time t−1 controlling for t−1 suicidal ideation allows more direct causal inferences, although the design is still correlational".

And finally, To, Fisher et al. (2012) controlled for the lagged outcome in both concurrent and lagged analyses as a solution for the serial autocorrelation present in ESM data. They referenced Ilies and Judge (2002) who in turn cited Bryk and Raudenbush (1992) for this parameterisation of the residual correlation.

In summary, although each of these papers used either change scores or lagged outcome models in some way, none commented on the possible flaws of these methods in multilevel data. The following sections of this chapter will outline potential problems with fitting change models in ESM data and present solutions with application to the recovery data set.

6.2 ISSUES SURROUNDING CHANGE MODELS: MOTIVATING EXAMPLE

As presented in Chapter 1.3, change models were a question of primary interest in the recovery study. However, the results of these models were not as expected. To answer the question of interest "what predicts a change in recovery?", the change in recovery from moment $i-1$ to i, calculated as a change score $recovery_i - recovery_{i-1}$, was regressed on level 1 diary constructs such as self-esteem and hopelessness at moment $i-1$:

$$recovery_{ijk} - recovery_{i-1,jk} = \beta_0 + \ \beta_1 selfesteem_{i-1,jk} + u_k + v_{jk} + e_{ijk}.$$

This resulted in the estimates of Table 2:5, presented graphically below in Figure 6:1.

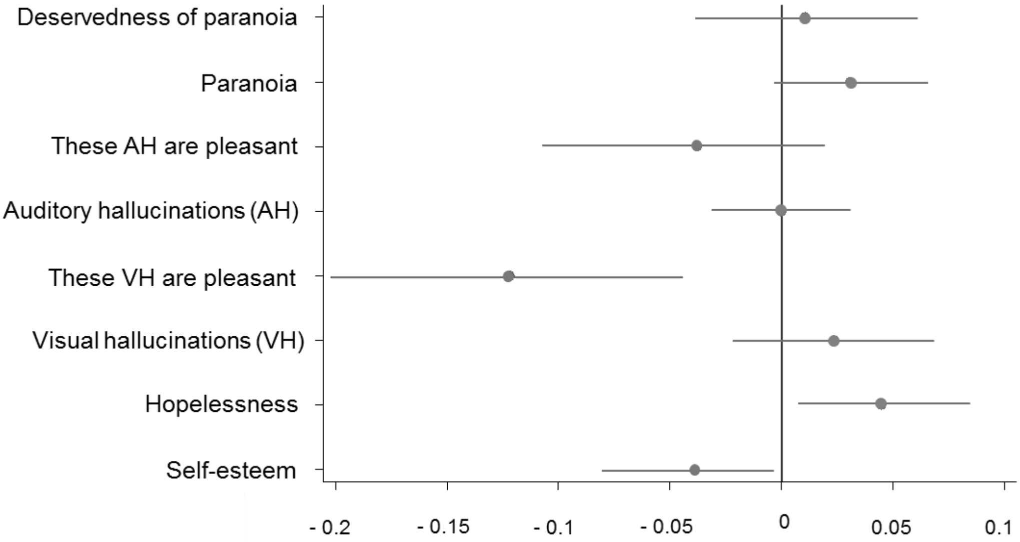


Figure 6:1 Parameter estimates of naïve change score models in the recovery data

The effect estimates from these models are quite unusual; it is suggested that an increase in self-esteem, for example, precedes a reduction in self-reported recovery, while an increase in recovery follows a greater feeling of hopelessness. Intuitively one would expect the opposite sign on the coefficient for each variable in the table. Moreover, subject-level and day-level variation in random intercepts was negligible, implying that all subjects have the same change score for mean values of x.

To understand this phenomenon in the fixed effects it is useful to dissect the relationship between X in Y. If examining X at timepoint $i-1$ and the change in Y from $i-1$ to i, i.e $corr(X_{i-1}, Y_i - Y_{i-1})$, we have that

$$corr(X_{i-1}, Y_i - Y_{i-1}) = \frac{cov(X_{i-1}, Y_i - Y_{i-1})}{\sigma_{X_{i-1}}\sigma_{\Delta Y}}$$

where $\Delta Y = (Y_i - Y_{i-1})$. From this, using the properties of covariance, it follows that

$$\begin{aligned}\frac{cov(X_{i-1}, Y_i - Y_{i-1})}{\sigma_{X_{i-1}}\sigma_{\Delta Y}} &= \frac{cov(X_{i-1}, Y_i)}{\sigma_{X_{i-1}}\sigma_{\Delta Y}} - \frac{cov(X_{i-1}, Y_{i-1})}{\sigma_{X_{i-1}}\sigma_{\Delta Y}} \\ &= \frac{corr(X_{i-1}, Y_i)\sigma_{Y_i}}{\sigma_{\Delta Y}} - \frac{corr(X_{i-1}, Y_{i-1})\sigma_{Y_{i-1}}}{\sigma_{\Delta Y}}\end{aligned}$$

which, if we assume the standard deviation of Y_i is approximately equal to the standard deviation of Y_{i-1}, i.e. $\sigma^2_{Y_i} \approx \sigma^2_{Y_{i-1}}$, is

$$= \frac{\left(corr(X_{i-1}, Y_i) - corr(X_{i-1}, Y_{i-1})\right)\sigma_Y}{\sigma_{\Delta Y}}.$$

Thus if the lagged relationship is weaker than the concurrent relationship, that is if $corr(X_{i-1}, Y_i) < corr(X_{i-1}, Y_{i-1})$, then the relationship between the change in Y and the lagged X will be negative $(i.e. corr(X_{i-1}, Y_i - Y_{i-1}) < 0)$. As the relationship between X and Y typically diminishes with each lag, this inequality would frequently hold and would explain any unexpected direction in coefficient estimates. To illustrate this, two examples of these lagged correlation pairs are presented in Table 6:1 for recovery (Y) and covariates self-esteem and hopelessness using the recovery data. For both these variables $corr(X_{i-1}, Y_i) < corr(X_{i-1}, Y_{i-1})$, thus explaining the irregular signs in Table 2:5.

Correlation pair	Self-esteem	Hopelessness
$corr(X_{i-1}, Y_{i-1})$	0.573	-0.588
$corr(X_{i-1}, Y_i)$	0.546	-0.543
$corr(X_{i-2}, Y_{i-1})$	0.545	-0.542
$corr(X_{i-2}, Y_i)$	0.533	-0.533

Table 6:1 Lagged correlation pairs for recovery data

6.2.1 Endogeneity

An alternative approach to using a change score as the outcome is to instead fit a model with the lagged outcome as a covariate. This, however, presents its own set of problems which is referred to as endogeneity.

To understand the potential problems with applying lagged outcome models to ESM data it is first useful to recall the assumptions underpinning multilevel models. As outlined in Section 2.1, it is assumed that the level 1 residuals of a three-level random intercept model are normally distributed with mean zero given the random intercepts and the set of covariates $\boldsymbol{x}$,

$$E\left(e_{ijk}\middle|\boldsymbol{x}, u_k, v_{jk}\right) = 0$$

and thus the covariates and residuals are uncorrelated, or 'exogenous'. Similarly at levels 2 and 3

$$E\left(v_{jk}\middle|\boldsymbol{x}, u_k, e_{ijk}\right) = 0$$

$$E\left(u_k\middle|\boldsymbol{x}, v_{jk}, e_{ijk}\right) = 0$$

implying the subject-level and day-level random intercepts are also independent of the covariates. If any of these assumptions are violated, wherein the covariates are correlated with the residuals or random effects, the covariates are considered 'endogenous'. Fitting a model with endogeneity at levels 1, 2 or 3 would lead to biased estimates of both the fixed effects and the random effect variances (Ebbes, Böckenholt et al. 2004; Kazemi and Crouchley 2006; Rabe-Hesketh and Skrondal 2012).

When studying momentary level change using a lagged outcome model, the exogeneity assumption is violated as $y_{i-1,jk}$ is correlated with the random effects. This can be seen by taking the first lag of equation (5)

$$y_{i-1,jk} = \beta_0 + \beta_1 y_{i-2,jk} + \beta_2 x_{i-2,jk} + u_k + v_{jk} + e_{i-1,jk}$$

which shows that the lagged outcome is dependent on both u_k and v_{jk}.

Two methods for overcoming this problem when investigating change are presented below.

6.2.2 Solution to endogeneity: The first-difference method

Rabe-Hesketh and Skrondal (2012) advocate a first-difference approach to circumvent the endogeneity problem. The method is commonly used in economic literature and is discussed in detail by Anderson and Hsiao (1981); (1982) as a two-level problem, with a focus on the coefficient of the lag outcome, rather than covariates x.

Presented in the context of ESM, the method is as follows. Beginning with a two-level lagged outcome model with both moment level x_{ik} and subject level x_k predictors

$$y_{ik} = \beta_0 + \beta_1 y_{i-1,k} + \beta_2 x_{1ik} + \beta_3 x_{2k} + u_k + e_{ik} \tag{8}$$

the first lag of equation (8) is taken

$$y_{i-1,k} = \beta_0 + \beta_1 y_{i-2,k} + \beta_2 x_{1i-1,k} + \beta_3 x_{2k} + u_k + e_{i-1,k} \tag{9}$$

followed by the difference of the two models (8) and (9)

$$y_{ik} - y_{i-1,k} = \beta_1\left(y_{i-1,k} - y_{i-2,k}\right) + \beta_2\left(x_{1ik} - x_{1i-1,k}\right) + e_{ik} - e_{i-1,k} \tag{10}$$

which results in the first-difference equation.

This first-difference approach eliminates the random intercept u_k solving the original endogeneity problem. However, the process creates the lagged first-difference $(y_{i-1,k} - y_{i-2,k})$ on the right hand side of equation (10) which is correlated with the new residuals $(e_{ik} - e_{i-1,k})$, which can be seen by taking the first lag of (10):

$$y_{i-1,k} - y_{i-2,k} = \beta_1\left(y_{i-2,k} - y_{i-3,k}\right) + \beta_2\left(x_{1i-1,k} - x_{1i-2,k}\right) + e_{i-1,k} - e_{i-2,k}.$$

The authors refer to Anderson and Hsiao (1981, 1982) to solve this problem, who suggest using either the second lag of the outcome $(y_{i-2,k})$ or the second lag of the first-difference $(y_{i-2,k} - y_{i-3,k})$ as an instrumental variable for the lagged first-difference. Either of these approaches would be appropriate as they fulfil the instrumental variable requirements of being correlated with $\left(y_{i-1,k} - y_{i-2,k}\right)$ while being uncorrelated with $(e_{ik} - e_{i-1,k})$. They state that this method will provide consistent estimates for the coefficients of time varying variables (x_{1ik} in equation (10)) and the lagged outcome, however not for the coefficients of time invariant variables, i.e. subject level variables (x_{2k}), or the random intercept variances. Hsiao (2003) (Section 4.3.3.c) is referenced for instruction on how to obtain consistent estimates for these parameters.

Rabe-Hesketh and Skronal outline the first difference procedure in Stata and recommend using the second lag of the predictor $(y_{i-2,k})$ as the instrumental variable using the

IVREGRESS command. This command (pg 931 in Stata manual (StataCorp 2013)) fits the model

$$\mathbf{y}_{\boldsymbol{i}} = \boldsymbol{x}_{\mathbf{1}\boldsymbol{i}}\boldsymbol{\gamma}_{\mathbf{1}} + \boldsymbol{x}_{\mathbf{2}\boldsymbol{i}}\boldsymbol{\gamma}_{\mathbf{2}} + \epsilon_{\mathrm{i}} \tag{11}$$

$$y_i = \mathbf{y}_{\boldsymbol{i}}\boldsymbol{\beta}_{\mathbf{1}} + \boldsymbol{x}_{\mathbf{1}\boldsymbol{i}}\boldsymbol{\beta}_{\mathbf{2}} + e_i \tag{12}$$

where y_i is the outcome, $\mathbf{y}_i$ the vector of endogenous predictors, $\boldsymbol{x}_{\mathbf{1}\boldsymbol{i}}$ the vector of exogenous predictors and $\boldsymbol{x}_{\mathbf{2}\boldsymbol{i}}$ the vector of instrumental variables, and e_i and ϵ_i represent the error terms. This command allows the authors to use the second lagged outcome as an instrumental variable in their first-difference method, the resulting model specified as

$$y_{i-1,k} - y_{i-2,k} = \gamma_1\left(x_{1ik} - x_{1i-1,k}\right) + \gamma_2 y_{i-2,k} + \epsilon_{ik} - \epsilon_{i-1,k} \tag{13}$$

$$y_{ik} - y_{i-1,k} = \beta_1(\widehat{y_{i-1,k} - y_{i-2,k}}) + \beta_2\left(x_{1ik} - x_{1i-1,k}\right) + e_{ik} - e_{i-1,k} \tag{14}$$

The command runs a two-stage least squares procedure where the predicted values of $(y_{i-1,k} - y_{i-2,k})$ in equation (13) are substituted in for the observed values of $(y_{i-1,k} - y_{i-2,k})$ in equation (14). The coefficient of interest is then β_2 of model (14) which can be interpreted as the concurrent effect of a change in x on a change in y.

The requirements of Stata's available commands for instrumental variable use in simultaneous equations, however, does not accommodate three-level data. Section 6.3.1 of this chapter will introduce how a two-step estimating procedure can be manually reproduced for three-level data, with an example using the recovery study data.

6.2.3 SOLUTION TO ENDOGENEITY: THE INITIAL CONDITIONS PROBLEM

A further problem of using the lagged outcome as a predictor to study change is referred to as the 'initial conditions problem' (Rabe-Hesketh and Skrondal 2012; Steele, Rasbash et al. 2013) which arises when the measurement period does not coincide with the true start of the observed process. Steele, Rasbash et al. (2013) posit that the process between which two consecutive observations are related reduces to the association between the current (y_{ijk}) and initial (y_{11k}) observations. This can be shown through substitution for the lagged outcome model. Starting with the first available moment $i = 2$ (as there is no lag for $i = 1$)

$$y_{21k} = \beta_0 + \beta_1 y_{11k} + u_k + v_{1k} + e_{21k},$$

the following moment $i = 3$ can be written, substituting in for y_{21k} above,

$$
\begin{aligned}
y_{31k} &= \beta_0 + \beta_1 y_{21k} + u_k + v_{1k} + e_{31k} \\
&= \beta_0 + \beta_1(\beta_0 + \beta_1 y_{11k} + u_k + v_{1k} + e_{21k}) + u_k + v_{1k} + e_{31k} \\
&= \beta_0 + \beta_0\beta_1 + \beta_1^2 y_{11k} + u_k(1+\beta_1) + v_{1k}(1+\beta_1) + \beta_1 e_{21k} + e_{31k}\,.
\end{aligned}
$$

Similarly for moment $i = 4$

$$
\begin{aligned}
y_{41k} &= \beta_0 + \beta_1 y_{31k} + u_k + v_{1k} + e_{41k} \\
&= \beta_0 + \beta_1(\beta_0 + \beta_1 y_{21k} + u_k + v_{1k} + e_{31k}) + u_k + v_{1k} + e_{41k} \\
&= \beta_0 + \beta_1(\beta_0 + \beta_0\beta_1 + \beta_1^2 y_{11k} + u_k(1+\beta_1) + v_{1k}(1+\beta_1) + \beta_1 e_{21k} + e_{31k}) \\
&\qquad + u_k + v_{1k} + e_{41k} \\
&= \beta_0 + \beta_0\beta_1 + \beta_0\beta_1^2 + \beta_1^3 y_{11k} + u_k(1+\beta_1+\beta_1^2) + v_{1k}(1+\beta_1+\beta_1^2) + \beta_1^2 e_{21k} \\
&\qquad + \beta_1 e_{31k} + e_{41k}
\end{aligned}
$$

and so on within each day j. Thus for any $i = 2, \dots, n_1$ the outcome y_{ijk} is dependent on the initial value of y with coefficient β_1^{i-1}.

The initial conditions problem arises for ESM data as the initial observation y_{11k} will be correlated with u_k and v_{jk}. However, this initial observation is never used as a response in a lagged outcome model as it has no lag; y_{01k} is not observed in the ESM study. In only including y_{11k} as a covariate, it is not influenced by any unmeasured subject-level or day-level heterogeneity which is assumed to effect all other outcomes y_{ijk} for $i > 1$. This assumption is unlikely to hold unless y_{11k} is considered the true start of the process of y, which is unrealistic in ESM data where the data are presumed to be representative of just a snapshot of human experience, and the first measurement represents the start of the observation rather than the start of the process.

The naïve modelling approach of assuming y_{11k} is exogenous will result in a misspecified model. As a consequence, the variation which would be explained by the unobserved heterogeneity for the initial observation will be attributed to the change in y, resulting in positively biased estimates of β_1 and negatively biased estimates of the random effects σ_u^2 (Kazemi and Crouchley, 2006). Furthermore, Kazemi and Crouchley state that the naïve model may produce negative variance estimates of the random effects and demonstrate via probability limits that this approach will result in inconsistent parameter estimates when the number of subjects is large and the number of repeated observations is small.

Steele et al (2013) propose a solution for the initial conditions problem for two-level models where observations i are nested within clusters k, in which a separate model is specified for the initial condition to allow y_{1k} to depend on u_k. This is then jointly estimated with the model for the remaining observations. For simple change models in ESM data, the parametrization from Crouchley, Stott et al. (2009) from a "one-factor decomposition" model specifies a shared random effect u_j in both equations, the variance of which can be separately estimated for the initial model

$$y_{1k} = \alpha_0 + \lambda_u u_k + e_{1k} \tag{15}$$

and subsequent model

$$y_{ik} = \beta_0 + \beta_1 y_{i-1,k} + u_k + e_{ik}, \text{for } i > 1 \tag{16}$$

where λ_u is described as a "random effect loading" (Steele, Rasbash et al. 2013; Steele 2014) and can either be estimated, in effect weighting the random effects of the initial model, or constrained to $\lambda_u = 1$. The e_{1k} are the residuals for the initial observation for subjects $k = 1, \dots, n_2$.

The authors argue that if this initial model is correctly specified, bias in the estimated coefficients of the main model will be avoided. This is corroborated by Kazemi and Crouchley (2006) who's "pragmatic approach" also specifies joint modelling the initial condition with the remaining observations, estimating separate but correlated random effects, and who further contend that this approach eliminates the problem of negative variance estimates present in the naïve approach.

In a three-level ESM scenario, 'change' may only be relevant when restricted to be within a day: a change in y overnight may be considered too long a gap, with no interpretable consequence of x on a change in y when x is measured the night before. In this case, when fitting a lagged outcome model the first measurement each day is only ever included as a predictor, rather than just the first measurement overall in a two-level scenario. The same issues apply in this context as for a two-level model - when the y_{1jk} for each day j are not allowed to depend on unobserved heterogeneity at levels 2 and 3, β_1 will be overestimated and the random effect variances underestimated. To combat this bias, the joint modelling of the initial conditions approach can be extended to three levels and rather than just modelling the very first observation for each subject as the response, the initial model will now contain the first observation for each day, for each subject. The following proposes two ways this might be achieved.

The most straightforward approach is to simply expand equation (15) to two levels and jointly model

$$y_{1jk} = \alpha_0 + \lambda_u u_k + \lambda_v v_{jk} + e_{1jk} \tag{17}$$

for observation 1 on each of days j and the remaining observations

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + u_k + v_{jk} + e_{ijk} \tag{18}$$

for $i > 1$ in which u_k and v_{jk} are shared random effects, while allowing for independent residual variation in each model. As with the two-level paradigm, the option to weight the random effects in the initial model exists, this time with the additional loading λ_v for the day-level random effects.

If assuming equal influence of unmeasured heterogeneity on both equations, i.e. $\lambda_u = \lambda_v = 1$, this approach is acceptable; the separate residual variance estimated for each equation reflects the difference in predicted values for the initial and subsequent models, whilst the constrained loadings assume the initial observations each day vary between-subject and between-days in the same manner as the remaining observations. However, to freely estimate either of these loadings is to assume that subject-level or day-level unmeasured heterogeneity differs for these first observations of the day. In Steele et al (2013)'s two-level model this can be explained as unmeasured heterogeneity prior to the start of the study influencing the initial measurement. This too can be argued for the initial observation in the proposed three-level setting, but not the first measurements of each of the remaining days: these values surely instead are affected by unmeasured heterogeneity in the same way as the surrounding observations.

When expecting a differential effect of unmeasured heterogeneity on the initial observation in this way, an alternative approach to the model of (17) and (18) would be to specify an initial model for observation 1 of day 1 only

$$y_{11k} = \alpha_{00} + \lambda_u u_k + \lambda_v v_{1k} + e_{11k} \tag{19}$$

to be jointly estimated with both a model for the first observation for each remaining day $j = 2, \dots, n_2$

$$y_{1jk} = \alpha_{01} + u_k + v_{jk} + e_{1jk} \tag{20}$$

and a model for the subsequent observations $i = 2, \dots, n_1$ on all days

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + u_k + v_{jk} + e_{ijk}. \tag{21}$$

This approach allows for unmeasured heterogeneity prior to the start of the study to influence the initial observation via equation (19), whilst also ensuring the first observation for the subsequent days is also effected by unmeasured heterogeneity via equation (20) but maintains that this effect is equal to that placed on the remaining observations modelled in equation (21). Joint modelling all three equations should remove the bias in β_1 and the variance of the random effects encountered in the naïve model.

6.2.3.1 Joint estimation methods

Two methods will be presented to jointly estimate the multiple multilevel equations defined above. The first approach specifies both the initial model for y_{11k} and the full model for y_{ijk} for $i > 1$ as one model, using dummy variables to differentiate between the two outcomes. The joint model under specification 1 (two equations) can then be defined as

$$y_{ijk} = \alpha_0 T1 + \beta_0 TG1 + \beta_1 TG1 * y_{i-1,jk} + u_k + v_{jk} + e_{11k} T1 + e_{ijk} TG1 \tag{22}$$

where dummy variables

$$T1 = \begin{cases} 1 \text{ for } i = 1 \\ 0 \text{ for } i > 1 \end{cases} \text{ and } TG1 = \begin{cases} 0 \text{ for } i = 1 \\ 1 \text{ for } i > 1 \end{cases}$$

indicate responses for the initial model (T1) and for the remaining observations for time points greater than 1 (TG1). Parameters β_0 and β_1 in this model denote the fixed intercept and slope for equation (18)(13), while α_0 represented the intercept for the initial model (equation (17)). The dummy variables in the residuals allow for separate estimates of the residual variance for each equation. This approach can also accommodate the second joint model specification of three equations with the inclusion of an additional dummy variable for the equation for observation 1 on days $j = 2, \dots, n_2$. This model can be fitted using the MIXED command in Stata which allows for complex level 1 variation, resulting in separate estimates for the residual errors of the initial and full models.

MIXED does not currently have the capability to estimate factor loadings for random effects, so estimating λ_u and λ_v is not possible in this package. As an alternative, the models can be jointly estimated as structural equation models, where the random effects are specified as latent variables (Steele 2014). Multilevel models can be estimated in this manner in Stata using the GSEM (Generalized structural equation model estimation) command. When constraining both $\lambda = 1$, MIXED and GSEM will provide identical results under their default estimation methods as both use maximum likelihood estimation. Estimation using GSEM additionally allows for the estimation random effects loadings for the initial condition model, thus λ_u and λ_v can be estimated. Restrictions to the syntax of the GSEM command, however, impose that while these weights can be estimated, $\hat{\lambda}_u = \hat{\lambda}_v$. To estimate these weights independently, the joint models can be specified in a Bayesian framework and fit using STAN (Carpenter, Gelman et al. 2016).

6.2.3.2 Joint estimation issues – Lagged covariates

The joint model described in equation (22) is a simplified lagged outcome model with no covariates other than the outcome y at time point $i - 1$. Inference from this model provides understanding on the association between consecutive measurements. A more common research question in ESM asks "What predicts change?". To answer this question moment-level, day-level or subject-level explanatory variables can be included into the model. Level 1 predictors, however, prove to be problematic.

As discussed at the beginning of this chapter, the time point at which level 1 predictors are entered into a change model results in subtly different research questions. Including a predictor measured at moment i as in equation (6) results in a joint model

$$\begin{aligned} y_{ijk} &= \alpha_0 T1 + \alpha_1 T1 * x_{ijk} + \beta_0 TG1 + \beta_1 TG1 * x_{ijk} + \beta_2 TG1 * y_{i-1,jk} \\ &\quad + u_k + v_{jk} + e_{11k} T1 + e_{ijk} TG1 \end{aligned}$$

where β_1 can be interpreted as the concurrent effect of x on y, controlling for the value of y at the previous moment. Alternatively, including a lagged predictor

$$\begin{aligned} y_{ijk} &= \alpha_0 T1 + \alpha_1 T1 * x_{i-1,jk} + \beta_0 TG1 + \beta_1 TG1 * x_{i-jk} + \beta_2 TG1 * y_{i-1,jk} \\ &\quad + u_k + v_{jk} + e_{11k} T1 + e_{ijk} TG1 \end{aligned}$$

β_1 can be interpreted as the effect of x on a change in y from moment $i - 1$ to i.

While the concurrent association is straightforward to fit, the formulation of the initial conditions model prevents the inclusion of lagged covariates: for the initial model $i = 1$, so for a lagged covariate $x_{i-1,jk} = x_{0jk}$, for which there is no observed value. With no substitute value the initial model

$$y_{11k} = \alpha_0 + \alpha_1 x_{i-1,jk} + u_k + v_{jk} + e_{11k}$$

becomes just a random intercept model

$$y_{11k} = \alpha_0 + u_k + v_{jk} + e_{11k}.$$

Although the initial value of x is no longer present as a predictor of the initial value of y, the original purpose of specifying this model was to allow y_{11k} to depend on the unmeasured heterogeneity represented by random effects u_k and v_{jk} that would otherwise be ignored. As such, not modelling the initial relationship of y and x does not detract from the objective of this method; if the initial value of y is allowed to be influenced by the unobserved heterogeneity prior to the start of the measurement period the problem of endogeneity should still be resolved. However, the examples in the reviewed literature only included contemporaneous or higher level covariates and as such did not encounter this problem . Though the initial conditions problem thus far has only specified the initial observation's lack of dependence on the random effects, the argument can be extended to the fixed effects: if valid parameter estimates require y_1 to depend on unmeasured heterogeneity, then should it not also depend on measured covariates x? In ignoring the x y relationship in the initial observation of each day the estimation of covariate effects in the main equation are likely to be underestimated, with too much variation in y attributed to unmeasured heterogeneity.

An alternative solution to omitting x in the initial model to ensure y_1 depends on this covariate is to use the concurrent value x_{1jk} as a proxy for the unmeasured lagged value x_{0jk}. While this proposed solution preserves the relationship of interest in the initial moments, it has been observed that concurrent relationship between x and y may be stronger than lagged relationships. As a consequence, the effect of x in the change in y may be overestimated.

The following section will apply the proposed three-level extensions of the first-difference and initial conditions methods to the recovery data set. Here the two-equation and three-equation specifications of the initial conditions models will be investigated and the extent to which misspecification of the initial model due to the unobserved lagged covariate affects model estimates. The results of both methods will be compared to the naïve approach to quantify the inconsistencies in parameter estimates obtained when ignoring the endogeneity problem.

6.3 RECOVERY DATA ANALYSIS

Bentall and colleagues' primary research question concerned whether any moment-to-moment changes in recovery could be predicted by ESM measured variables. These variables included measures of self-esteem, hopelessness, visual and auditory hallucinations, paranoia and a measure of how deserving of these paranoid beliefs the participants felt. For details on how these scales were calculated from the original ESM items please refer to Chapter 1.3.

	Fixed Effects			Random Effects			
Lagged Variable	Coeff.	SE	P value	Level	Variance	SE	N
Self Esteem	0.103	0.026	<0.001	Person	0.687	0.145	1910
Recovery	0.295	0.023	<0.001	Day	0.065	0.018	
				Beep	0.477	0.018	
Hopelessness	-0.081	0.023	<0.001	Person	0.684	0.144	1879
Recovery	0.299	0.023	<0.001	Day	0.061	0.017	
				Beep	0.477	0.018	
Visual Hallucination	-0.020	0.027	0.469	Person	0.7560	0.157	1834
Recovery	0.310	0.022	<0.001	Day	0.051	0.016	
				Beep	0.482	0.019	
Auditory Hallucination	0.048	0.021	0.021	Person	0.785	0.165	1805
Recovery	0.284	0.022	<0.001	Day	0.064	0.017	
				Beep	0.467	0.018	
Paranoia	-0.137	0.026	<0.001	Person	0.785	0.165	1912
Recovery	0.292	0.023	<0.001	Day	0.064	0.017	
				Beep	0.467	0.018	
Deservedness	0.033	0.032	0.303	Person	0.389	0.135	715
Recovery	0.534	0.033	<0.001	Day	0.011	0.029	
				Beep	0.591	0.041	

Table 6:2 Naïve lagged outcome models for recovery data

The original analysis, presented in Chapter 1.3, was repeated with the lagged outcome as a predictor rather than using a change score as the response. Each predictor was entered into separate three-level random intercept models with the lagged outcome as the only other covariate. The results of these analyses are presented in Table 6:2.

In contrast to the models using a change score as the outcome, the results of Table 6:2 show directions of association more as would be expected: positively oriented scales (e.g. self-esteem) having positive associations with recovery, and negatively oriented scales (e.g. hopelessness, hallucinations) being negatively associated with recovery. Comparing the number of observations used in each of the analyses with the number used in the concurrent associations (Table 2:3) it is clear that using lagged predictors significantly reduces the number of observations available for analysis, with most analyses dropping around 550 observations. Note that the deservedness model has far fewer observations than the other measures as this scale was created as a combination of items that branched from a previous item; the items were only to be completed if the participant scored higher than 1 on the paranoia symptom items, thus were coded as missing for occasions where the participant was not currently experiencing paranoia. In addition to the differences in the fixed effects, the results of Table 6:2 also demonstrate marked differences in the variance of the random intercepts compared to the change outcome models. In each of the lagged outcome models there is substantial variation in the subject-specific intercepts, and though comparatively much smaller, still non-zero estimates of between-day variance, confirming what has been shown throughout this thesis.

As has been argued in this chapter, including the lagged outcome as a covariate in these models has violated the exogeneity assumption of random effects models. The first-difference and initial conditions methods will be demonstrated on this data to both identify the extent of the bias produced in the fixed and random effects and attempt to rectify this problem. The results of the two methods will be presented below in Sections 6.3.1 and 6.3.2, followed by a discussion of the benefits and drawbacks of the methodologies.

6.3.1 First-difference approach

Several new variables were required to implement the first difference approach. For the first-difference, lagged variables were created (restricting each lag to be within-day) which were then subtracted from the original variable. Holding the lags within-day ensured that the first-difference of the day was between moments y_{1jk} and y_{2jk} rather than overnight between moments $y_{10,j-1,k}$ and y_{1jk}. The lagged difference then was created by computing the second lag (i.e. y_{i-2}) and subtracting it from the first lag. See Table 6:3 for details.

#	Lag #	Difference	
0	y_{ijk} : lag 0	$y_{ijk} - y_{i-1,jk}$:	First-difference (D)
1	$y_{i-1,jk}$: 1st lag (L)	$y_{i-1,jk} - y_{i-2,jk}$:	Lagged difference (LD)
2	$y_{i-2,jk}$: 2nd lag (L2)	$y_{i-2,jk} - y_{i-3,jk}$:	Second lagged difference (LD2)

Table 6:3 First-difference notation

Working from the lagged outcome model of the original analysis, for example for self-esteem

$$recovery_{ijk} = \beta_0 + \beta_1 self\text{-}esteem_{i-1,jk} + \beta_2 recovery_{i-1,jk} + u_k + v_{jk} + e_{ijk} \tag{23}$$

to implement the first-difference method the Rabe-Hesketh and Skrondal rationale was then followed: the first lag of equation (23) was taken

$$\begin{aligned} recovery_{i-1,jk} = {} & \beta_0 + \beta_1 self\text{-}esteem_{i-2,jk} + \beta_2 recovery_{i-2,jk} \\ & + u_k + v_{jk} + e_{i-1,jk} \end{aligned} \tag{24}$$

followed by taking difference of equations (23) and (24)

$$\begin{aligned} recovery_{ijk} - recovery_{i-1,jk} = {} & \beta_1\left(selfesteem_{i-1,jk} - selfesteem_{i-2,jk}\right) \\ & + \beta_2\left(recovery_{i-1,jk} - recovery_{i-2,jk}\right) + e_{ijk} - e_{i-1,jk.} \end{aligned} \tag{25}$$

Equation (25) represents the first-difference model. As discussed, the lagged difference of recovery in the right hand side of this model $\left(recovery_{i-1,jk} - recovery_{i-2,jk}\right)$ is correlated with the error term $(e_{ijk} - e_{i-1,jk})$ and so an instrumental variable will be used in its place: $recovery_{i-2,jk}$, the second lag of the outcome. We then have the model

$$\begin{aligned} recovery_{i-1,jk} - recovery_{i-2,jk} = {} & \gamma_1\left(selfesteem_{i-1,jk} - selfesteem_{i-2,jk}\right) \\ & + \gamma_2 recovery_{i-2,jk} + \epsilon_{ijk} - \epsilon_{i-1,jk} \end{aligned} \tag{26}$$

$$\begin{aligned} & recovery_{ijk} - recovery_{i-1,jk} \\ & \quad = \beta_1\left(selfesteem_{i-1,jk} - selfesteem_{i-2,jk}\right) \\ & \quad + \beta_2\left(\widehat{recovery_{\imath-1,jk} - recovery_{\imath-2,jk}}\right) + e_{ijk} - e_{i-1,jk} \end{aligned} \tag{27}$$

analogous to the model displayed in equations (11) and (12), where the predicted values of equation (26) are used in place of the observed values of $\left(recovery_{i-1,jk} - recovery_{i-2,jk}\right)$ in equation (27).

When fitting these models using two-step procedure the standard errors for β_1 and β_2 are likely to be inflated, so the process will be bootstrapped and robust standard errors will be presented.

Table 6:4 presents the results of the first-difference models with separate models fit for each of the measures.

Lagged difference **Predictor**	Coefficient	Bootstrapped SE	P value	Bootstrapped 95% CI	N
LD2.Self-Esteem	-0.131	1.203	0.913	(-0.249, 2.227)	1493
LD2.Recovery	0.303	4.815	0.950	(-9.135, 9.741)	
LD2.Hopelessness	0.090	0.541	0.868	(-0.971, 1.151)	1447
LD2.Recovery	0.172	2.368	0.942	(-4.469, 4.813)	
LD2.Visual Hallucination	0.053	2.105	0.980	(-4.072,4.179)	1427
LD2.Recovery	0.318	16.588	0.985	(-32.194, 32.830)	
LD2.Auditory Hallucinations	0.040	1.101	0.971	(-2.117, 2.198)	1404
LD2.Recovery	0.479	13.675	0.972	(-26.324, 27.281)	
LD2.Paranoia	0.194	5.675	0.973	(-10.930, 11.317)	2000
LD2.Recovery	0.288	22.342	0.990	(-43.501, 44.077)	
LD2.Deservedness	0.090	7.168	0.990	(-13.959, 14.139)	502
LD2.Recovery	-0.110	29.868	0.997	(-58.649, 58.430)	

Table 6:4 Change model results - first-difference method

For each model the fixed effect estimates for the second lagged difference in the predictor $(x_{i-1} - x_{i-2})$ are given alongside the estimates for the second lagged difference in the outcome $(y_{i-1} - y_{i-2})$, i.e. parameters β_1 and β_2 in equation (27). The predictor estimates should be interpreted as the effect of a change in predictor at one time interval on the change in recovery at the following interval, for example the effect of a change in self-esteem from beep 1 to beep 2 on the change in recovery from beep 2 to beep 3.

In contrast to the original analysis, none of the covariates were significantly associated with a change in recovery. The trends appeared to be in the same direction for example, a change in self-esteem at one interval predicts a fall in recovery at the following interval, a change in hopelessness predicting a subsequent increase in the feeling of recovery, but none were significant at the 5% level.

6.3.1.1 Ongoing sign problem within first-difference method

As discussed, the first-difference method reduces the bias caused by the violation of the endogeneity assumption. However, it is not clear whether the method succumbs to the sign problem observed when fitting a change score as the outcome, as presented in Section 6.2. Table 6:4 displays the fixed effects estimates relating to the second lagged difference in the predictor, e.g. β_1 of $\beta_1\left(selfesteem_{i-2,jk} - selfesteem_{i-1,jk}\right)$ from equation (27). Although the estimates appear to follow the same trends as the naïve lagged outcome models of Table 6:2, the confidence intervals around the non-significant estimates suggest that the inverse trend might instead be true.

Theoretically the issue with the sign of β is still present in the first difference method. For the change score model it was shown that the correlation between a lagged covariate and a change score was dependent on the comparative strength of the concurrent relationship between X and Y and the lagged relationship. Considering the current problem as the correlation between the lagged difference of X and the first-difference of Y, $corr(X_{i-1} - X_{i-2}, Y_i - Y_{i-1})$, it follows as before that

$$corr(X_{i-1} - X_{i-2}, Y_i - Y_{i-1}) = \frac{cov(X_{i-1} - X_{i-2}, Y_i - Y_{i-1})}{\sigma_{\Delta X}\sigma_{\Delta Y}}$$

$$= \frac{cov(X_{i-1} - X_{i-2}, Y_i) - cov(X_{i-1} - X_{i-2}, Y_{i-1})}{\sigma_{\Delta X}\sigma_{\Delta Y}}$$

$$= \frac{cov(X_{i-1}, Y_i) - cov(X_{i-2}, Y_i) - cov(X_{i-1}, Y_{i-1}) + cov(X_{i-2}, Y_{i-1})}{\sigma_{\Delta X}\sigma_{\Delta Y}}$$

$$= \frac{corr(X_{i-1}, Y_i)\sigma_{X_{i-1}}\sigma_{Y_i} - corr(X_{i-2}, Y_i)\sigma_{X_{i-2}}\sigma_{Y_i} - corr(X_{i-1}, Y_{i-1})\sigma_{X_{i-1}}\sigma_{Y_{i-1}} + corr(X_{i-2}, Y_{i-1})\sigma_{X_{i-2}}\sigma_{Y_{i-1}}}{\sigma_{\Delta X}\sigma_{\Delta Y}}$$

which, if it is assumed again that the standard deviations of the current and lagged variables are approximately equal, i.e. $\sigma_{X_i} \approx \sigma_{X_{i-1}} \approx \sigma_{X_{i-2}}$ and $\sigma_{Y_i} \approx \sigma_{Y_{i-1}}$,

$$\begin{aligned} = \big[\left(corr(X_{i-1}, Y_i) + corr(X_{i-2}, Y_{i-1})\right) \\ - \left(corr(X_{i-2}, Y_i) + corr(X_{i-1}, Y_{i-1})\right)\big] \frac{\sigma_X\sigma_Y}{\sigma_{\Delta X}\sigma_{\Delta Y}} \end{aligned} \quad (28)$$

where $\Delta Y = Y_i - Y_{i-1}$ and now $\Delta X = X_{i-1} - X_{i-2}$. The sign of the lagged difference correlation $corr(X_{i-1} - X_{i-2}, Y_i - Y_{i-1})$ is thus dependent on the strength of the correlations $corr(X_{i-1}, Y_i) + corr(X_{i-2}, Y_{i-1})$ versus $corr(X_{i-2}, Y_i) + corr(X_{i-1}, Y_{i-1})$.

This relationship will depend heavily on the rate of decline in correlations between lagged observations. The greater the reduction in rate, the faster $corr(X_{i-2}, Y_i)$ will tend to zero and the smaller its contribution will be to the second phrase of the equation. With a large enough decay in the lagged relationship between X and Y this will reduce the problem to one between the sum of two lagged correlations versus a concurrent correlation. If it is assumed that the lagged correlation between X and Y at any two successive time points is approximately equal (i.e. that $corr(X_{i-2}, Y_{i-1}) \approx corr(X_{i-1}, Y_i)$) then in this case the original relationship $corr(X_{i-1} - X_{i-2}, Y_i - Y_{i-1})$ will be zero when

$$\left(corr(X_{i-1}, Y_i) + corr(X_{i-2}, Y_{i-1})\right) - corr(X_{i-1}, Y_{i-1}) = 0$$

i.e. when

$$2 * corr(X_{i-1}, Y_i) - corr(X_{i-1}, Y_{i-1}) = 0$$

$$corr(X_{i-1}, Y_i) = \frac{1}{2} corr(X_{i-1}, Y_{i-1})$$

If the lagged relationship between X and Y is less than half the concurrent relationship then this method exhibits the same problem as described in Section 6.2.3.2.

Such a strong reduction in association by only the second lag may not occur in ESM data, however. The intensive sampling scheme of many ESM studies results in observations taken relatively close together, which may preserve relationships over several lags. Correlation between subsequent measures in the recovery data, for example, is relatively stable with only a small decline at each lag (see Table 6:1), so for this example the full expression of equation (28) should be considered.

6.3.2 INITIAL CONDITIONS MODEL

As an alternative to the first-difference method, the initial conditions approach, whereby equations for the first measurement and subsequent observations are jointly modelled, will also be applied to study predictors of change in the recovery data.

Recall the initial conditions models presented in Section 6.2.3, with the addition of a lagged predictor $x_{i-1,jk}$. Model 1 specifies three equations, the first containing observation 1 on day 1 only, the second containing the first observations on each subsequent day and the third containing all remaining observations:

$$y_{11k} = \alpha_{00} + \alpha_{01} x_{01k} + \lambda_u u_k + \lambda_v v_{1k} + e_{11k} \tag{19}$$

$$y_{1jk} = \alpha_{10} + \alpha_{11} x_{0jk} + u_k + v_{jk} + e_{1jk} \tag{20}$$

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + \beta_2 x_{i-1,jk} + u_k + v_{jk} + e_{ijk}. \tag{21}$$

Model 2 specifies an initial equation containing the first observation on each day and a main equation containing all subsequent observations:

$$y_{1jk} = \alpha_0 + \alpha_1 x_{0jk} + \lambda_u u_k + \lambda_v v_{jk} + e_{1jk} \tag{17}$$

$$y_{ijk} = \beta_0 + \beta_1 y_{i-1,jk} + \beta_2 x_{i-1,jk} + u_k + v_{jk} + e_{ijk.} \tag{18}$$

Each model has the option to constrain the random effect loadings $\lambda_u = \lambda_v = 1$ or to estimate them from the data. It was argued that the rationale for weighting the random effects in the initial equation in a two-level model does not extend to the three-level specification of Model 2. As such, results will be presented for the three-equation specification of Model 1. Finally, two options were presented for the unmeasured x_{0jk} in the initial equations: this term can either be set to $x_{0jk} = 1$ and a variance components model is fitted to the first observations of each day, or the first observed measurement x_{1jk} can be used as a proxy for x_{0jk}.

The aim of this section was to compare the parameter estimates of the naïve approach with the joint modelling approach and examine how estimates vary under different conditions for x_{0jk}. Comparisons will be presented on one variable, self-esteem, for clarity, to demonstrate differences in estimates.

Fixed effects

$x_{0jk} = 1$				$\lambda_u = \lambda_v = 1$		$\lambda_u = \hat{\lambda}_u; \lambda_v = \hat{\lambda}_v$		
	Naive model			**IC Model**		**IC Model***		
$i = 1, Day = 1$	Coeff.	SE	N	Coeff.	SE	Coeff.	SE	N
α_{00} Intercept			1910	4.501	0.201	4.577	0.228	2105
$i = 1, Day > 1$								
α_{10} Intercept				4.943	0.140	4.950	0.147	
Lagged variable at $t > 1$								
β_0 Intercept	3.525	0.156		3.920	0.163	3.899	0.183	
β_2 Self-esteem	0.103	0.026		0.074	0.025	0.075	0.025	
β_1 Recovery	0.295	0.023		0.214	0.021	0.219	0.026	

Random effects

Level	Variance	SE		Variance	SE	Variance	SE	
Person	0.687	0.145		0.925	0.181	0.983	0.215	
Day	0.065	0.018		0.102	0.019	0.102	0.019	
Residuals	0.477	0.018						
$(i = 1, Day = 1)$				0.930	0.226	0.463	0.016	
$(i = 1, Day > 1)$				0.690	0.088	0.917	0.095	
$(i > 1)$				0.460	0.017	0.463	0.311	
						Coeff.	SE	
λ_u						1.020	0.189	
λ_v						-0.026	1.182	

Table 6:5 Comparison of naive model estimates with initial conditions method. Three equation specification joint modelled with $x_0 = 1$ in initial equations.

* Free estimation of λ_u and λ_v required the models to be fitted using STAN. As such, mean and standard errors for the parameter distributions are presented rather than point estimates and associated standard errors.

When fitting a random components model for the initial equations (i.e. $x_{0jk} = 1$) the change in results from the naïve model are as expected: estimates for lagged recovery β_1 are reduced, with greater variation in the random effects. Interestingly, estimates of lagged self-esteem β_2 are also reduced, reflecting the lack of dependence in this covariate for the first observations as anticipated.

Wald tests of the random effect loadings suggest that there is more subject-level random variation at time point 1 ($z = 5.40,\ p < 0.001$), while λ_v is not significantly different than zero at the 5% level ($z = -0.022, p = 0.982$)

Fixed effects

$\boldsymbol{x_{0jk} = x_{1jk}}$				$\lambda_u = \lambda_v = 1$		$\lambda_u = \hat{\lambda}_u; \lambda_v = \hat{\lambda}_v$		
	Naive model			**IC Model**		**IC Model***		
$\boldsymbol{i = 1, Day = 1}$	Coeff.	SE	N	Coeff.	SE	Coeff.	SE	N
α_{00} Intercept			1910	4.508	0.120	4.573	0.235	2099
α_{01} Self-esteem				0.239	0.136	0.231	0.162	
$\boldsymbol{i = 1, Day > 1}$								
α_{10} Intercept				4.968	0.30	4.960	0.129	
α_{11} Self-esteem				0.292	0.048	0.293	0.050	
Lagged variable at $\boldsymbol{t > 1}$								
β_0 Intercept	3.525	0.156		3.812	0.156	3.779	0.180	
β_2 Self-esteem	0.103	0.026		0.126	0.026	0.125	0.026	
β_1 Recovery	0.295	0.023		0.236	0.021	0.241	0.027	

Random effects

Level	Variance	SE	Variance	SE	Variance	SE
Person	0.687	0.145	0.790	0.157	0.836	0.179
Day	0.065	0.018	0.088	0.018	0.088	0.018
Residuals	0.477	0.018				
$(i = 1, Day = 1)$			0.927	0.232	0.469	0.017
$(i = 1, Day = 1)$			0.258	0.073	0.602	0.076
$(i > 1)$			0.467	0.017	0.469	0.325
					Coeff.	SE
λ_u					0.981	0.225
λ_v					-0.083	1.256

Table 6:6 Comparison of naive model estimates with initial conditions method. Three equation specification joint modelled with $x_0 = x_1$ in initial equations.

* Free estimation of λ_u and λ_v required the models to be fitted using STAN. As such mean and standard errors for the parameter distributions are presented rather than point estimates and associated standard errors.

When substituting x_1 for x_0 in the initial equations the effect of self-esteem on change in recovery is increased in the main equation, with again more variation in the random effects.

As with the models for $x_0 = 1$, Wald tests suggest that only the subject-level random effect loading is significant, in this parameterisation of x_0 the initial equation estimating less subject-level variation than the main equation.

The models with $x_{0jk} = 1$ and $x_{0jk} = x_{1jk}$ both perform as expected. Comparing the corresponding models' AIC suggests that using x_{1jk} as a substitute results in better model fit (e.g. comparing models where $\lambda_u = \lambda_v = 1$: For $x_{0jk} = 1$ AIC = 4919.89, while for $x_{0jk} = x_{1jk}$ AIC = 4878.25). However, as neither method correctly specifies this initial equation, the extent to which the fixed and random effect estimates may still be biased remains unclear.

6.4 SUMMARY

Studying predictors of change in ESM requires careful consideration. While the data structure allows for the exploration of momentary patterns due to the intensive sampling and range of data collected at each time point, many straightforward analyses are not appropriate. Using a change score as the dependent variable can result in the reversal of fixed effect directions depending on the strength of concurrent and lagged relationships. Using dynamic, or lagged outcome, models as an alternative violates assumptions of multilevel models, resulting in biased estimates of both the fixed and random effects. The first-difference method and initial conditions method for overcoming this bias were presented, with extensions to a three-level data structure. Consideration of these methods in the context of ESM concluded that neither provide a definitive resolution to the problems encountered: the first-difference method may still succumb to the sign problem as in its solution to endogeneity it computes a change score in the outcome; and including time lagged variables as predictors of change when joint modelling the initial condition may induce further bias due to misspecification in the initial equation. As such, ESM research into predictors of change should be carefully considered to ensure that this is the most appropriate question. If pursued, the limitations of each method should be clearly stated.

7 Power and sample size

One of the most important questions when conducting any study is that of power and sample size: how many participants are required to detect a clinically significant effect in the data. For ESM studies where multiple observations can be taken across several days or weeks for each subject, the total sample size N is no longer just the number of participants. Instead this N is partitioned into three: the number of subjects n_3, the number of days n_2 and the number of measurements within days n_1, resulting in a total $N = n_1 \times n_2 \times n_3$. Throughout, this total sample size will be referred to in terms of the number of subject and the sampling scheme, which consists of the choice of number of measurements per day and the number of days of observation. As each component of N can be varied to contribute to the overall sample size, defining the optimum combination to detect a meaningful difference at the design stage of the study is a complex problem.

The following chapter will outline power and sample size calculations for ESM studies. Firstly the extent to which sample size is considered in current practice will be presented using the papers identified in the systematic review. This will be followed by an introduction to the concept of statistical power and a presentation of closed form sample size formulae available for two- and three-level data. Empirical power calculation using Monte Carlo simulation will then be discussed. The final section of this chapter will present the results of using Monte Carlo simulation to estimate power in an ESM context for different research questions. Code developed for the implementation of such simulations in Stata will also be presented.

7.1 Power in current ESM research

In addition to research questions and statistical models, the systematic review described in Chapter 3 recorded whether power was considered a priori. Only one study of the 74 reviewed provided justification of their sampling scheme. Meyers et al. (2012) stated that for their hierarchical linear model an a priori power calculation was carried out using Gpower (Faul, Erdfelder et al. 2007), which for a multiple regression with two predictors required a sample size of 89 participants. They further referenced an unpublished report by Kreft (1996), in which the author finds that 25 repeated observations provide sufficient power for 60 subjects, as justification for their sampling scheme of five observations a day over five days for their 99 participants. In addition, the authors report that a post-hoc power calculation was carried out which found their design had 31% power to detect a small effect size, 77% power to detect a medium effect size and 95% power to detect a large effect size.

While it is promising that at least one study considered power while designing their ESM study, the justifications made by Meyers et al are weak. Firstly, their a priori power calculation to decide on the number of subjects required was based on a single level rather than multilevel model, ignoring any correlation between observations within subjects. Moreover, the description of this calculation stated that a maximum of two predictors were used in their data analysis, however their most complex, moderation model involves an interaction which appears to be unaccounted for.

The post hoc power analysis is also problematic. Power was given for three effect sizes, 'small', 'medium' and 'large', however no explanation was given as to what the values of these effect might be. It is presumed these are reference to Cohen (1988)'s small, medium and large criteria, but no reference is given. Furthermore, it was unclear which model motivated this post-hoc power calculation and for which parameter power was estimated. Two two-level random intercept models were used for the most complex analysis, investigating whether "thin ideal internalization" or feminist beliefs, both subject-level variables, moderated the relationship between social comparisons and "body image disturbance" (a composite score of State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991) and the Body Checking Questionnaire (BCQ; Reas et al., 2002)) which were repeatedly measured in the ESM diary. The hypotheses suggest the two cross-level interactions were of main interest, with estimates of $\beta = 0.025$ and $\beta = 0.00$ given for thin-ideal internalization and feminist beliefs respectively, however it is unclear whether these regression coefficients were standardised or not and no standard errors were provided. The conclusions of their post-hoc power calculation therefore remain unclear. The authors briefly comment on the "relatively small" sample size in the limitations of the study, but do not expand on whether it was sufficiently powered to detect the effects of the moderators.

7.2 Considerations for power and sample size calculations

Studies are powered a-priori so as to be able to detect some pre-specified effect. This effect size should correspond to a clinically meaningful value the researchers hypothesise exists in this population. Defining a null hypothesis of no effect and an alternative hypothesis of some effect $\beta \neq 0$, four scenarios are possible, listed in Table 7:1. Two scenarios describe the ideal results that can occur: either this difference exists in the population and it is detected in the sample, in which the null hypothesis is correctly rejected, or the difference does not exist and it is not found. The remaining two scenarios are less desirable: false negative and false positive results. A false negative result occurs when the expected difference exists in the population but it is not detected in the sample, the hypothesis test thus returning a negative result where a positive should be found. This is also known as a type II error and is denoted by γ. A false positive result is obtained when the difference does not exist in the population but it is detected in the sample. This is referred to as a type I error, or significance level, and is denoted by α. Ideally when designing a study researchers should aim to minimise the probability of both types of errors. More often than referring to the type II error, researchers are interested in $1-\gamma$, or the probability of detecting an effect where one exists. This is known as the power of a study and in aiming to minimise γ, $1-\gamma$ is maximised. Convention typically dictates a significance level of $\alpha = 0.05$ and power of $1-\gamma = 0.8$ or 0.9 corresponding to a 5% probability of failing to observe a true effect and an 80% or 90% probability of detecting an effect when one is present. While these values are somewhat arbitrary, they reflect the trade-off between feasibility and ideal circumstances.

	Null hypothesis true	**Null hypothesis false**
Reject the null	Type I error False positive	True positive
Do not reject the null	True negative	Type II error False negative

Table 7:1 Possible scenarios of statistical hypothesis tests

Power is related to sample size through the test statistic of the parameter of interest. In a regression model the fixed effect parameters are subject to a Wald test in which the test statistic

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

is used to test the null hypothesis $\beta = 0$ versus the alternative $\beta \neq 0$. The test statistic is larger for larger effect sizes and for smaller $SE(\hat{\beta})$, that is more precise estimates of $\hat{\beta}$. As Cohen (1988) describes, the reliability of the estimate may depend on factors such as measurement error, however it will always depend on sample size: the larger the sample size, the smaller the standard error and the more precise the estimate. Thus power, effect size and sample size are inherently related.

When designing a study the relationship between power and sample size can be used to ensure that enough participants are sought to enable an effect to be detected in the sample. Collecting data on too few subjects may result in an imprecise estimate, whereas collecting data on too many subjects would waste both time and resources. To decide on an N a priori, power formulae or sample size calculations can be used. A general form for power for which sample sizes can be derived is

$$1-\gamma=\Phi\left(\frac{\delta}{se(\delta)}-z_{\alpha/2}\right)$$

where $(1-\gamma)$ is the power, δ is the difference between two groups, $se(\delta)$ its standard error and $z_{\alpha/2}=\Phi^{-1}(1-\alpha/2)$ for significance level α.

7.3 POWER FOR LONGITUDINAL DATA

When data are longitudinal and multiple observations are taken for each subject, sample size N contains both the number of subjects n_2 and the number of observations per subject n_1. Power calculations for specific tests need to reflect the correlation now present time point to time point. For two levels, this within-subject correlation is defined as $\rho=corr(y_{ik},y_{i'k})$ for observations $i\neq i'$. Variation in y is also now partitioned into between-subject variation σ_u^2 and residual variation σ_e^2.

Diggle (2002) provides basic sample size calculations for comparing two independent groups A and B for measurements on a continuous, time varying covariate x_{ik} within cluster. He assumes the same number of observations are taken for each subject k, simplifying $x_{ik}=x_i$. For comparing two groups with a continuous response he states that the number of subjects needed per group ($n_2/2$), with n_1 observations per person, for a type 1 error = α and power $1-\gamma$ is

$$\frac{n_2}{2}=\frac{2\left(z_{\alpha/2}+z_\gamma\right)^2\sigma_e^2(1-\rho)}{n_1\,s_x^2\delta^2} \tag{29}$$

where ${\sigma_e}^2=var\left(e_{ij}\right)$, $\rho=corr(y_{ij},y_{i'j})$ for $i\neq i'$, $\delta=\beta_{1B}-\beta_{1A}$ when β_{1A} and β_{1B} are the effect of x on y for groups A and B, and z_p is the pth quantile of the standard Gaussian distribution. Here $s_x^2=\ \Sigma_i(x_i-\bar{x})^2/n_1$, the within-subject variance of x_i.

This formula shows that the number of subjects increases with larger residual variation σ_e^2 but decreases with within-subject variation in x, group difference and within-subject correlation in y.

Fitzmaurice, Laird et al. (2012) expands on this simple scenario, presenting a sample size formula for a random slope model comparing group differences in time trends. Here the random slope is included to allow the time trends to vary between subjects. The notation for these equations switches to covariate t_{ik} to represent time, rather than Diggle's time-varying covariate x_{ik}. The total number of subjects can be calculated

$$n_2 = \frac{\{z_{\alpha/2} + z_{(1-\gamma)}\}^2 \sigma_\beta^2}{\pi(1-\pi)\delta^2} \tag{30}$$

where

$$\sigma_\beta^2 = \sigma_e^2 \left\{ \sum_{i=1}^{n_1} (t_i - \bar{t})^2 \right\}^{-1} + \sigma_{u1}^2 \tag{31}$$

denotes the within- and between-subject variation in slopes, δ the treatment effect of interest and π the proportion of subjects in group 1.

Snijders (2005) posits that sample sizes for multilevel designs can be calculated by the "sample size for a simple random design, multiplied by the design effect", defining this design effect (DE) as

$$DE = \frac{\text{SE}^2 \text{ under this design}}{\text{SE}^2 \text{ under standard design}}$$

If DE < 1 he states the multilevel design is more efficient than the simple design, if DE > 1 the multilevel design is less efficient. For two-level designs he defines the DE when estimating different parameters. These formulae are presented in Table 7:2. The notation is as before, with the extension to random slope models requiring the between-subject random variance σ_u^2 to be expressed in terms of variation in random intercepts σ_{u0}^2 and variation in random slopes σ_{u1}^2.

Estimating parameter	Design Effect
Population mean β_0, Random intercept model	$1 + (n_1 - 1)\rho \geq 1$
Level 1 coefficient β_1, Random intercept model	$1 - \rho \leq 1$
Level 1 coefficient β_1, Random slope model	$\dfrac{n_1\sigma_{u1}^2 s_{x1}^2 + \sigma_e^2}{\sigma_{u1}^2 s_{x1}^2 + \sigma_{u0}^2 + \sigma_e^2}$
Level 2 coefficient β_2, Random intercept model	$1 + (n_1 - 1)\rho \geq 1$

Table 7:2 Design effects for estimating fixed effects

Using this reasoning, Snijders (2005) suggests that if the parameters of interest are the population mean or level 2 coefficients then using a sample size derived from a simple random design would yield more efficient estimates then a multilevel design. For level 1 coefficients in a random intercept model a multilevel design is more efficient, while in a random coefficient model the design effect will depend on the variance of the random effects. Finally, Snijders states that while power will depend on sample size at each level, a greater number of higher-level units are preferable to more observations within units, and although small level 1 units will not be problematic for estimating fixed, this will result in low power for testing random slope variances.

7.3.1 Power for three-level data

For three-level data the sample N can now be partitioned into three, with $i = 1, \ldots, n_1$ nested within $j = 1, \ldots, n_2$ nested within $k = 1, \ldots, n_3$. Similarly, the total model variance σ^2 can be partitioned into σ_u^2 at level 3, σ_v^2 at level 2 and σ_e^2 at level 1.

Heo and Leon (2008) provide formulae for power and sample size calculations for three-level models to detect a level 3 treatment effect $\delta = \bar{y}(1) - \bar{y}(0)$, where for group $g = 0, 1$,

$$\bar{y}(g) = \frac{1}{n_3 n_2 n_1} \sum\nolimits_{k=1}^{n_3} \sum_{j=1}^{n_2} \sum\nolimits_{i=1}^{n_1} y_{ijk}$$

for a balanced design. They define the variance of the group mean as

$$Var\big(\bar{y}(g)\big) = \frac{f\sigma^2}{n_3 n_2 n_1}$$

where $\sigma^2 = \sigma_u^2 + \sigma_v^2 + \sigma_e^2$ and f is design effect or variance inflation factor,

$$f = 1 + n_1(n_2 - 1)\rho_2 + (n_1 - 1)\rho_1.$$

The ρ are defined as the ICC for level 2 and level 1,

$$\rho_2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}$$

$$\rho_1 = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}.$$

Finally, the power calculation for test statistic $T = \frac{\hat{\delta}}{se(\hat{\delta})}$ is given as

$$1 - \gamma = \Phi\left\{\Delta\sqrt{\frac{n_3 n_2 n_1}{2f}} - z_{\alpha/2}\right\} \tag{32}$$

where $\Delta = \delta/\sigma$ is the standardised effect size, α is the two-sided significant level, γ the probability of a type II error and Φ the cumulative distribution function of a standard normal distribution.

Rearranging (32), the sample size for level 3, level 2 and level 1 are given by

$$n_3 = \frac{2f\{z_{\alpha/2} + z_{(1-\gamma)}\}^2}{n_2 n_1 \Delta^2}$$

$$n_2 = \frac{2\{1 + (\rho_1 - \rho_2)n_1 - \rho_1\}\{z_{\alpha/2} + z_{(1-\gamma)}\}2}{n_1 n_3 \Delta^2 - 2\rho_2 n_1\{z_{\alpha/2} + z_{(1-\gamma)}\}^2}$$

and

$$n_1 = \frac{2(1 - \rho_1)\{z_{\alpha/2} + z_{(1-\gamma)}\}^2}{n_2 n_3 \Delta^2 - 2\{(n_2 - 1)\rho_2 + \rho_1\}\{z_{\alpha/2} + z_{(1-\gamma)}\}2}.$$

A closed form power function to detect a time varying treatment effect for a three-level random slope model has been derived by Heo, Xue et al. (2013). Their paper outlined the sample size calculation for a longitudinal cluster randomized trial for a balanced design. For occasions $i = 1, \ldots, n_1$ nested within individuals $j = 1, \ldots, n_2$ within cluster $k = 1, \ldots, 2n_3$ they present the model

$$y_{ijk} = \beta_0 + \beta_1 x_k + \beta_2 t_{ijk} + \beta_3 x_k t_{ijk} + u_k + v_{0jk} + v_{1jk} t_{ijk} + e_{ijk}$$

where the level 3 intervention indicator x_k takes the value 0 or 1 representing a control or intervention cluster respectively, t_{ijk} is a level 1 time variable with equal unit spacing, and the random slope v_{1k} allows for subject-specific time trends. All random terms are assumed to be normally distributed with cluster-level random intercept $u_k \sim N(0, \sigma_u^2)$, subject-level random intercept $v_{0jk} \sim N(0, \sigma_{v_0}^2)$ and random slope $v_{1jk} \sim N(0, \sigma_{v_1}^2)$, and residual error $e_{ijk} \sim N(0, \sigma_e^2)$. It is assumed that all four random effects are independent. The parameter of interest is β_3, representing the difference in mean slopes between the two intervention groups.

The OLS estimate of β_3 is given by $\hat{\beta}_3 = \hat{\eta}_1 - \hat{\eta}_0$ where $\hat{\eta}_g$ is the OLS estimate of the slope for intervention group $g = 0, 1$. They derive the test statistic T to test $\beta_3 = 0$

$$T = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} = \frac{(\hat{\eta}_1 - \hat{\eta}_0)\sqrt{n_3 n_2 n_1 s_t^2}}{\sqrt{2\{(1-\rho_1)\sigma^2 + n_1 s_t^2 \sigma_{v_1}^2\}}} \tag{33}$$

where $s_t^2 = \sum_{i=1}^{n_1} \left(t_{ijk} - \bar{t}\right)^2 / n_1$ is "the population variance of time variable t" and $\rho_1$is the ICC under the fixed slope model. The power function is then expressed as

$$1 - \gamma = \Phi\left\{\Delta \sqrt{\frac{n_3 n_2 n_1 s_t^2}{2\{(1-\rho_1) + \rho_{v_1} n_1 s_t^2\}}} - z_{\alpha/2}\right\} \tag{34}$$

where $\Delta = \beta_3/\sigma$ is the effect size and $\rho_{v_1} = \sigma_{v_1}^2 / (\sigma_u^2 + \sigma_v^2 + \sigma_e^2) = \sigma_{v_1}^2/\sigma^2$ is the ratio of the random slope variance to the sum of the other variances. This power function increases with larger effect sizes and correlation between level 1 units, but decreases with the random slope variance.

The sample sizes at level 3 and level 2 can then be calculated as

$$n_3 = \frac{2\{(1-\rho_1) + \rho_{v_1} n_1 s_t^2\}\{z_{\alpha/2} + z_{(1-\gamma)}\}^2}{n_2 n_1 s_t^2 \Delta^2}$$

$$n_2 = \frac{2\{(1-\rho_1) + \rho_{v_1} n_1 s_t^2\}\{z_{\alpha/2} + z_{(1-\gamma)}\}^2}{n_3 n_1 s_t^2 \Delta^2}.$$

As s_t^2 is a function of n_1, the sample size for level 1 must be determined iteratively, with

$$n_1 = \frac{(1-\rho_1)}{s_t^2\left[n_3 n_2 \Delta^2 / \left\{2\left(z_{\alpha/2} + z_{(1-\gamma)}\right)^2\right\} - \rho_{v_1}\right]}.$$

The formulation of these equations is such that power and the level 1 sample size are only determined from the product of the level 2 and level 3 sample size, $n_3 n_2$, rather than their individual terms. Consequently, desired power for level 1 sample size can be based on various combinations of cluster size and units within clusters, thus allowing for greater flexibility in design.

The authors compared these formulae to empirical MLE based estimates of sample size using simulation and found the results to be almost identical. Furthermore, they investigated the fallout from designing a study powered for a more simple model than used in the analysis. This was determined by evaluating the effect of random slope variance on level 3 sample size. They found that even for small variance $\sigma_{v_1}^2$ the ratio of this to the total variance ρ_{v_1} can be large and have a considerable effect on sample size. As such, studies designed under a fixed slope model can be dramatically underpowered when there is between-subject variation in slopes.

While useful for the design of cluster randomised trials, the power formula derived by Heo and Leon (2008) and Heo, Xue et al. (2013) correspond to detecting an effect at level 3. In ESM research momentary level effects are typically of interest, that is differences at level 1. Cunningham and Johnson (2012) consider designs where treatment or intervention is randomised at either level 3, level 2 or level 1. For the simple scenario testing a treatment effect at level 1 with no covariates they derive the variance estimator $var(\hat{\beta})$ required for a Wald test of treatment effect $\beta \neq 0$ when randomisation occur at each of the three levels. The design effects are then presented as

Randomisation at level 3: $DE = 1 + (n_1 - 1)\rho_1 + n_1(n_2 - 1)\rho_2$ (35)

Randomisation at level 2: $DE = 1 + (n_1 - 1)\rho_1 - n_1\rho_2$

Randomisation at level 1: $DE = 1 - \rho_1$ (36)

7.3.2 Empirical power

Power formulae for three-level models are at present quite limited, and although the formulae presented thus far are by no means exhaustive, demonstrate how closed form expressions are limited to certain designs. Indeed, while ESM may be used to deliver an intervention many studies are observational and do not depend on randomisation. The power formulae described for three-level models thus far would not be appropriate in this case. Literature for power calculations in three-level models is limited and very much lacking in a non-RCT framework; an alternative approach to investigate power and sample size is to use Monte Carlo simulation.

Instead of being reliant on designs where power formulae are derived or for scenarios where when no closed form exists, empirical power estimates can be obtained via simulation (Landau and Stahl 2013). Here data are generated under a set of conditions and empirical power is estimated in terms of the proportion of samples for which a pre-specified null hypothesis is rejected. Heo and Leon verified the accuracy of their formulae (presented above, based on power function (32)) using this type of simulation. The level 3 sample size was simulated for increasing levels of $n_2,\ n_1,\ \rho_2,\ \rho_1$ and standardised effect size Δ. The tabulated results allow the choice of a level 3 sample size based on the other sample size values, ICCs and for a particular effect size.

Maas and Hox (2005) use simulation to investigate bias in fixed and random effects estimates with different level 1 and 2 sample sizes. They corroborate the findings of Snijders (2005) that a larger number of groups is preferable to a large group size but found that parameter estimates and level 1 variance estimates are generally unbiased with a smaller sample size. Problems are instead observed in the estimates of the random effects where the level 2 variance may be underestimated and the standard error of variance components will be too small. Maas and Hox also considered the effect of small sample sizes, simulating data with 10 level 2 units with a cluster size of 5, equivalent to an ESM study with 10 participants and 5 diaries per person, with varying levels of intraclass correlation. They found that regression coefficients and level 1 variance components exhibited negligible bias, but the level 2 variance was estimated as much too large with bias up to 25%, and the standard errors for both the regression coefficients and variance components were too small, particularly at Level 2.

7.3.3 AVAILABLE SOFTWARE

There are several software packages available to compute power in two level models such as PinT (Power analysis IN Two-level designs) (Snijders and Bosker 1993), and Bolger and Laurenceau (2013) give details for simulation using MPlus. For three level models Browne, Golalizadeh Lahi et al. (2009) have developed MLPowSim which generates either R or MLWiN code for up to three level models and can accommodate both balanced and non-balanced data sets. Though the code is comparatively fast to run in R, the program requires estimates of model and covariate variance at each of the different levels, information which may be unavailable in ESM research. There are currently no commands for three level power calculations in Stata.

7.4 Determining sample size for ESM studies

Due to the multilevel nature of ESM data, sample size can be thought of as being partitioned into three levels: level 1 – the number of moments to take measurements on each day; level 2 – the number of days to be observed over; and level 3 – the number of participants to be recruited. Powering an ESM study thus requires a balance of sample sizes at all levels. Literature suggests that increasing the highest number of units has a greater effect on power (Maas and Hox 2005; Snijders 2005), implying that for ESM studies recruiting more participants is the most efficient way to increase power. However, in practice constrains to sample size may be unavoidable at each level when designing an ESM study, which limit the flexibility of simply increasing level 3 units. The number of participants may have an upper limit prior to starting study for several reasons. Firstly, study size may be dependent on resources and time available for recruitment, with smaller scale studies struggling to recruit large numbers of participants, for example. Alternatively, the proportion of eligible participants may be low for rare conditions or diseases, or when researching a combination of behaviours and health status, such as drug use and bipolar disorder for example. Moreover, recruitment for certain populations, such as those exhibiting 'risky behaviours', may limit the number of available participants. Equally at level 1, the number of observations per day may be capped as researchers may wish to follow a pre-specified sampling regime, either designed to lessen the burden of this intensive sampling methodology or on advice from a research group (Delespaul 1995). At level 2, a pre specified study period may be required on the basis of restricted time or funding, or to limit drop out or missing data due to participant fatigue. Conversely, a set time period may be required in order to capture a minimum number of events or phenomena. Each of these factors contribute to the overall sample size of an ESM style study and will restrict how the three components combine to produce a study design with sufficient power.

This simulation study aimed to determine sufficient power for varying sample sizes at levels 1, 2, and 3 to reflect the possibility of at least one element of the sample size will be fixed by design. The results can then be interpreted by holding sample size at one level fixed and varying the remaining two to compare power with different combinations of n contributing to the full sample size. For example, for a sampling scheme of 10 prompts per day the effect of varying the number of days and participants could be compared. Alternatively, if recruitment was expected to be limited to 30 subjects the simulation results could be used to inform whether more measurements per day or additional study days would provide greater power.

Power graphs will be presented for two scenarios: a simple association model and a group difference model. The simple association model will investigate power for varying sample sizes when estimating the association between two moment-level variables

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_k + v_{jk} + e_{ijk}. \tag{37}$$

A group difference model will compare the means of two groups, where groups $G_k = 0, 1$ are defined at level 3

$$y_{ijk} = \zeta_0 + \zeta_1 G_k + u_k + v_{jk} + e_{ijk} \tag{38}$$

This could be used to explore how a momentary level measure differs between two treatment groups, for example. Although many more scenarios are possible and more complex models can be accommodated using this procedure, these two simple examples were chosen to demonstrate the method and to highlight the difference in sample sizes required when the parameter of interest is measured at different levels.

7.4.1 Simulation protocol

As the primary purpose of the simulation procedure is to investigate power due to sample size rather than test for the consequences of model misspecification, the data generating model for the outcome will match the estimation model for each scenario, i.e. the data generating model for y in the association example uses equation (37) and the data generating model for y in the group difference example will use equation (38). For the association example the covariate x will be created using a three-level variance components model to allow for the subject-level, day-level and residual variation in this level 1 predictor

$$x_{ijk} = \alpha_0 + u_{x,k} + v_{x,jk} + e_{x,ijk}$$

where $u_{x,k} \sim N\left(0, \sigma^2_{x_{L3}}\right)$, $v_{x,jk} \sim N\left(0, \sigma^2_{x_{L2}}\right)$ and $e_{x,ijk} \sim N(\bar{x}, \sigma^2_{x_{L1}})$, and $\bar{x}$ is the expected mean of x at level 1.

Sample size for each simulation will vary by number of participants n_3 from 10 to 60 in intervals of 10, number of days n_2 from 6 to 10 in intervals of 1 and number of observations n_1 from 4 to 10 in intervals of 2.

Level of sample size	Number of units within each level
Number of participants (n_3)	10, 20, 30, 40, 50, 60
Number of days (n_2)	6, 7, 8, 9, 10
Number of observations (n_1)	4, 6, 8, 10

Table 7:3 Sample sizes for simulation procedure

These sample sizes were chosen to investigate power in smaller ESM sized studies representative of those within mental health research where recruitment is often an issue. The number of days was varied between six and 10 on the advice of the Maastricht group (Delespaul 1995) that a week is adequate to observe sufficient variation in mood and symptoms. The group also suggest sampling 10 semi-random beeps per day in order to capture events of interest. Across a week, however, this is quite an intensive sampling procedure and has been criticized as potentially effecting responses (Robins, Fraley et al. 2009). The number of measurement moments was therefore varied between four and 10 per day to investigate whether sufficient power is possible with a less burdensome sampling scheme.

A priori parameter estimates are also required for the simulation procedure. Typically these estimates are informed by published studies, prior data or expert knowledge. In ESM research, however, many of the required values will be unavailable: although there is a growing literature of ESM studies in various fields, reporting of model parameters is still widely varied with model variance estimates, for example, often underreported. Therefore, in addition to varying the sample size at each level, the extent to which uncertainty in a priori estimates effect power will also be investigated: for each parameter, small, medium and large values will be defined and varied within each sample.

Fixed effect estimates will be varied based on advice of Cohen (1988) for effect sizes. Effect sizes for the association model will be set at $r = 0$ for the null case, $r = 0.1$ (small effect), $r = 0.3$ (medium effect), and $r = 0.5$ (large effect). These effect sizes are then scaled by the standard deviation of the predictor and outcome to create the regression coefficient β_1 to be used in the data generating model (37) such that

$$\beta_1 = r\frac{\sigma_y}{\sigma_x} \tag{39}$$

where $\sigma_y = \sqrt{\sigma_u^2 + \sigma_v^2 + \sigma_e^2}$ and $\sigma_x = \sqrt{\sigma_{x_{L3}}^2 + \sigma_{x_{L2}}^2 + \sigma_{x_{L1}}^2}$, with $\sigma_{x_{L\#}}^2$ denoting the variance in x at level $\# = 1, 2, 3$.

For the group difference model (38), Cohen's d was used

$$d = \frac{\bar{y}_1 - \bar{y}_0}{\sigma_y}$$

where $\bar{y}_0$ denotes the mean for group 0, $\bar{y}_1$ the mean of group 1 and σ_y the standard deviation of either group (as they are assumed to be equal). The effect size d can be transformed into a regression coefficient to be used in the data generating model using

$$\zeta = \frac{d}{\sqrt{d^2 + (1/\pi(1-\pi))}}$$

where π is the proportion of subjects in group 1 and $1 - \pi$ the proportion of subjects in group 0 (Cohen 1988). This effect size will be varied based on Cohen's small, medium and large criteria, where a small effect size $d = 0.2$, a medium effect size $d = 0.5$ and a large effect size $d = 0.8$.

Although widely used as definitive measures of effect sizes, Cohen warns the small/medium/large criteria are to be evaluated in context, and only used as convention when no other definitions are available. As effect sizes are depending on population and ESM can be applied in a wide range of areas these categorisations are considered appropriate. However, in practice effect sizes for power calculations should be based on clinically meaningful values.

Estimates for both the model variances and predictor variances will be varied based on unit spread in the common 7 point Likert scale. Values will be chosen such that 95% of points will lie 1 scale unit either side of the mean, 1.5 units either side of the mean and 3 units either side of the mean. This 'small', 'medium' and 'large' categorisation correspond to variance estimates of $\sigma^2 = 0.26, 0.59$ and 2.34 respectively. The model variance, predictor variance and effect size estimates to be used for simulation are summarized in Table 7:4.

	Effect size r	Effect size d	Model variances $\sigma_u^2, \sigma_v^2, \sigma_e^2$	Predictor variances $\sigma_{x_{L3}}^2, \sigma_{x_{L2}}^2, \sigma_{x_{L1}}^2$
Small	0.1	0.2	0.26	0.26
Medium	0.3	0.5	0.59	0.59
Large	0.5	0.8	2.34	2.34

Table 7:4 Effect sizes and variance estimates for simulation procedure

First, power will be estimated for complete data. This assumes in an ESM context that all diaries are returned completed. However, evidence suggests that missing data is prevalent in ESM studies and can occur at the item, moment or day level. A detailed account of missing data in current research as well as the recovery data example can be found in Chapter 4. Following the main simulations in this section, the effect of missing data on power will be explored. The systematic review of Chapter 3 found that missing data rates varied widely in practice from less than 1% up to 70%, with a median of 20% moment nonresponse. Examining missing data in the recovery study data uncovered trends in the pattern of missing diaries, with a greater proportion uncompleted at the start and end of the day as well as towards the end of the week. Using the recovery data to motivate parameter estimates, this missing data process will also be modelled in the power simulations where data are deleted using the logistic process

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 s_j \tag{40}$$

where π_{ijk} is the probability that observation i is missing.

Three missing data scenarios will therefore be explored: a randomly drawn 20% missing and 70% missing, and missing based on time trends of the logistic process of equation (40) with coefficients $\beta_0 = -0.78, \beta_1 = -0.56, \beta_2 = 0.05$ and $\beta_3 = 0.28$.

The number of simulations for each combination of n and parameter estimates was based on a power calculation for the precision of the proportion of samples for which the null hypothesis is rejected. Each set of simulations will be carried out to determine the sample size necessary at each level to achieve a minimum of 80% power. The confidence interval for the proportion p of significant estimates is defined as

$$p \pm z\, se(p)$$

or confidence width $w = 2 * z\, se(p)$ where $se(p) = \sqrt{}(p(1-p)/S)$, with S the number of simulations and $z = 1.96$ for a 95% confidence interval. To restrict the confidence interval to be no wider than 1% $(w = 0.01)$ when estimating $p > 0.8$,

$$S \geq \frac{4z^2\big(p(1-p)\big)}{w^2}$$
$$\geq 24586.24$$

To allow for a slightly wider confidence interval, letting $w = 0.05$, the number of simulations required is far less, with $S \geq 983.4496$. Due to the speed of the simulation procedure in Stata the latter option was used.

Each simulation will proceed through the following steps.

1. Three-level data will be generated under the assumption that the estimated parameter values represent the true population estimates, with the response variable generated as per the analysis model
2. The appropriate analysis model (equation (37) for level 1 associations and equation (38) for group differences) will be fitted to this generated data with fixed and random effect estimates stored and statistical test of an appropriate null hypothesis computed.
3. For the total number of simulations S, the proportion of samples which reject the null hypothesis will provide an empirical estimate of power.

This process will be repeated for each combination of n_3, n_2 and n_1 given in Table 7:3 for varying effect sizes, model variances and predictor variances to provide an estimate of power under each condition. The code for each of the two scenarios (level 1 association and level 3 group difference) is provided in Appendix 5: Power and sample size. The results will then be presented in graphically to allow for easy comparison of varying each n.

Note that the data generating model and analysis model will be identical within each example, as the purpose of this simulation study is to estimate empirical power based on various sample sizes rather than evaluating the performance of the analysis model through precision of parameter estimates.

Finally, to examine the effect of missing data, a missing value generating model will follow the data generating model to eliminate a proportion of diaries based on the missing data levels of 20% and 70%, corresponding to the median and maximum levels of missing diary data observed in the systematic review.

7.4.2 Results of simulations

The results of the power simulations for each research question will be presented in this section. For each question, the empirical power estimates will be presented for each combination of sample size when varying effect sizes and variance estimates. Results will be presented using power graphs with corresponding power tables filed in Appendix 5: Power and sample size. In order to isolate the effect of each element on power, when varying each component all others will remain fixed. For example, the degree to which effect size influences power will be presented first, holding both predictor and model variances constant. The variances will then be examined in turn, for a fixed effect size. For each example, optimum sample sizes will be considered the smallest combination of n for which 80% power is achieved.

7.4.2.1 Simple association model

Simulations for an effect size of zero were first run to compare the empirical type I error rate to the specified significance level of $\alpha = 0.05$. Over all sample sizes the proportion rejected ranged between 0.037 and 0.072. This is considered within an acceptable range, indicating little bias in the small samples.

7.4.2.1.1 Results of varying effect sizes

Figure 7:1 depicts the varying levels of power achieved by Monte Carlo simulation for the association example with the smallest effect size $(r = 0.1)$ and the smallest predictor and model variances $(\sigma^2_{x_{L3}} = \sigma^2_{x_{L2}} = \sigma^2_{x_{L1}} = 0.26$ and $\sigma^2_u = \sigma^2_v = \sigma^2_e = 0.26)$. The five graphs correspond to the five levels of n_2 (6, 7, 8, 9 and 10 days). The lines on each graph reflect the change in power for increasing n_1 (4, 6, 8 and 10 moments per day) for the different levels of n_3 (10, 20, 30, 40, 50 and 60 subjects). The most apparent feature of the graphs is the high level of power for larger levels of n_3; even at the lowest combination of number of days and moments, 40 participants provide sufficient power to detect an effect size of 0.1 $(1 - \gamma = 0.86)$, and increasing the number of moments to 6 per day or increasing the number of days from 6 to 7 provides adequate power for as few as 30 participants, $(1 - \gamma = 0.93)$ and $(1 - \gamma = 0.80)$ respectively.

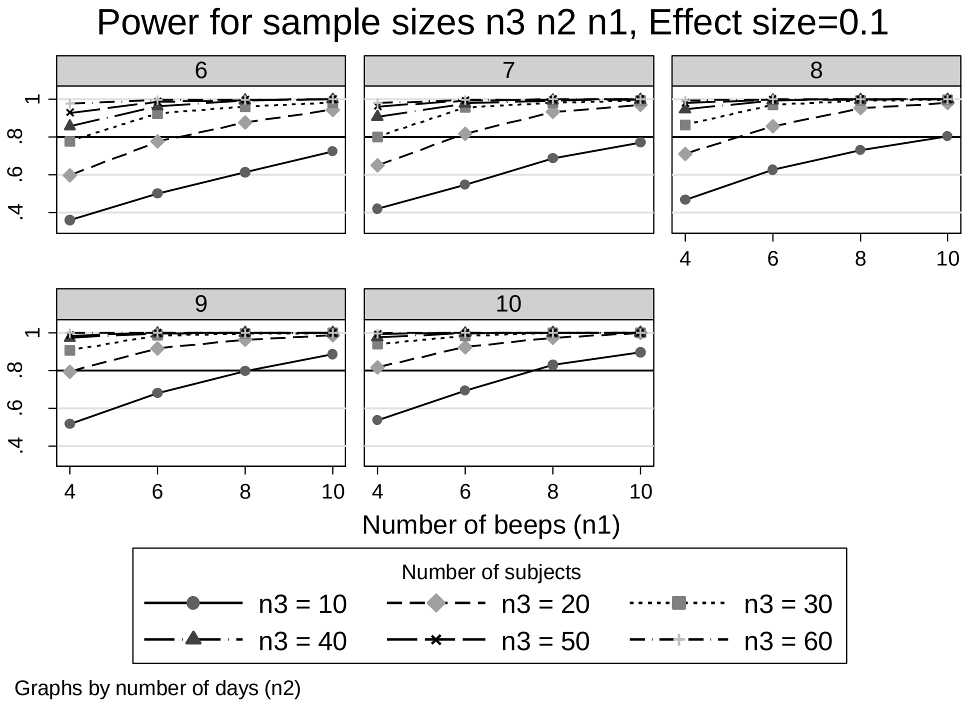


Figure 7:1 Power for association with effect size 0.1 and small predictor and model variances – Association model.

Also clearly depicted are the similar profiles across the graphs, indicating that varying the number of days is of little consequence when powering for a level 1 association. The only significant change here is four moments per day can provide adequate power with only 20 participants when increasing the number of days to nine ($1 - \gamma$ = 0.79) or 10 ($1 - \gamma$ = 0.82). The simulations demonstrate that increasing the number of participants or the number of measurements within days is more influential on power than increasing the number of days.

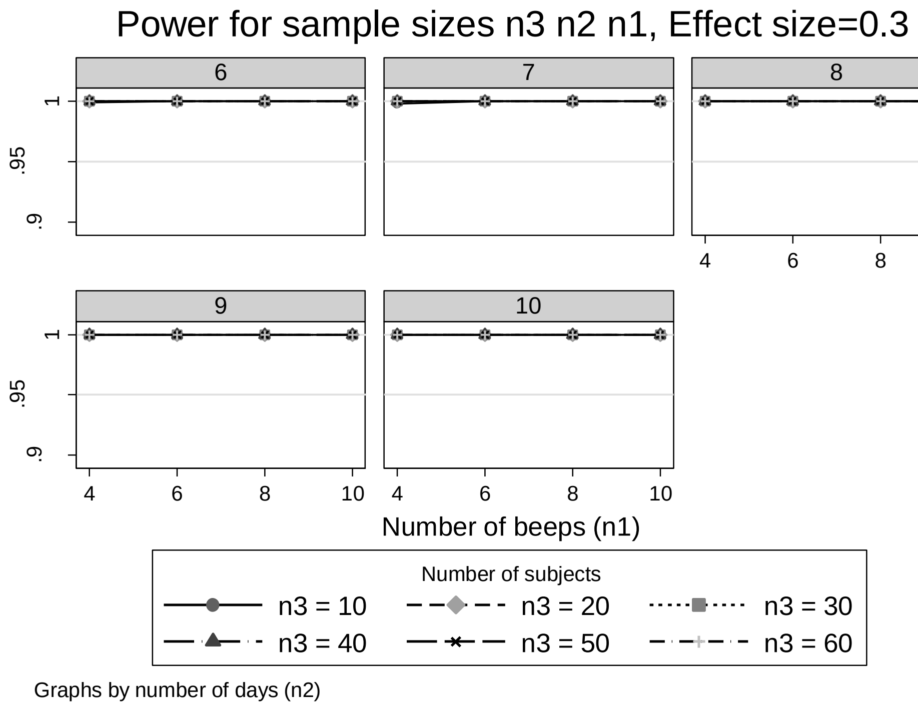


Figure 7:2 Power for association with effect size 0.3 with small predictor and model variances – Association model

Figure 7:2 presents the power graphs for a 'medium' effect size of 0.3 where both the model variances and predictor variances are fixed at the lowest level of 0.26. High power is demonstrated for all sample sizes in this example, even the smallest combination of n having near 100% power ($1 - \gamma$ = 0.999).

As such high power was achieved with a medium effect size, to efficiently utilize time simulations with a 'large' effect size of 0.5 were not carried out.

7.4.2.1.2 Results of varying model variances

To investigate how power changes when variances estimates are unknown a priori, model variances were varied from 'small' to 'medium' and 'large' as defined above in Section 7.4.1. Firstly, the level 3 variance σ_u^2 was increased to 2.34, the largest variance estimate, for a small effect size and holding the remaining variance estimates 'small'. The results of this increase can be seen in Figure 7:3.

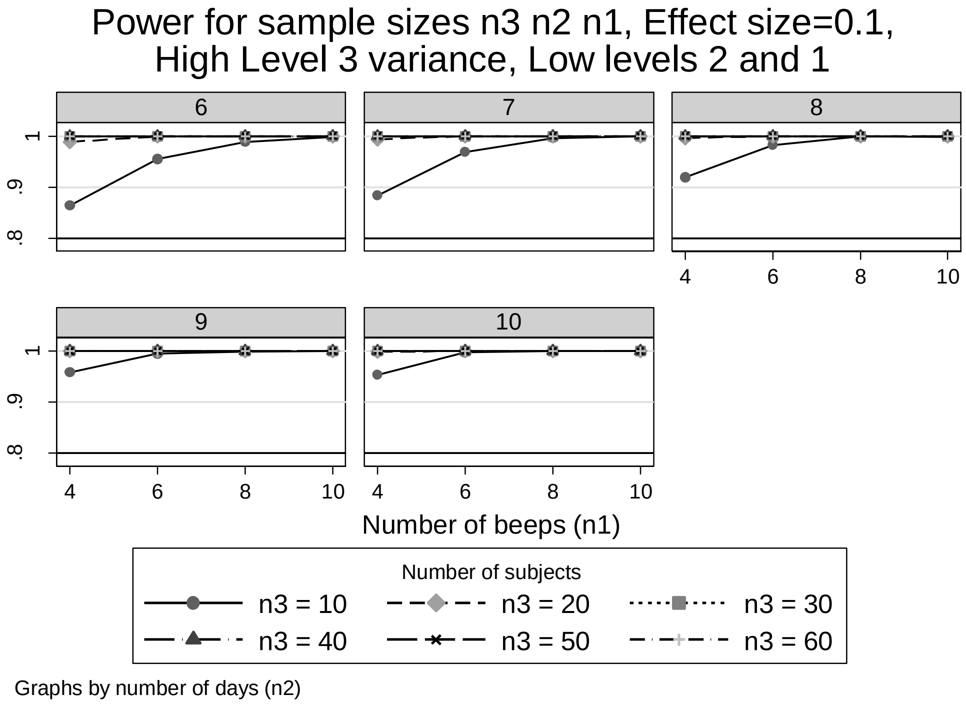


Figure 7:3 Power for effect size 0.1, large level 3 model variance, all remaining variances small – Association model

Comparing the power graphs of Figure 7:3 to Figure 7:1, increasing level 3 variance to $\sigma_u^2 = 2.34$ while holding all other model and predictor variances equal to 0.26 dramatically increases power for all levels of n. Power is now over 80% for all combinations of sample size.

Increasing level 2 variation to 2.34 while holding the other parameter estimates 'small' resulted in the power estimates depicted in Figure 7:4.

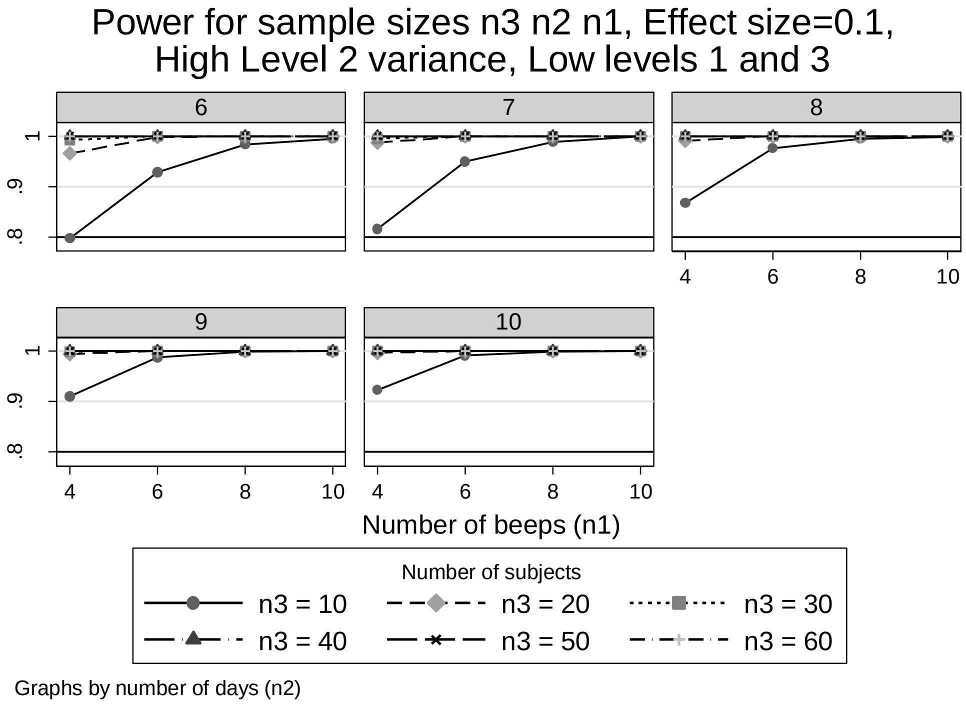


Figure 7:4 Power for effect size 0.1, large level 2 model variance, all remaining variances small – Association model

As with the level 3 variance, increasing model variation at level 2 resulted in an increase of power for all sample sizes. With even the very lowest combination of n (4 measurements a day for 6 days for 10 people) providing sufficient power to detect a small effect size $(1-\gamma = 0.798)$. Although their formula were derived for a binary predictor, the design effect given by Cunningham and Johnson (2012) for level 1 randomisation supports these results: in equation (36) if either the level 3 variance or level 2 variance contributing to the numerator of $\rho_1$increases, the design effect reduces and so power increases.

Finally, level 1 variance was increased. Figure 7:5 illustrates the power for varying sample sizes when the residual variance was increased to the highest variance estimate of 2.34. For each graph, power has decreased compared to Figure 7:1; a similar profile observed for six days originally is only achieved with nine days of observation in this scenario. Adequate power is achievable with as few as 30 participants but requires either a greater number of measurements per day over a shorter study period or with more days of observation when sampling fewer times a day. With large level 1 variation 80% power is only reached with 20 participants when taking 10 measurements a day for seven days or eight measurements a day for eight days.

A summary of the effect of varying model variances at each level are presented in Table 7:5.

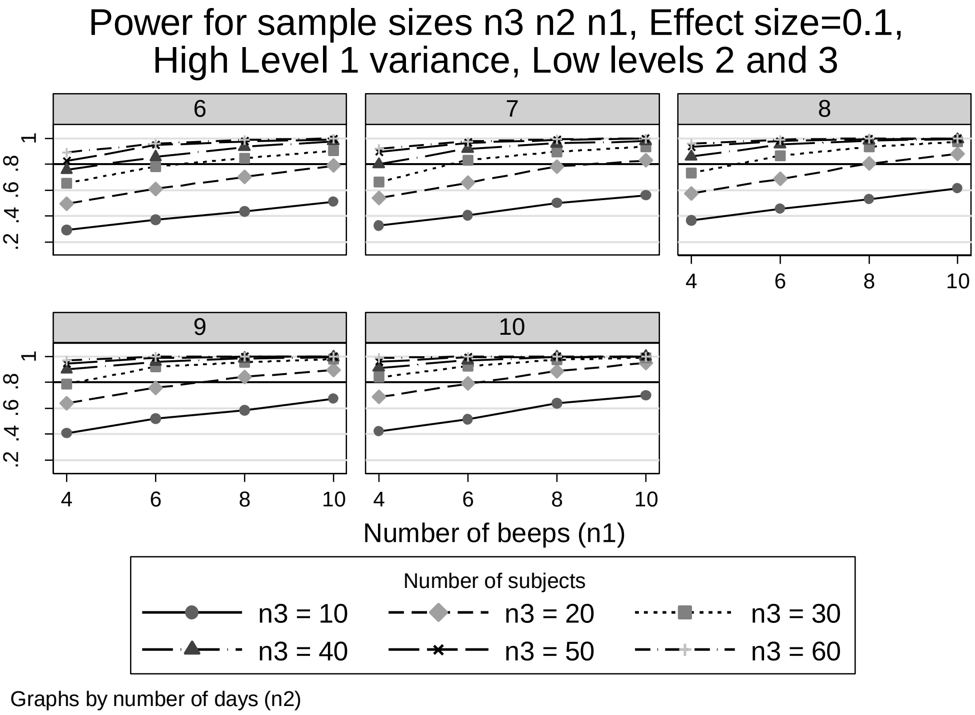


Figure 7:5 Power for effect size 0.1, large residual variance, all remaining variances small – Association model

Similar patterns in power were observed with the medium variance estimates, suggesting it is the proportion of variance at each level, rather than the magnitude, which influences power.

7.4.2.1.3 Results of varying predictor variances

Predictor variances at levels 3, 2 and 1 were also varied between high ($\sigma^2_{x_{L\#}} = 2.34$), medium ($\sigma^2_{x_{L\#}} = 0.56$) and low ($\sigma^2_{x_{L\#}} = 0.26$) while holding all other variances estimates constant at 0.26. The results are presented in Figure 7:6.

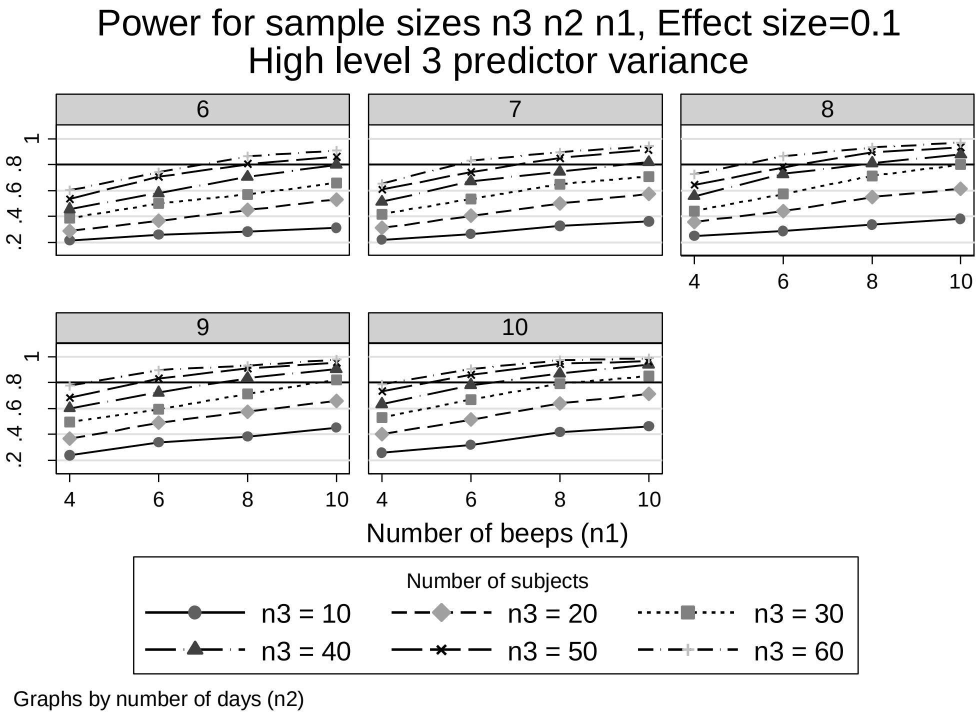


Figure 7:6 Power for effect size 0.1, large level 3 predictor variance, all remaining variances small – Association model

For an effect size of 0.1, increasing the level 3 predictor variance to 2.34 resulted in the power estimates shown in Figure 7:6. Comparing these graphs to Figure 7:1, power has been reduced for all n; with the maximum number of measurement per day, 80% power is only achievable with a week of observation for 40 participants or more. Unlike with low predictor variances, adequate power is unobtainable with 20 participants and only possible for 30 participants with more intensive sampling for longer study periods (10 measurements a day for nine days $(1 - \gamma = 0.82)$.

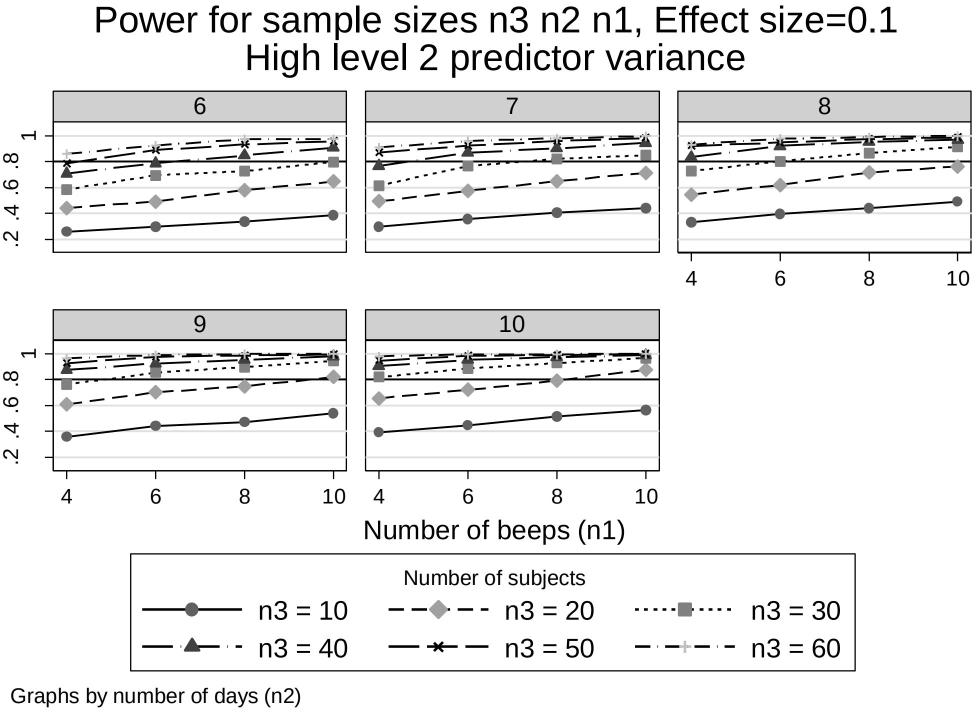


Figure 7:7 Power for effect size 0.1, large level 2 predictor variance, all remaining variances small – Association model

When the level 2 predictor variance is increased to 2.34 power is again reduced, although to a lesser extent. For a week of observation, eight measurements a day provide 82% power for detecting an effect size of 0.1 with 30 participants, or with 10 measurements a day only six days are required for approximately 80% power ($1 - \gamma = 0.797$).

Finally, the level 1 predictor variance was increased to 2.34. The results of these simulations are presented in Figure 7:8. These graphs show markedly different results to the previous two scenarios, with power significantly increased for all level of n. Under these conditions only a minimum sample size of six measurements a day is required for 10 subjects at any sampling length for adequate power. For 20 participants or more any sampling scheme or length of study will result in sufficient power to detect a small effect size.

As with the model variances, medium predictor variances had similar effects on power to the large variances, again indicating the proportion of variance at each level influences power more than the amount of variation.

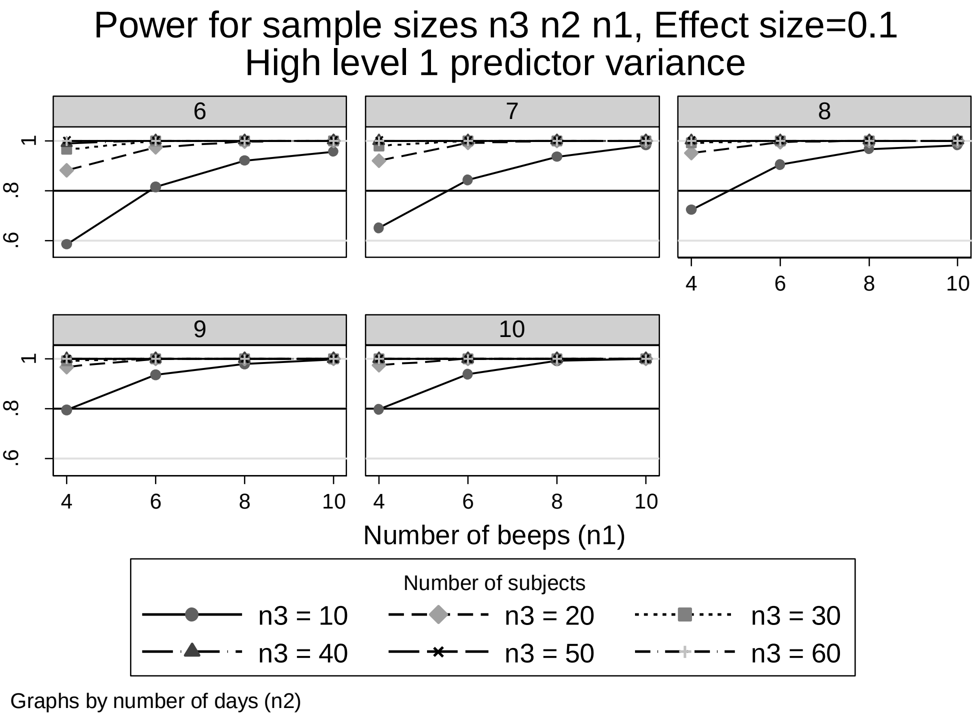


Figure 7:8 Power for effect size 0.1, large level 1 predictor variance, all remaining variances small – Association model

The results of varying model and predictor variances are summarized in Table 7:5.

Effect size	Predictor variance			Model variances			
r	$\sigma^2_{x_{L3}}$	$\sigma^2_{x_{L2}}$	$\sigma^2_{x_{L1}}$	σ^2_u	σ^2_v	σ^2_e	Summary
0.1	Equal variance			Highest proportion at level 3			Greatly increases power
				Highest proportion at level 2			Greatly increases power
				Highest proportion at level 1			Slightly reduces power
0.1	Highest proportion at level 3			Equal variance			Reduces power
	Highest proportion at level 2						Reduces power
	Highest proportion at level 1						Increases power

Table 7:5 Results of varying model and predictor variances for simple association model with fixed effect size

7.4.2.1.4 Missing data

The simulations thus far have assumed fully observed data. As this may not be the case, the results of this section describe how power varies under different combinations of sample sizes for varying amounts of missing data.

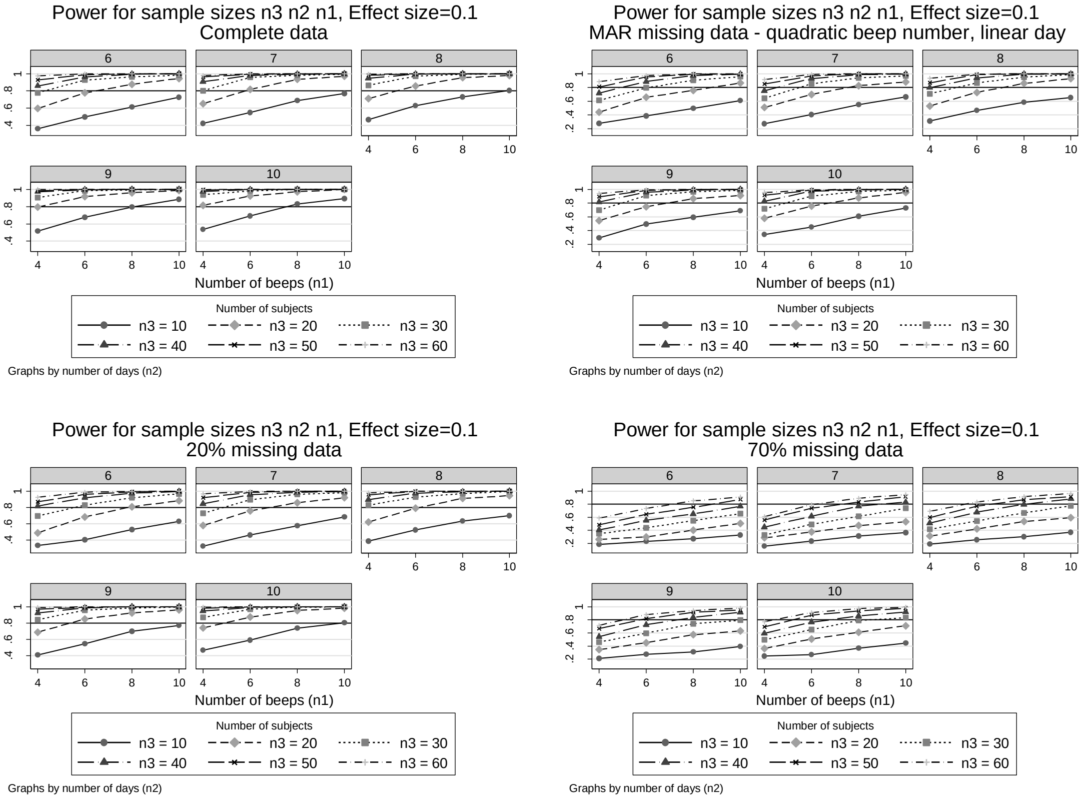


Figure 7:9 Power for varying levels of missing data. Effect size 0.1, small model and predictor variances – Association model

The graphs for 20% missing data and missing data patterns over time based on the recovery data show very similar results. Overall, the proportion of missing data was similar in the two samples, with time varying nonresponse resulting in an average of 34% missing data. In both of these scenarios power is reduced slightly compared to the complete sample for lower numbers of level 3 units, while in samples with 50 or 60 participants power remains high with six or more observations a day. In both scenarios approximately 80% power is achieved with 30 participants when sampling at least six times a day over six days. Similar levels of power are only possible with fewer participants when sampling eight times a day for six days under 20% missing data (20 participants) and ten times a day for ten days with ten subjects. Adequate power is not achieved for any combination of n_1 and n_2 for ten subjects under the missing data pattern observed in the recovery study data, but is possible with 20 participants when sampling ten times a day for six days or eight times a day for seven days.

A much more marked difference in power occurs in the extreme case of 70% missing data. Power is dramatically reduced to detect an effect size of 0.1 across all combinations of n. In this scenario a minimum of 80% power is first achieved sampling eight times a day for six days with 60 participants. Adequate power with fewer participants is only possible during a week of observation when sampling ten times a day, though a minimum of 40 subjects are required.

7.4.2.2 Group difference model

This section presents the results of the power simulations for a model investigating a level 3, or subject level, group difference in means. Power graphs will again display the results of varying the effect size and the changes to power when varying the model variances σ_u^2, σ_v^2 and σ_e^2.

7.4.2.2.1 Results of varying effect sizes

Figure 7:10 illustrates the first scenario where variance estimates are set to low $(\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26)$ and power is estimated for Cohen's smallest effect size of 0.2. The probability of being in group 1 was fixed as 0.33.

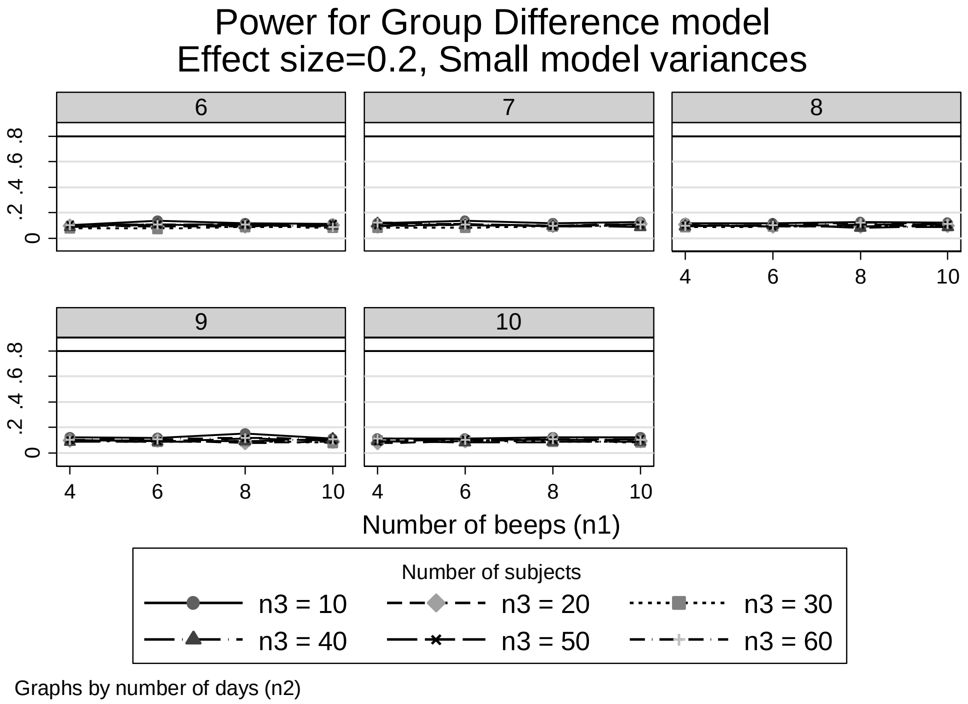


Figure 7:10 Power for effect size 0.2, small model variances – Group difference model

For all n power is very low, fluctuating between 8% $(n_3 = 30, n_2 = 6, n_1 = 6)$ and never exceeding 15% $(n_3 = 10, n_2 = 9, n_1 = 8)$.

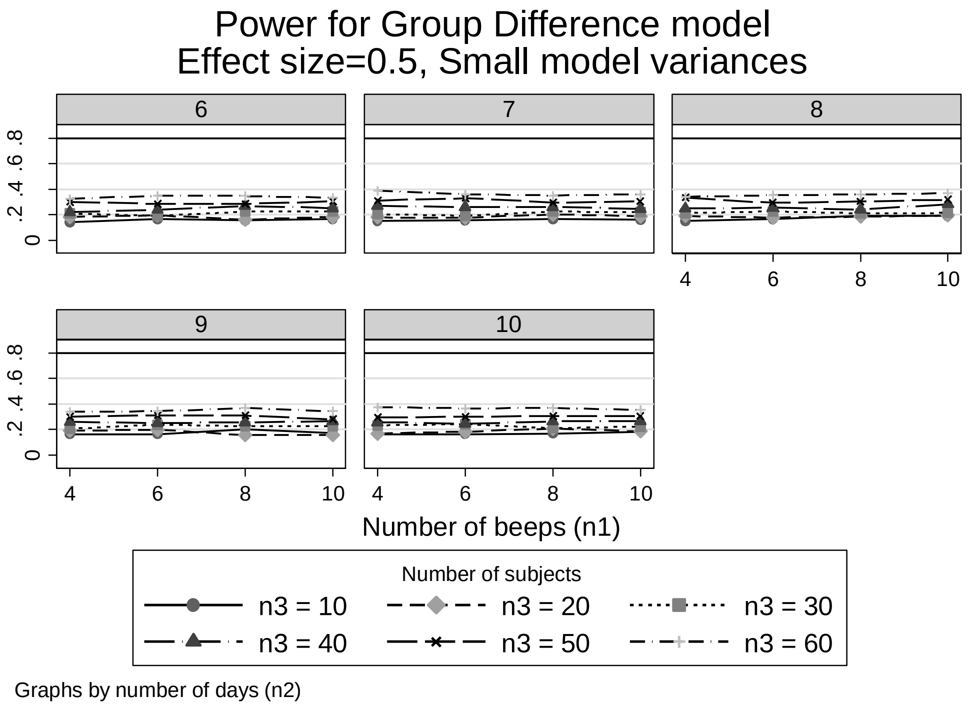


Figure 7:11 Power for effect size 0.5, small model variances - Group difference model

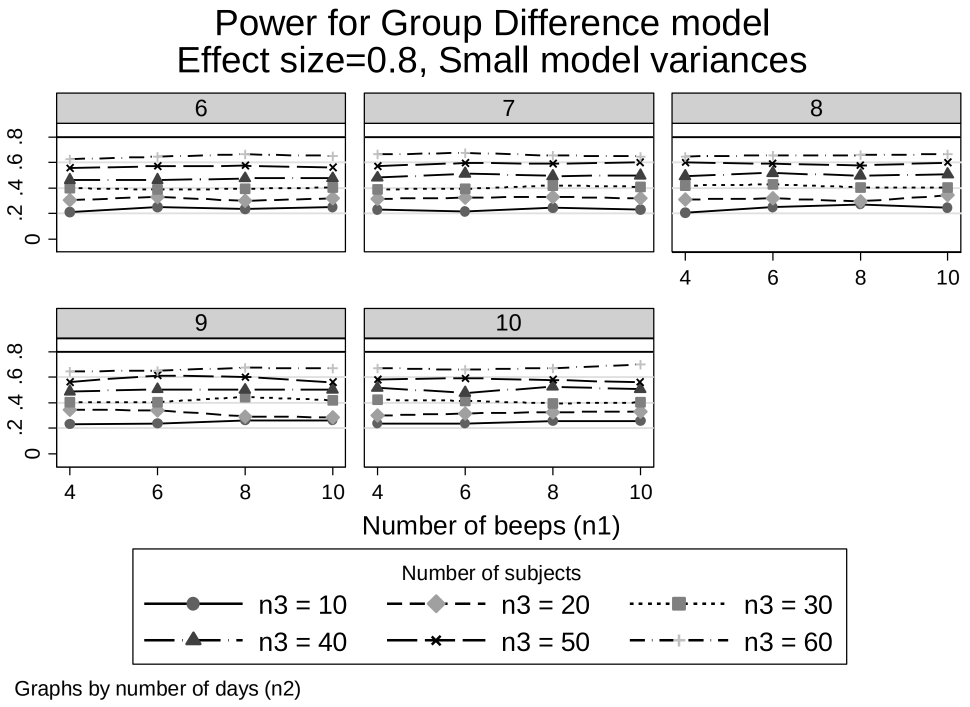


Figure 7:12 Power for effect size 0.8, small model variances - Group difference model

Sufficient power is not achieved for any of the three effect sizes for any combination of n. In all three scenarios a similar power profile is observed: flat within-day lines with very little variation between days, indicating neither the number of measurements a day nor the duration of the study influences the power to detect a subject-level effect. Only the number of participants seemingly effects power, with power increasing for a higher number of participants for larger effect sizes.

The number of subjects was increased above the originally specified maximum of 60 to establish how many subjects would be required to detect a large effect size with the current sampling schemes. Figure 7:13 displays the power graphs for $n_3 = 70, 80, 90$ and 100. With these sample sizes the smallest combination for which 80% power is achievable is with 90 participants, sampled eight times a day for six days $(1 - \gamma = 0.803)$. For a week of observation any sampling scheme produces approximately 80% power for 90 subjects and is only possible with 80 subjects when monitoring for 10 days.

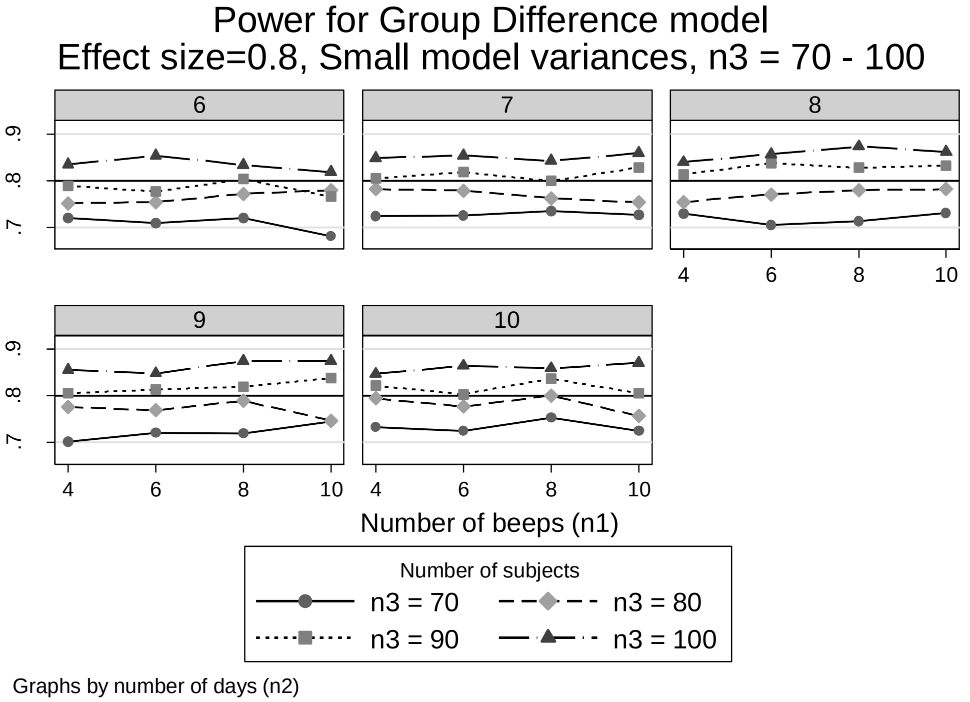


Figure 7:13 Power for effect size 0.8, small model variances, larger n3 - Group difference model

7.4.2.2.2 Results of varying model variances

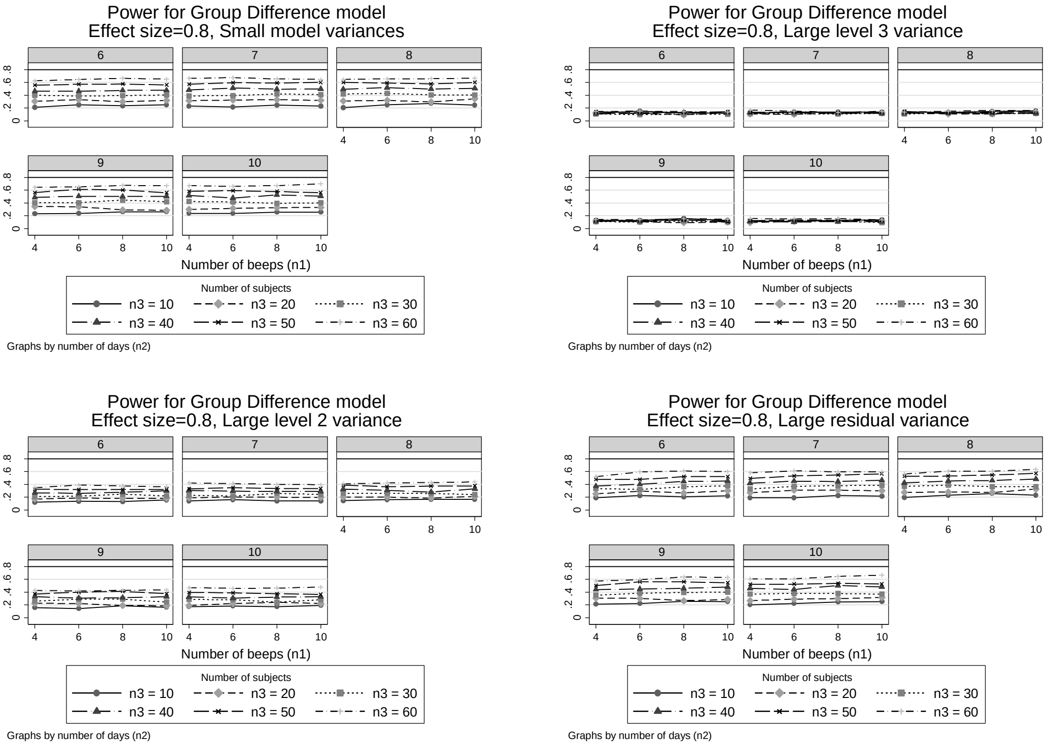


Figure 7:14 Varying model variances for effect size 0.8 – Group difference model

The previous simulations were based on small model variances with equal variation at each level. Varying the model variances σ_u^2, σ_v^2 and σ_e^2 in the group difference model resulted in the power graphs of Figure 7:14. Each graph demonstrates the effect of increasing the model variance from small ($\sigma^2 = 0.26$) to large ($\sigma^2 = 2.34$) at each level while holding the other two constant. The first graph, for reference, depicts the power for a model with effect size 0.8 where all variances are set to the lowest variance.

Increasing between-subject model variance has the greatest effect on power, reducing estimates to less than 20% for all n. Increasing level 2 model variation also reduces power, though to a lesser extent. Little change is witnessed after increasing residual variation. These results are supported by the design effect given by Cunningham and Johnson (2012) for randomisation at level 3 (equation (35)): increasing the level 3 or level 2 model variance in this formula gives a larger design effect and thus results in less power.

7.4.2.2.3 Missing data

The power graphs of Figure 7:15 demonstrate the change in power for each level of n when diary data is not complete. As with the association model, the proportion of missing diaries was varied between 20%, 70% and under the missing data pattern observed in the recovery data. The top left panel illustrates power for complete data under the same fata generating model, for reference. Comparing this to the remaining three, the introduction of moment nonresponse results in little change to the empirical power estimates.

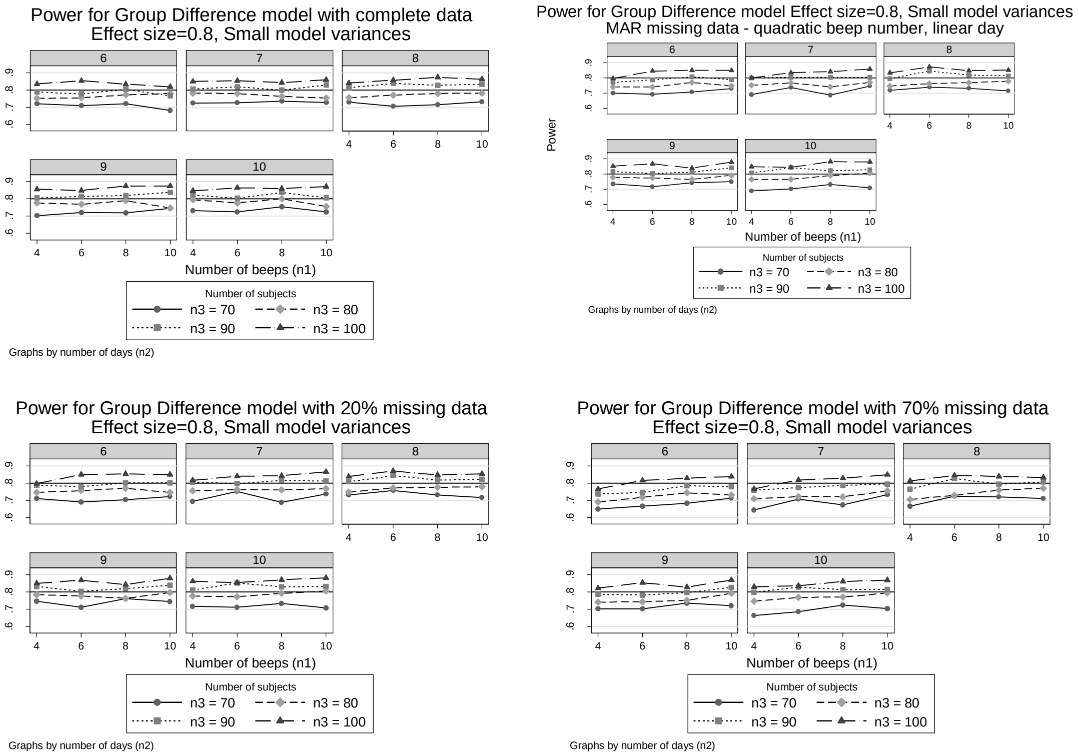


Figure 7:15 Power under difference proportions of missing diary data - Group difference model

Power was also simulated when the predictor x is subject to missing data. The graphs of Figure 7:16 show the effect of 10%, 30% and 50% missing data in the level 3 grouping variable in addition to 20% missing diary data. The graphs demonstrate the greater effect of missing data at level 3 has on power than moment nonresponse.

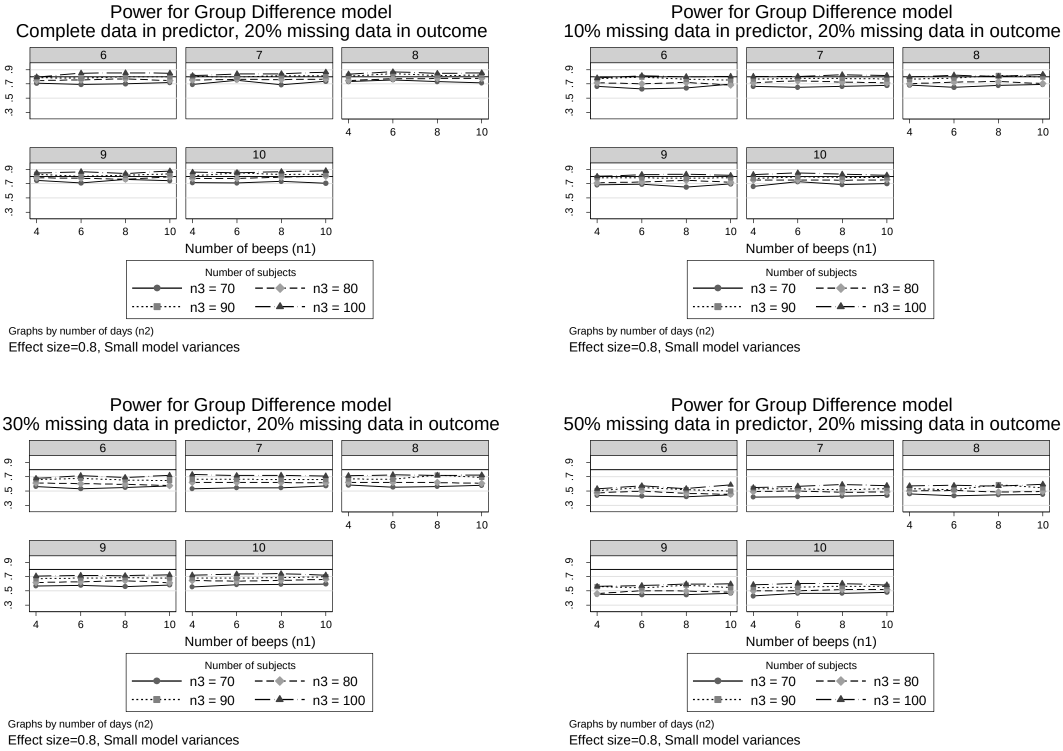


Figure 7:16 Power with missing data in both predictor and outcome - Group difference model

7.5 SUMMARY

This chapter has presented closed form equations for power calculations in one-, two- and three-level models as well as code for empirical power using Monte Carlo simulation. The methodological literature for power in three-level models is growing, however, the closed form power calculations and design effects, though developed for three-level random intercept and some random slope models, are unsuitable for many ESM studies as they have been derived in a trials framework and require randomisation at one of the three levels. As much of the ESM literature is observational, power calculations by simulation is more appropriate.

For a simple association model the simulations demonstrated that even for Cohen's small effect size 80% power is achievable with as few as 20 subjects under certain designs. Increasing model variances at level 3 or level 2 dramatically increases power to detect this effect. Similarly, increasing the level 1 variance of the predictor increases power, while increasing variation at level 3 or 2 of the predictor reduces power. Rather than the size of the variance effecting power it is the proportion of total variation which commands the change; 'medium' levels of variance producing similar patterns. These results suggest that if parameter estimates are unavailable a priori, conservative power estimates can be achieved by specifying the majority of model variance at level 1 while minimising momentary level variation in the predictor in the data generating model.

For group differences, the power simulations showed that when examining effects measured at higher levels many more subjects are required for adequate power. In addition, power reduces dramatically when the largest proportion of variance is between-subjects. When subject level differences are of interest and a priori estimates are unknown, conservative power calculations should be based on smaller effect sizes and larger level 3 variance estimates.

Finally, in both cases missing data was seen to influence power. While nonresponse is more often published than variance estimates, the reporting is often unclear and the proportion of missing data varies greatly from one study to another (see Chapter 4). Furthermore, the reported figures often only pertain to the diary data and not the amount of completed baseline data, for example. As such, basing missing data levels on previous research may not adequately reflect the amount of missing data that will arise in the proposed study. Care should be taken to both estimate the likely proportion of nonresponse in the particular study population and reduce missing data as part of the study design, particularly with respect to subject-level data.

8 DISCUSSION

This thesis aimed to develop methods to improve the design, reporting, analysis and exploration of ESM studies, building on areas identified as weak following a methodological review of current ESM literature. Improvements to design were pursued through the exploration of power for three-level data, where optimal combinations of sample sizes at each level were investigated for varying effect sizes, and the impact of uncertain a priori estimates on power was tested. Improved methods of reporting were considered for missing data to highlight how nonresponse in ESM is a more complex issue than simple adherence to protocol, and in understanding the missing data mechanisms, model assumptions can be satisfied. Extensions to current analysis practices in ESM were developed to answer questions relating to change and variability at level 1, identifying methodological weaknesses and proposing solutions. Finally, models for exploring patterns in ESM data were developed to allow for a detailed examination of the variation in momentary measures both with-day and across all days of observation.

The methods identified and developed for these areas will be discussed in detail and novel contributions of this thesis will be highlighted. The strengths and limitations of the methods will be examined in terms of their application to ESM data.

8.1 DESIGN - POWER AND SAMPLE SIZE

Two options for calculating sample sizes for ESM studies were presented in Chapter 7: sample size formula and empirical power using Monte Carlo simulation. Although closed form solutions are available for some three-level designs, many ESM research questions require models more complicated than the current literature provides for. The chapter aimed to develop code for simulating power in Stata for application in an ESM setting.

As sample size for ESM studies constitutes three levels: the number of subjects, the number of days observed and the number of measurements per day, power will depend on the combination of all three n. As such, simply increasing the number of participants is not the only way to increase power. An optimal balance between number of subjects and sampling scheme will ensure that there is sufficient power to detect an effect whilst minimising any stress due to the intensive nature of the methodology and ensuring cost effectiveness of the study. The power simulations can thus be used to find such a combination, specific to each ESM study, the resulting graphs allowing for an easy comparison of power for all combinations of n. The simulations for the two example research questions reinforced the results of other multilevel simulation studies that additional units at the highest level have the greatest influence on power. For ESM studies however, the number of subjects may be limited; these graphs demonstrate the benefits of using Monte Carlo simulation to calculate power, where increasing either the number of days of observation or the number of measurements each day can be compared to achieve sufficient power. For example, if ESM is being used to study a specific small population and is only likely to recruit 20 participants, the power graphs can be used, fixing $n_3 = 20$, to aid in the choice of the sampling scheme. Alternatively, for larger available populations, if participant burden needs to be minimised, the number of days and the number of subjects can be compared for $n_1 = 4$. Participant burden and reactivity bias as a result is a serious concern of ESM ((Larson R and Csikszentmihalyi 1983; Delespaul 1995)). Designs that aim to minimise this should clearly be sought when possible.

The main barrier to powering ESM studies is the lack of available a priori information required for computations. Estimates of expected effect sizes as well as predictor variances and model variances for all three levels of the data are necessary for empirical power calculations. However, it is unlikely that researchers will have prior knowledge of how predictor and response variables will vary between- and within-subjects before collecting their data. ESM is implemented expecting moment to moment variability in measures but quantifying this a priori may be unachievable, even with expert opinion, as published studies often lack this information. If prior data are available, estimates may vary by population (for example treatment effects of depression in clinical and non-clinical populations) and so may not be transferable for the purpose of sample size calculations for future studies. The Monte Carlo simulation results presented how sensitive power is to these estimates and how estimates can be chosen prospectively to obtain conservative power for optimal sample size combinations. For both the research questions of momentary-level association and subject-level group differences, power estimates were found to be heavily reliant on effect estimates and so in both cases a priori effect sizes should not be over-estimated. When variance estimates are unknown, conservative power estimates for momentary associations can be obtained by specifying the highest proportion of model variance at level 1, while minimising the variance of the predictor at this level. For subject-level group differences many more subjects are required for acceptable power than for level 1 associations, with power highly dependent on model variances. Conservative power estimates should be based on a higher proportion of variance at level 3.

Varying the effect size demonstrated how influential this parameter is on power for level 1 associations: the power to detect a 'medium' effect size of 0.3, as suggested by (Cohen 1988), was achievable with even the lowest combination of sample sizes; greater variation in power was found with a 'small' effect size of 0.1. Cohen proposed these categorisations with the caveat that they are by no means definitive or applicable to all scenarios. It remains unclear whether these effect sizes are appropriate references for ESM data. Unfortunately, collecting effect sizes was out of the scope of the systematic review discussed in Chapter 3 and there is currently very little literature reviewing ESM studies to which effect sizes can be compared for this population. In the recovery study results presented in Chapter 1.3, rescaling the regression coefficients as effect sizes, absolute concurrent effect sizes ranged between 0.10 and 0.48. In contrast, the second motivating example (Tyler, Jones et al. 2015) much smaller level associations were found between mood and symptoms scores, with effect sizes ranging from 0.06 to 0.14. Care should be taken when considering clinically relevant effect sizes for powering ESM studies, resisting the urge to simply use Cohen's standardized effect sizes, as powering for too high a value may result in insufficient power to detect the true effect size. Moreover, any statistically significant estimates found as a result might overestimate the effect size or show it to be in the wrong direction (Gelman and Carlin 2014).

The simulation results of varying model variances for both scenarios were compared to how power would change using the design effect equations of Cunningham and Johnson (2012). Though these formulae corroborated the findings of the simulations, the comparison is perhaps not appropriate for the level 1 association example. Cunningham's design effect was developed for randomisation at level 1, thus a binary predictor with no higher levels of variation. Though the formula might make an adequate comparison for the continuous predictor of the simulation scenario, this momentary ESM measure has variation at all three levels of the data which is not accounted for when deriving the design effect. An alternative explanation of the influence of the model and predictor variance estimates on power can be considered by how the regression coefficient β, and in turn dependent variable y, was calculated. The effect size was transformed into a regression coefficient used to generate y through scaling by the ratio of the model and predictor variances. Increasing any part of the model variance (that is, between-subject variance, between-day variance or residual variance) will thus increase β if the components of the predictor variance are held constant. As such, increasing model variances will lead to greater power for smaller sample sizes as sufficient power to detect a large effect is achieved with smaller n. Conversely, increasing any of the predictor variances while holding variance components of the model constant will decrease the size of the regression coefficient β. In these circumstances a greater sample size will be required to detect β and thus the power graphs will show a reduction in power.

Based on this rationale, an increase in any component of model variance should increase power. However, this was not found to be the case: increasing residual model variance resulted in a reduction in power. In addition to the effect residual variance has on β, increasing the residual variance results in greater variation in y at level 1, and thus the standard errors for level 1 predictors will be larger. Larger standard errors lead to smaller test statistics which means the null hypothesis is less likely to be rejected and thus power decreases.

8.1.1 Implications of missing data

The power simulations demonstrated how varying levels of missing data can affect the power to detect a momentary effect, and how sample sizes chosen on the basis of complete data may be underpowered in the presence of moment nonresponse. The extent to which this will influence power depends on the quantity of missing data and so alongside realistic estimates of momentary nonresponse when calculating prospective power, every effort should be taken to limit the amount of uncompleted diaries during the study.

Fitting a model with a subject-level predictor and a moment-level outcome, there are two potential sources of missing data: missing observations in x or y. In the association example both variables were measured in an ESM diary, so when defining missing data at the moment level, when x was missing so too was y. A level 3 variable, however, is only measured once per subject and thus if missing, no data from that participant can be used in the model. Missing data in the predictor of the group difference model, therefore, can result in a much larger reduction in overall sample size. One might assume baseline data to be complete in ESM studies as it would typically be completed with a researcher rather than independently as a self-completed form. However, in the recovery data example three subjects who went on to complete the ESM did not have baseline recovery scores. Underestimating the amount of subject-level missing data could result in significant loss in power and so if powering for complete baseline data it is imperative that this information is gathered for all participants.

The level 1 association analysis demonstrated that power to detect an association between two momentary-level measures is typically high, even in small sample sizes. The presence of missing data naturally reduces power and the impact of this loss of power is dependent on the proportion of uncompleted diaries. Although measures can be taken to prevent missing data during the study, it is important to note that it may also be introduced by design. The simple association example presented investigated power for a concurrent association where both predictor and outcome are measured at the same time point. If a diary is uncompleted then both variables are missing. If instead a lagged association is of interest, the predictor is entered into the model at moment $i - 1$. In this scenario the number of usable measurements per day automatically reduces: the first observation each day is dropped as there is no lag covariate measured. When taking 10 measurements a day, the effective sample size is reduced by 10% using a lagged analysis, with the proportion increasing, naturally, for less intensive sampling. Thus the impact of additional missing data may result in a substantial loss of power. This is compounded by the fact that momentary nonresponse effects the model twice in a lagged analysis as two time points are modelled simultaneously, thus an observed response for the predictor will be dropped when the outcome is missing and vice versa.

8.1.2 Adaptation to more complex scenarios

The simple association model was given as an example of how to use Monte Carlo simulation to inform sample sizes at levels 1, 2 and 3 for a prospective ESM study. However, this method is not limited to such simple models and can be adapted to accommodate any model of interest. The drawback of more complex models, however, is more a priori parameter estimates are required. For example, to estimate empirical power for varying sample sizes for time trend models, estimates of the random slope variances and covariances at each level will be required, in addition to the estimates required of the simple model. As an example, recall the across-day random slope model used to examine whether daily trends vary one day to the next (Model 2b of Table 5:3). The data generating model for this question requires five variance and two covariance estimates for the random effects, where random slopes are applied to daily trends at both the subject-level and day-level. If nonlinear terms are added, even more estimates will be required. The scarcity of published random effect variance estimates in the ESM literature in general, but particularly with more complex statistical models, will make estimating these parameters incredibly difficult. Of the papers reviewed in Chapter 3 for example, 14 applied multilevel random slope models to their data but only four presented the resulting variance estimates. These studies were based in completely different areas, with the random slopes used to test between-subject differences in smoking (Shadel, Martino et al. 2012), exercise (LePage, Price et al. 2012) (studies 1 and 2) and headaches (Kikuchi, Yoshiuchi et al. 2012). Whilst just a sample of published studies, this demonstrates that gathering conclusive evidence of random effect variances may prove difficult. The multilevel modelling literature suggests that, for randomised designs at least, power for a three-level design will decrease with increasing between-subject variation in slopes (Heo, Xue et al. 2013). Thus for conservative power estimates, prospective Monte Carlo power simulations should be based on larger variance estimates of random coefficients. The extent to which these estimates influence sample sizes can then be investigated as for the simple association model.

The simulations for the group difference model demonstrated that when the parameter of interest is measured at level 3 many more subjects are required to achieve 80% power compared to a level 1 association model, particularly for small effect sizes. While these simulations were based on a simple model, this result has implications for more complex research questions. A documented question of interest in ESM studies as found in Chapter 3 was that of moderation. This involves investigating how the relationship between two variables differs by a third. For example, the relationship between self-esteem and ESM recorded recovery may by moderated by whether the subjects reported feeling recovered at the start of the study. The analysis model now contains a cross-level interaction, with self-esteem measured at level 1 and the moderator measured at level 3. The results of the group difference model suggest that many more subjects will be required to detect the interaction than data may have been collected on if the study was only powered to detect the level 1 association. This reinforces the necessity of defining research questions during the design process rather than after data collection.

8.1.3 Limitations of work on sample size and power

Many sample size formula and simulation studies for multilevel data use the ICC, or proportion of variance at each level, rather than the individual variance terms in their power estimates. A decision was made to compare the magnitude of the variance at each level rather than specifying an ICC as the code created for simulation procedure required individual variance terms for each level to simulate the random effects. Varying the magnitude of variation is equivalent to an increase/decrease in the proportion of variance at each level; however, in retrospect it may have been more logical to choose variance estimates that summed to 1 to reflect the proportion of variance as to equate to ICCs from the literature. An argument against this parameterisation was in the choice of proportion to vary for the simulations: the proportion of variance at each level of the model differed greatly in the recovery data set to the bipolar data set, for example. As such a pragmatic choice of proportions could not be determined and the small/medium/large variance estimates based on the measurement scale were used instead, and while not derived from the data these values were similar in magnitude to the variance components of the recovery data ($\sigma_u^2 = 1.627, \sigma_v^2 = 0.217, \sigma_e^2 = 0.496$, Table 5:2).

A limitation of Monte Carlo simulation over closed form expressions is the time required to calculate power. While the association models were quicker to run than the group difference models, both took several hours to complete the 984 simulations on each permutation of n. More complex models require even longer, some time trend models taking days to complete. This may be in part due to the computational capabilities of the software chosen. Simulations were conducted in Stata as there are currently no programs written in this software for three-level power calculations. One of the aims of this piece of work was to write such a program that easily executed in this common statistical software. Although computationally more intensive, empirical power estimates via simulation provides a much more flexible framework for power in a ESM setting.

8.2 Reporting - Missing data

Missing data in ESM research can be defined in greater detail than just a single proportion relating to number of diaries completed. Nonresponse was found to occur at the item-level, moment-level and day-level, and summarising this information requires detailed consideration. Participant fatigue is a concern with this methodology, both in terms of individual diary completion and the repeated sampling method; graphs and tables were presented to demonstrate how missing data can be described at each level of the data, these simple summaries providing insight into patterns of nonresponse within a diary in addition to patterns both within-day and across the sampling period.

Item nonresponse is rarely mentioned in the literature, with the papers from the systematic review as well as the papers by Silvia, Kwapil et al. (2013) and Messiah, Grondin et al. (2011) referring only to moment nonresponse. Unlike with PDAs and mobile phone applications where programming can ensure questions cannot be skipped, paper diary methods are susceptible to item nonresponse where selective completion of items is possible. Though perhaps an antiquated issue with the recent developments in mobile technology, paper diaries were used in approximately 30% of the 74 reviewed papers of Chapter 3 which suggests it is still a popular data collection method. Indeed, there are certain scenarios where it is necessary to use paper diaries, such as in prisons or inpatient wards. Although electronic data collection is becoming more accessible, it can be an expensive option where smartphones need to be provided or custom software is developed. As such, paper diaries are a convenient option for studies with a more limited budget such as student-led projects. Furthermore, while digital data collection methods may eliminate sporadic item completion, they can still be left unfinished partway through the questionnaire resulting in missing data for the remaining items.

Data collection methods that restrict the ability to skip items, however, are not a faultless solution to item nonresponse. Requiring all items to be completed may lead to non-informative responses; mindlessly completing items without regard to the question. This may present in uniform completion of scale extremes, consistently scoring 7, for example, on 7 point Likert scales. This problem is not limited to smartphone collection, as it is easily possible with the traditional paper diary method. The interpretation of such flat response data will be reliant on the researcher examining variation in measures, either graphically or model based, and working with the participant after the diaries are returned to assess whether these values were truly representative of their experience. One benefit of the smartphone technology is that data are uploaded to external servers either in real time or at the end of each day and researchers can potentially review participants' data during the data collection period for excessive missing data or for responses flattening out. This would enable intervention to occur during the study to assess whether the data are representative of the respondent's current state or if they are finding the sampling procedure too intensive, potentially reducing the amount of missing data. The practicalities of this, however, are questionable for large scale studies.

Moment nonresponse was found to be the typically reported figure in current ESM literature, where studies described the proportion of completed prompts as a measure of adherence to protocol. The definition of a 'completed' prompt however, is rarely discussed. As described, selective item completion is possible in many ESM data collection methods, and so the interpretation of these adherence figures is questionable. Possible options for moment nonresponse were thus proposed as either missing all items in an ESM diary for one sampling moment or having missing data on the outcome measure at one sampling moment, thus rendering all items for that time point unusable. The latter argument holds for concurrent analysis, where outcome and covariates are measured at the same time point, however, a lagged analysis would still be able to include covariates at moment $i - 1$ when the outcome is missing in this diary. As such, it may be more useful to present only completely missing diary data when summarising moment nonresponse. In either case, a clear description of what is considered a 'complete' diary should be provided when summarizing adherence to protocol or nonresponse rates to enable the reader to understand the assumptions being made.

In many of the papers reviewed, a largely uncited cut-off of approximately 33% complete data was required to be considered a valid response, with those individuals completing less than a third of signalled prompts excluded from the analysis. Delespaul (1995) is most often attributed to this cut-off, however the book provides no justification for this value. Using maximum likelihood estimation of multilevel modelling there is no minimum requirement for completed observations as all available information will contribute to analysis. Consequently, this seemingly arbitrary figure of 33% is not enforced from any statistical necessity, instead it is likely that this is used as an indicator of validity of response where completing fewer than a third of responses suggests the items that have been completed are unlikely to be a true representation of a subject's current state. With this in mind, a minimum number of completed items per booklet might be a consideration when defining moment nonresponse, with sporadic completion treated as missing in order to avoid unrepresentative results. Discretion on the part of the researcher is required as to where to place the cut off to ensure responses are valid and representative of the true sample.

8.2.1 Predictors of nonresponse and missing data mechanisms

Further to the summarization of missing data, predictors of missing data were also explored in the recovery data. Although routinely conducted in other contexts such as randomised controlled trials, investigation of missing data mechanisms in ESM data appears to be lacking in practice. It was demonstrated how observed items can be used to predict missing items within an ESM questionnaire as well as how items from previous diaries can predict subsequent moment nonresponse. Predictors of nonresponse are not limited to previous time points though, forward lagged covariates could also easily be modelled, predicting missing data at the previous time point. Time trends in momentary nonresponse were explored following the suggestion of time dependent missingness from the descriptive statistics. Quadratic trends were observed within-day, where diaries were more likely to be missing both at the start and end of the day, and weekly trends indicate that there is a decline in response as the week progresses. This pattern was in contrast to what has been found in the limited literature investigating ESM nonresponse (Silvia, Kwapil et al. 2013), where missing diaries were less likely at the start and end of the day. Identifying these patterns can aid in the design of future studies, either through the choice of more appropriate sampling hours for certain populations or planning for researchers to check in with participants as the week progresses to prevent too much missing data.

In addition to the benefits for future studies, identifying predictors of nonresponse is key to satisfying the missing at random assumption made by maximum likelihood estimation. However, a drawback of satisfying this MAR assumption is that the interpretation of the coefficient of interest is now necessarily changed. In the example of the recovery data, both the linear and quadratic terms for beep number were found to predict missingness and so any subsequent analysis model should include these terms. Regressing recovery on self-esteem, for example, should now control for this quadratic time trend, altering the interpretation of the original research question as a proportion of the variation in the outcome can now be explained by the additional covariates. This can be avoided when using multiple imputation as the imputation model and analysis models can be separately specified, and thus the estimation and interpretation of the coefficients of the analysis model are unaffected by the variables in the imputation model.

In addition to maximum likelihood, multiple imputation was discussed as a method for addressing missing data in ESM studies. The dangers of misspecification in the imputation model were presented with reference to Black, Harel et al. (2011) who demonstrated how variance estimates can be biased when single level rather than multilevel imputation models are used. As software is currently unavailable for three-level multiple imputation, attempts to use two-level imputation on ESM data may result in biased estimates. Furthermore, maximum likelihood estimates were found to be unbiased for fixed effects even with large amounts of missing data (Black, Harel et al. 2011). Another scenario in which multiple imputation might be valuable is when there is missing data in baseline, rather than ESM, measured variables. If baseline variables such as age or gender are to be included as model covariates or as predictors of missing ESM data, any missing data will result in all ESM data being lost from the analysis model. Imputing these values will be much more straightforward as they come from level 3 variables and so a multilevel multiple imputation model is not required. An alternative method to multiple imputation in this case is to use a full information maximum likelihood approach in order to use all available data.

A potential limitation of maximum likelihood is the assumption that the missing data are missing at random. In ESM data it is not unreasonable to assume that a diary may be missing due to heightened symptoms, for example, and thus the missing data would be classed as MNAR. As non-ignorable missingness in empirically untestable, the argument in favour of more complex missing data methods must rest on how plausible it is that the probability the data are missing depends on the unobserved values. For ESM, diaries might be ignored during a hallucinatory event or those feeling depressed may be less inclined to complete a questionnaire. However, given the intensive sampling of the procedure and the expected correlation between items at different time points, ESM provides a lot of information which can be used to explain this missing data. It might be possible to predict this missingness from scores from the previous diary, for example, where rising symptom scores are followed by an uncompleted diary. However, if symptoms arise suddenly and are not captured by previous questionnaires then data will not be available to predict nonresponse. Low level symptoms resulting in nonresponse would also eliminate any chance of predicting missing data, resulting in non-ignorable missingness.

ESM benefits from typically having a debriefing session with participants at the end of the sampling period. During this session the researcher can discuss any sections of nonresponse and gain information as to why the diaries were left unanswered. This information can then be included in any statistical models to satisfy the MAR assumptions of maximum likelihood methods. The only data that would be unexplained would be from subjects who dropped out mid study or refused to meet with a researcher following their participation in the study. To capitalise on this post-data collection session researchers must be made aware of the assumptions being made regarding missing data, and the importance of gathering all potential information relating to nonresponse. Where follow up sessions are not possible and non-ignorable missingness likely, serious consideration is required to determine if ESM is the most appropriate data collection method for the study: a data collection method that is expected to consistently underreport symptoms is flawed and more appropriate alternatives should be adopted.

8.3 ANALYSIS - CHANGE MODELS

An in depth examination of the analysis of change was prompted by the unusual effect estimates obtained after fitting a change score model to the recovery data. This highlighted an important problem which can occur when using a change score as the dependent variable in a multilevel model: the strength of the concurrent versus lagged association of X on Y will influence the direction of association between the lagged X and a change in Y. This issue does not arise in concurrent change models, such as described in Mata, Thompson et al. (2012) where X is measured at moment i, but affects models interested in the change in Y following X where the lagged X is used as a covariate. To resolve this problem, a multilevel model with y_{ijk} as the dependent variable and the time lagged outcome $y_{i-1,jk}$ as a covariate was presented as an alternative to using a change score. Two methods were then discussed for overcoming the methodological issues arising from this model formulation: the joint modelling of the initial condition and subsequent outcomes and the first-difference procedure using instrumental variables. Both methods overcome the problem of endogeneity, the first-difference method removing the random effects from the model whilst preserving the within-subject interpretation of the fixed effects, while the alternative solved the problem by appropriately modelling the initial condition, allowing it to depend on unobserved heterogeneity. Both methods contain strengths and weaknesses in their approaches and while the techniques do not allow for a direct comparison of results, this section will aim to discuss the two methods in terms of their application to change models for intense longitudinal data.

Though rectifying the complication arising from including the lagged outcome as a covariate, it remains unclear whether the first-difference method, with its creation of a change score in the response, succumbs to the sign issue originally encountered. The results of the recovery data analysis using this method suggest the problem still exists as coefficients were in the direction originally observed using the change score model. The standard errors for the first-difference method coefficients were much larger, however, indicating effects in either direction were plausible. It was shown algebraically that the direction of the coefficient using the first difference method will depend on the strength of the lagged relationship between X and Y and the concurrent relationship. Though discussing the method in detail, Rabe-Hesketh and Skrondal (2012) do not comment on the possibility of this problem, nor does Steele (2014) in her application of the method.

Although the method requires the specification of an additional model for the initial outcome, the joint modelling the initial conditions results in easy to interpret model coefficients. The inference is directly comparable to that of the naïve lagged outcome model and standard random effects models: β_1 representing the effect of a unit increase of X in a change in Y, the sign of β_1 indicating the direction of the change in Y and its magnitude the strength of the relationship with X. The first-difference method, on the other hand, results in a change score in the fixed effects predicting a change score in the outcome. The interpretation of this coefficient is less straightforward as the change in X may be in either direction. However, the direction of the coefficient indicates the direction of the resulting change in Y: if $\beta > 0$ the change in Y is in the same direction as the change in X; conversely, if $\beta < 0$ the change in Y is in the opposite direction to the change in X. A benefit of the initial conditions method over the first-difference method is the between-subject and between-day variance estimates are still calculated. As the first-difference method removes the random effects, σ_u^2 and σ_v^2 are not estimated in the model for change. Although in many cases the fixed effect estimates are of primary importance, removal of the random effects eliminates the possibility of comparing predictors of change between participants and between days as random slopes are also removed.

8.3.1 LIMITATIONS OF THE INITIAL CONDITIONS METHOD

Kazemi and Crouchley (2006) state that when joint modelling to avoid endogeneity, the initial model needs to be correctly specified in order to produce valid estimates. One concern for using this method to analyse change in ESM data comes with the inability to include the lagged predictor in the initial model. To consider the effect of x on a subsequent change in y, x_{i-1} is included as a covariate. However, for the initial model where $i = 1$, $x_{i-1} = x_0$ is not observed. Two options were presented to overcome this issue: one can omit x as a covariate in the initial model or use x at moment i rather than $i - 1$. Omitting x when it is expected to predict the outcome will result in the initial model being misspecified. The argument made by Steele for using the initial model is to allow y to depend on the unobserved heterogeneity it would otherwise miss due to $y_1$only being used as a covariate in the model. In this regard, omitting x in the initial model does not detract from the purpose of using this method. However, Steele's example did not consider a lagged covariate and as such did not encounter this issue. In the extension of the method proposed for ESM data, in addition to $y_1$in the naïve model lacking influence from unobserved heterogeneity, it also lacks influence from x. Thus, Steele's argument requires x to be included in the initial model. Not including x would mean that any variation in y_1

due to this covariate would be falsely attributed to between-subject or between-day variation, and could lead to over inflated variance estimates of the random effects.

The alternative option presented was to include the concurrent predictor x_i in the initial model. Although this is obviously not the relationship the main model intends to examine, this may be a valid substitution for x_{i-1} when the two are expected to be substantially correlated. When this is the case, the lagged relationship is preserved in the concurrent relationship. As demonstrated in Chapter 6 however, the concurrent relationship between two variables is often stronger than the lagged relationship, and so using x_1 in place of x_0 in the initial equation may result in a stronger association estimated in the main equation with less variation in the random effects.

The basis for the initial conditions method is that the relationship between the current outcome and the lagged outcome can be expressed through that of the initial outcome with coefficient β_1^{i-1}, and that without this initial outcome being influenced by the unobserved heterogeneity that effects all other responses, this coefficient as well as the random effect variances will be biased. The extent of this bias, however, will depend on two things: the magnitude to which the correlation between observations diminishes with increasing lags (in effect the size of β_1) and the number of lags being calculated (i.e. n_1 for $i = 1, \dots, n_1$). For $|\beta_1| < 1$ increasing values of i will cause β_1^{i-1} to tend to zero, reducing the influence of the initial condition on y_{ijk}. Thus for larger n_1, the extent to which including a lagged outcome as a covariate biases the model decreases (Bhargava and Sargan 1983; Kazemi and Davies 2002). This asymptotic bias means that for short studies or for few repeated measures, endogeneity is of particular concern as the effect of the initial outcome will be strong when i is small. However, for ESM studies when n_1 can be large should the lagged outcome still be considered a problem? The results of the recovery data analysis unfortunately cannot be used to comment on this as it has been argued that these models may have biased estimates due to initial model misspecification. However, in Appendix 4: Predicting the analysis of this change model has been reproduced examining the effect of a concurrent predictor on outcome, controlling for previous outcome (as described by equation (22), Chapter 6). For this specification, x_i in the initial equation is observed and so the effect of number of measurements n_1 on bias can be isolated. These results suggest that bias from the initial condition does exist with 10 measurements a day (Naïve model: $\beta_1 = 0.283\ (0.021), \sigma_u^2 = 0.510\ (0.108), \sigma_v^2 = 0.058(0.015)$; IC model: $\beta_1 = 0.224\ (0.020), \sigma_u^2 = 0.631\ (0.127), \sigma_v^2 = 0.080\ (0.016)$), though the effect of self-esteem does not appear affected (Naïve model: $\beta_2 = 0.323\ (0.024)$; IC model: $\beta_2 = 0.333\ (0.024)$).

Finally, one should consider whether the question of change is even appropriate for ESM research due to unequally spaced measurements and the resulting interpretation of the fixed effects. Some sampling schemes, such as interval sampling, will produce equally spaces time points. However, Delespaul (1995) advises against such a regimented scheme as interval designs may result in participants becoming aware of the routine and thus changing their behaviour in anticipation of the alarm. This anticipation can potentially also affect mood and symptoms, increasing feelings of anxiety or stress for example, biasing the results of the study. The alternative signal based designs are either truly random prompts or pseudo random prompts where the alarm is emitted randomly within a fixed time interval. The latter is suggested as optimal – capturing moments across the whole sampling period while remaining unpredictable. Change models in a truly random design, where the length of time between prompts varies, are likely to produce meaningless results: the gradient on a linear trend varies between time points when intervals differ, resulting in large standard errors and uninterpretable effect sizes. This is compounded when different random patterns are generated for each participant. Pseudo random prompts, on the other hand, vary randomly within some fixed interval. So whilst the time between alerts does vary, it can be argued that for small enough fixed intervals this difference is negligible and one can assume the time between measurements is equal. This approximation is employed in regular longitudinal studies; six month follow-ups, for example, are rarely ever taken exactly six months from the start of the study.

8.3.2 FUTURE WORK ON CHANGE MODELS

This work has highlighted some interesting aspects of change models in an ESM setting. The initial conditions problem, particularly, could be pursued further in an attempt to quantify the bias introduced by including the lagged outcome as a covariate. Simulation could be used to investigate how the number of time points per day impacts on the bias of this variable to determine whether the joint modelling process if necessary for larger n_1.

Similarly, the extent of model misspecification for the initial equation could be investigated via simulation when using a lagged predictor. The bias introduced when either omitting x from this initial equation or using x_1 as a proxy for the unmeasured x_0 could be compared to that of simply fitting the naïve model, to ascertain whether joint modelling with a misspecified initial model actually induces more bias than simply ignoring the initial conditions problem altogether.

8.4 Investigation – Time trends using random slope models

The momentary variation in ESM measures is a potentially rich data source that can be investigated. Current methods for examining variation identified in the review underutilize intensively collected data, using methods that require aggregation to a higher level. These methods reduce moment to moment variation to one measure, ignoring potential trends at the momentary level. The random slope models presented aimed to investigate this momentary level variation more fully and tease out inference of time trends and variation where ‘time’ is measured at more than one level.

ESM presents an interesting area for the development of standard growth models; with three levels of variation and two distinct measures of time, this data structure requires care with regards to the choice of random effects and interpretation of random slopes. The explorations of random slope specifications yielded a total of 32 base models that can be extended to accommodate more complex trends in the fixed effects, through cross-level interactions and non-linear time trends, as well as complex random variation to study group differences. In practice not all of these models will be useful and model specification should be driven by the research question. However, the benefits of outlining all permutations of random effects allows for a greater understanding of how time trends can be studied at the momentary- and day-level, and how variation in these trend can differ both between- and within-subjects.

The random slope models presented in this chapter focused on explaining level 1 variation in terms of trends; the correlation between successive observations expressed through a fixed effect of time with random effects in place to measure between- and within-subject heterogeneity. An alternative approach is to allow this momentary level correlation to be absorbed in the residual variance. Multilevel models assume the residuals to be independent and identically distributed, conditional on the random effects, however, this assumption may be relaxed to allow for the correlation between successive time points to instead be expressed through an appropriate residual covariance structure. Autoregressive covariance structure can be implemented without the inclusion of random slopes to capture the correlation between observations, replacing a random slope model with a random intercept model with AR(1) residuals, for example. Used in conjunction with the random slope models, a non-independent covariance structure can be used to explain any additional correlation not accounted for by the time trends.

There is a trade-off between using a more complex random slope model to account for the momentary correlation and a comparatively simpler model where the variation is absorbed into the level 1 covariance matrix. Although the latter seemingly represents a more straightforward option, there are two serious points to consider. Firstly, as discussed in Chapter 2, many autoregressive covariance structures require equally spaced time points, which although applicable when beep number is used as a proxy for time, does not hold in many ESM settings where measurements are collected pseudo-randomly and thus will be unequally spaced. Moreover, commonly used structures such as AR(1) can impose strict conditions on the correlation between time points, and although more relaxed options are available, misspecification of covariance structure can lead to over-estimated random effect variances, leading to inaccurate inference (Ferron, Dailey et al. 2002; Sivo, Fan et al. 2005; Kwok, West et al. 2007). Secondly, expressing the level 1 correlation in terms of an autoregression parameter such as ρ from an AR(1) covariance structure, while providing a measure of variation, limits the discussion of variation to one question: how are successive observations related? Random slope models allow for a more detailed investigation of this at all three levels of data through linear and nonlinear trends, and with the addition of complex variation at level 1, can be an effective method for comparing group differences while addressing the popular question of 'fluctuation' of outcome.

Though providing a more informative picture of variation than autoregressive covariance structures, a limitation of the models presented was the relative simplicity of the nonlinear time function adopted. The intention of these models was not to perfectly describe underlying trends, but to instead offer options for investigation. More forms of describing nonlinear trends are applicable in this framework, for example higher order polynomials or exponential functions of time. Alternatively when greater flexibility is desired, piecewise linear functions (Bryk and Raudenbush 1992; Snijders and Bosker 1999) or spline functions can be adopted in which separate polynomials are fitted to time intervals (Pan and Goldstein 1998). Though the detail to which trends are studied is a personal choice, it is important to model nonlinear trends when they occur in the data. Bauer and Cai (2009) demonstrate how spurious random slopes and cross-level interactions can be observed when nonlinear trends at level 1 are ignored.

The time trend models described in this thesis aimed to explore variation in ESM measures in greater depth than currently seen in practice. This goal has certainly been achieved; however, the selection of models presented is by no means exhaustive and the method applied can be adapted to different scenarios. The time trend models presented have mainly focused on how one measure varies over the observation period, however, identifying how one measure varies in terms of another may also be of interest. For subject-level grouping variables this was explored using the complex variation models of Section 5.3.2.3 where variation or fluctuation in ESM measures were estimated for each group in the level 1 residuals. Not investigated was the scenario of comparing variation in two or more level 1 variables. For example, one may wish to examine how ESM measured self-esteem and recovery covary over the sampling period. For this question a new set of methods is not required, instead the models developed can simply be further extended. MacCallum, Kim et al. (1997) describe how trends on different measures can be compared using multivariate multilevel models. Here a new dependent variable is defined containing both outcomes, and dummy variables are used to indicate to which outcome each part of the model corresponds to. This formulation is equivalent to the process described for the joint modelling approach for the initial conditions problem in Chapter 6, where instead of different outcomes variables, the same response variable at different time points was joint modelled. As an example, to compare the weekly trends of two variables, self-esteem and recovery, the multivariate model specifies the random slopes (day number) for both outcome variables as multivariate normal, thus allowing for the estimation of the covariance of the intercepts and slopes for each outcome. Covariances are estimated for all pairs of random effects, i.e. the intercept and slope covariance $cov(u_{0kl}, u_{1kl})$ for each outcome l as well as the covariances between outcome intercepts $cov(u_{0k1}, u_{0k2})$, outcome slopes $cov(u_{1k1m}, u_{1k2})$ and outcome intercept/slope covariances $cov(u_{0k1}, u_{1k2})$ and $cov(u_{0k2}, u_{1k1})$. Inference would be drawn from the fixed effects of the model, describing the weekly trends of self-esteem and recovery separately, and the random effect covariance of the two slopes $cov(u_{1k1}, u_{1k2})$, describing the level of association between the trends of both variables.

This multivariate parameterisation can accommodate the variety of models presented to investigate complex time trends and allows for different time trend models for each outcome through alternative model expressions for each indicator.

The practicalities of fitting such complex models in practice, however, may be questionable. The simple weekly trends model comparing the trends of two variables requires four variance and six covariance parameters to be estimated at level 3 in addition to the level 2 and level 1 variance estimates. More complex models with a greater number of random slopes will require many more random effect variance and covariance estimates and estimating the covariance matrix when comparing multiple outcomes under these conditions may require very large sample sizes.

8.5 Applications outside of ESM

Although the methods described in this thesis have been developed with ESM as the focus, the applications of this work are not limited to just this area. Intensive longitudinal data can arise in a variety of settings, with real-time data collection becoming much more feasible thanks to developments in mobile phone technology. An ongoing study, Cloudy with a chance of pain (uMotif 2016), is currently in progress, for example, using a mobile phone application to investigate how arthritic pain varies with weather conditions. Unlike the ESM studies presented previously, this study is open to anyone in the UK with arthritic pain who has a smartphone. Current location is continuously recorded to link to weather services and a daily questionnaire relating to symptoms is administered each evening, monitoring each participant for six months. The exploration of trends in this data can apply the models developed for time trends in this thesis; instead of daily and weekly trends, variation can be assessed over much longer periods. The challenge of linking the pain data to weather conditions will be an extension of these methods, in particular the multivariate trend models discussed above for assessing the covariance of two continuous measures. Such unregulated data collection, however, will inevitably pose new problems. Missing data will likely be a much greater issue: six months of daily monitoring will almost certainly be limited by participant fatigue with additional intermittent data collection within that period. Establishing trends with such sporadic data will likely present many new challenges, but the ideas established in this thesis can be used as a starting point for new methodological development.

8.6 Summary

ESM can be a useful tool for examining real-time variation in symptoms and behaviours. Whether applied in observational research or in an experimental design, this thesis has developed methodology that can be used in future ESM studies to ensure the best use of resources through efficient design and full utilization of this three-level data structure through appropriate analysis methods. To summarize, the novel contributions of this work to the field of ESM have been:

- A review of current practice

Though this is not the first review of ESM studies, to the author's knowledge no other published work evaluates the statistical methods chosen to answer specific research questions.

- An exploration of missing data

Nonresponse has been defined at the item level, moment level and day level, in contrast to the standard practice of a single figure for adherence. Through these definitions missing data can be thoroughly explored, and suggestions for efficient summaries of missing data have been presented.

- Optimal use of multilevel models

This has been demonstrated in the extension of two-level models to three-level data. This has included three-level time trend models for examining momentary variation both within-, between- and across-days through careful consideration of random slope placement, and three-level extensions to first-difference and joint modelling methods for studying predictors of change. Moreover, it has been established that change scores are not suitable for the analysis of ESM data and that the joint modelling and first-difference methods for the alternative lagged outcome model may not be applicable in an ESM context.

- Stata code for three-level power calculations

Stata programs were written for the Monte Carlo simulation of power for three-level data. These programs calculate power for varying sample sizes at the moment-, day- and subject-level and resulted in graphs which can be used to choose an appropriate combination of n to adequately power specific research questions.

9 REFERENCES

Anderson, T. W. and C. Hsiao (1981). "ESTIMATION OF DYNAMIC-MODELS WITH ERROR-COMPONENTS." Journal of the American Statistical Association **76**(375): 598-606.

Anderson, T. W. and C. Hsiao (1982). "Formulation and estimation of dynamic models using panel data." Journal of Econometrics **18**(1): 47-82.

Andriod, S. A. Retrieved 12/8/13, from https://sites.google.com/site/sleepasandroid/home.

Bauer, D. J. and L. Cai (2009). "Consequences of unmodeled nonlinear effects in multilevel models." Journal of Educational and Behavioral Statistics **34**(1): 97-114.

Ben-Zeev, D., M. A. Young, et al. (2012). "Real-time predictors of suicidal ideation: mobile assessment of hospitalized depressed patients." Psychiatry Research **197**(1-2): 55-59.

Bhargava, A. and J. D. Sargan (1983). "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods." Econometrica **51**(6): 1635-1659.

Black, A. C., O. Harel, et al. (2011). "Missing data techniques for multilevel data: implications of model misspecification." Journal of Applied Statistics **38**(9): 1845-1865.

Bolger, N., A. Davis, et al. (2003). "Diary methods: Capturing life as it is lived." Annual Review of Psychology **54**: 579-616.

Bolger, N. and J. Laurenceau (2013). Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research, Guilford Publication.

Bolt, D. M., M. E. Piper, et al. (2012). "Why two smoking cessation agents work better than one: role of craving suppression." Journal of Consulting & Clinical Psychology **80**(1): 54-65.

Bradburn, N. M., L. J. Rips, et al. (1987). "Answering autobiographical questions: The impact of memory and inference on surveys." Science **236**(4798): 157-161.

Browne, W. J., M. Golalizadeh Lahi, et al. (2009). "A guide to sample size calculations for random effect models via simulation and the MLPowSim software package." University of Bristol, UK. Retrieved October **29**: 2010.

Bruehl, S., X. Liu, et al. (2012). "Associations between daily chronic pain intensity, daily anger expression, and trait anger expressiveness: an ecological momentary assessment study." Pain **153**(12): 2352-2358.

Bryk, A. S. and S. W. Raudenbush (1992). Hierarchical linear models: applications and data analysis methods, Sage Publications.

Buckner, J. D., R. D. Crosby, et al. (2012). "Immediate antecedents of marijuana use: an analysis from ecological momentary assessment." Journal of Behavior Therapy & Experimental Psychiatry **43**(1): 647-655.

Carpenter, B., A. Gelman, et al. (2016). "Stan: A probabilistic programming language." J Stat Softw.

Carpenter, J. and M. Kenward (2007). "Guidelines for handling missing data in Social Science Research."

Carpenter, J. and M. Kenward (2012). Multiple Imputation and its Application, Wiley.

Carpenter, J. R., H. Goldstein, et al. (2011). "REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types." Journal of Statistical Software **45**(5).

Carpenter, J. R., M. G. Kenward, et al. (2007). "Sensitivity analysis after multiple imputation under missing at random: a weighting approach." Statistical Methods in Medical Research **16**(3): 259-275.

ClinTouch. from http://www.clintouch.com/.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, L. Erlbaum Associates.

Collins, L. M., J. L. Schafer, et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." Psychological Methods **6**(4): 330.

Cook, J. E., J. E. Calcagno, et al. (2012). "Friendship trumps ethnicity (but not sexual orientation): comfort and discomfort in inter-group interactions." British Journal of Social Psychology **51**(2): 273-289.

Crouchley, R., D. Stott, et al. (2009). "Multivariate Generalised Linear Mixed Models via sabreStata (Sabre in Stata) Version 1 (Draft)." Centre for e-Science, Lancaster University.

Csikszentmihalyi, M. (2014). Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi, Springer Netherlands.

Cunningham, T. D. and R. E. Johnson (2012). "Design effects for sample size computation in three-level designs." Statistical Methods in Medical Research: 0962280212460443.

Delespaul, P. A. E. G. (1995). Assessing Schizophrenia in Daily Life: The Experience Sampling Method, UPM, Universitaire Pers Maastricht.

Demiralp, E., R. J. Thompson, et al. (2012). "Feeling blue or turquoise? Emotional differentiation in major depressive disorder." Psychol Sci **23**(11): 1410-1416.

Diggle, P. (2002). Analysis of Longitudinal Data, Oxford University Press.

Diggle, P. J. (1988). "An approach to the analysis of repeated measurements." Biometrics **44**(4): 959-971.

Donders, A. R. T., G. J. M. G. van der Heijden, et al. (2006). "Review: a gentle introduction to imputation of missing values." Journal of clinical epidemiology **59**(10): 1087-1091.

Ebbes, P., U. Böckenholt, et al. (2004). "Regressor and random-effects dependencies in multilevel models." Statistica Neerlandica **58**(2): 161-178.

Elavsky, S., P. C. M. Molenaar, et al. (2012). "Daily physical activity and menopausal hot flashes: applying a novel within-person approach to demonstrate individual differences." Maturitas **71**(3): 287-293.

Faul, F., E. Erdfelder, et al. (2007). "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences." Behavior research methods **39**(2): 175-191.

Ferron, J., R. Dailey, et al. (2002). "Effects of misspecifying the first-level error structure in two-level models of change." Multivariate Behavioral Research **37**(3): 379-403.

Fitzmaurice, G. M., N. M. Laird, et al. (2012). Applied Longitudinal Analysis, Wiley.

Forbes, E. E., S. D. Stepp, et al. (2012). "Real-world affect and social context as predictors of treatment response in child and adolescent depression and anxiety: an ecological momentary assessment study." Journal of Child & Adolescent Psychopharmacology **22**(1): 37-47.

Gardiner, J. C., Z. Luo, et al. (2009). "Fixed effects, random effects and GEE: What are the differences?" Statistics in Medicine **28**(2): 221-239.

Gelman, A. and J. Carlin (2014). "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." Perspectives on Psychological Science **9**(6): 641-651.

Giesbrecht, G. F., T. Campbell, et al. (2012). "Psychological distress and salivary cortisol covary within persons during pregnancy." Psychoneuroendocrinology **37**(2): 270-279.

Giesbrecht, G. F., N. Letourneau, et al. (2012). "Affective experience in ecologically relevant contexts is dynamic and not progressively attenuated during pregnancy." Arch Womens Ment Health **15**(6): 481-485.

Goldschmidt, A. B., S. G. Engel, et al. (2012). "Momentary affect surrounding loss of control and overeating in obese adults with and without binge eating disorder." Obesity **20**(6): 1206-1211.

Goldstein, H. (2009). REALCOM-Impute: Multiple Imputation using MLwiN, User Guide http://www.bristol.ac.uk/cmm/software/realcom/imputation.html, University of Bristol

Goldstein, H., M. J. R. Healy, et al. (1994). "MULTILEVEL TIME-SERIES MODELS WITH APPLICATIONS TO REPEATED-MEASURES DATA." Statistics in Medicine **13**(16): 1643-1655.

Gottman, J. M. (1990). "Time-series analysis applied to physiological data."

Granger, C. W. (1969). "Investigating causal relations by econometric models and cross-spectral methods." Econometrica: Journal of the Econometric Society: 424-438.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, Volume 5, number 4, NBER**:** 475-492.

Hedeker, D., R. J. Mermelstein, et al. (2009). "Modeling mood variation associated with smoking: an application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data." Addiction **104**(2): 297-307.

Henquet, C., J. van Os, et al. (2010). "Psychosis reactivity to cannabis use in daily life: an experience sampling study." The British Journal of Psychiatry **196**(6): 447-453.

Heo, M. and A. C. Leon (2008). "Statistical power and sample size requirements for three level hierarchical cluster randomized trials." Biometrics **64**(4): 1256-1262.

Heo, M., X. N. Xue, et al. (2013). "Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials with random slopes." Computational Statistics & Data Analysis **60**: 169-178.

Hox, J. J. (2002). Multilevel Analysis: Techniques and Applications, Lawrence Erlbaum Associates.

Hsiao, C. (2003). Analysis of Panel Data, Cambridge University Press.

Ilies, R. and T. A. Judge (2002). "Understanding the dynamic relationships among personality, mood, and job satisfaction: A field experience sampling study." Organizational behavior and human decision processes **89**(2): 1119-1139.

Jackson, J., P. Kuppens, et al. (2011). "Expression of anger in depressed adolescents: The role of the family environment." Journal of abnormal child psychology **39**(3): 463-474.

Kazemi, I. and R. Crouchley (2006). Modelling the initial conditions in dynamic regression models of panel data with random effects. Panel Data Econometrics: Theoretical Contributions and Empirical Applications. B. H. Baltagi. **274:** 91-117.

Kazemi, I. and R. Davies (2002). "The asymptotic bias of MLEs for dynamic panel data models." Statistical modelling in society, Proceedings of the 17th IWSM, Chania, Greece: 391-395.

Kikuchi, H., K. Yoshiuchi, et al. (2012). "Diurnal variation of tension-type headache intensity and exacerbation: An investigation using computerized ecological momentary assessment." Biopsychosoc Med **6**(1): 18.

Koval, P. and P. Kuppens (2012). "Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia." Emotion **12**(2): 256-267.

Kramer, I., C. J. P. Simons, et al. (2014). "A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial." World Psychiatry **13**(1): 68-77.

Kreft, I. G. (1996). Are multilevel techniques necessary? An overview, including simulation studies, California State University Press, Los Angeles.

Kuppens, P., D. Champagne, et al. (2012). "The Dynamic Interplay between Appraisal and Core Affect in Daily Life." Front Psychol **3**: 380.

Kwok, O., S. G. West, et al. (2007). "The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study." Multivariate Behavioral Research **42**(3): 557-592.

Landau, S. and D. Stahl (2013). "Sample size and power calculations for medical studies by simulation when closed form expressions are not available." Statistical Methods in Medical Research **22**(3): 324-345.

Larson R and M. Csikszentmihalyi (1983). The Experience Sampling Method. Directions for Methodology of Social and Behavioral Science. **15:** 41-56.

LePage, M. L., M. Price, et al. (2012). "The effect of exercise absence on affect and body dissatisfaction as moderated by obligatory exercise beliefs and eating disordered beliefs and behaviors." Psychol Sport Exerc **13**(4): 500-508.

Litt, M. D., N. L. Cooney, et al. (1998). "Ecological momentary assessment (EMA) with treated alcoholics: methodological problems and potential solutions." Health Psychology **17**(1): 48.

Little, R. J. (1993). "Pattern-mixture models for multivariate incomplete data." Journal of the American Statistical Association **88**(421): 125-134.

Little, R. J. (1995). "Modeling the drop-out mechanism in repeated-measures studies." Journal of the American Statistical Association **90**(431): 1112-1121.

Little, R. J. A. and D. B. Rubin (1987). Statistical analysis with missing data, Wiley.

Maas, C. J. M. and J. J. Hox (2005). "Sufficient Sample Sizes for Multilevel Modeling." Methodology: European Journal of Research Methods for the Behavioral and Social Sciences **1**(3): 86-92.

MacCallum, R. C., C. Kim, et al. (1997). "Studying multivariate change using multilevel models and latent curve models." Multivariate Behavioral Research **32**(3): 215-253.

Mak, T. N., C. J. Prynne, et al. (2012). "Assessing eating context and fruit and vegetable consumption in children: new methods using food diaries in the UK National Diet and Nutrition Survey Rolling Programme." Int J Behav Nutr Phys Act **9**: 126.

Mata, J., R. J. Thompson, et al. (2012). "Walk on the bright side: physical activity and affect in major depressive disorder." Journal of Abnormal Psychology **121**(2): 297-308.

McCabe, K. O. and W. Fleeson (2012). "What is extraversion for? Integrating trait and motivational perspectives and identifying the purpose of extraversion." Psychol Sci **23**(12): 1498-1505.

Menne-Lothmann, C., N. Jacobs, et al. (2012). "Genetic and environmental causes of individual differences in daily life positive affect and reward experience and its overlap with stress-sensitivity." Behav Genet **42**(5): 778-786.

Messiah, A., O. Grondin, et al. (2011). "Factors associated with missing data in an experience sampling investigation of substance use determinants." Drug Alcohol Depend **114**(2): 153-158.

Molenberghs, G., H. Thijs, et al. (2004). "Analyzing incomplete longitudinal clinical trial data." Biostatistics **5**(3): 445-464.

Morren, M., S. van Dulmen, et al. (2009). "Compliance with momentary pain measurement using electronic diaries: A systematic review." European Journal of Pain **13**(4): 354-365.

Muller, A., J. E. Mitchell, et al. (2012). "Mood states preceding and following compulsive buying episodes: an ecological momentary assessment study." Psychiatry Res **200**(2-3): 575-580.

Munsch, S., A. H. Meyer, et al. (2012). "Binge eating in binge eating disorder: a breakdown of emotion regulatory process?" Psychiatry Research **195**(3): 118-124.

Myers, T. A., D. R. Ridolfi, et al. (2012). "The impact of appearance-focused social comparisons on body image disturbance in the naturalistic environment: the roles of thin-ideal internalization and feminist beliefs." Body Image **9**(3): 342-351.

Myin-Germeys, I., A. Klippel, et al. (2016). "Ecological momentary interventions in psychiatry." Current opinion in psychiatry **29**(4): 258-263.

Oorschot, M., T. Lataster, et al. (2012). "Temporal dynamics of visual and auditory hallucinations in psychosis." Schizophr Res **140**(1): 77-82.

Oorschot, M., T. Lataster, et al. (2012). "Mobile assessment in schizophrenia: a data-driven momentary approach." Schizophrenia Bulletin **38**(3): 405-413.

Palmier-Claus, J., J. Ainsworth, et al. (2012). "The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application." BMC psychiatry **12**(1): 172.

Palmier-Claus, J. E., I. Myin-Germeys, et al. (2011). "Experience sampling research in individuals with mental illness: reflections and guidance." Acta Psychiatrica Scandinavica **123**(1): 12-20.

Palmier-Claus, J. E., P. J. Taylor, et al. (2012). "Affective variability predicts suicidal ideation in individuals at ultra-high risk of developing psychosis: an experience sampling study." British Journal of Clinical Psychology **51**(1): 72-83.

Pan, H. and H. Goldstein (1998). "Multi-level repeated measures growth modelling using extended spline functions." Statistics in Medicine **17**(23): 2755-2770.

Peters, E., T. Lataster, et al. (2012). "Appraisals, psychotic symptoms and affect in daily life." Psychological Medicine **42**(5): 1013-1023.

Rabe-Hesketh, S. and A. Skrondal (2012). Multilevel and Longitudinal Modeling Using Stata, Volumes I and II, Third Edition, Taylor & Francis.

Raudenbush, S. W., A. S. Bryk, et al. (2004). "HLM 6 for Windows [Computer software]." Lincolnwood, IL: Scientific Software International.

Robins, R. W., R. C. Fraley, et al. (2009). Handbook of Research Methods in Personality Psychology, Guilford Publications.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys, Wiley.

Rubin, D. B. (1996). "Multiple imputation after 18+ years." Journal of the American Statistical Association **91**(434): 473-489.

Russell, J. A., A. Weiss, et al. (1989). "Affect grid: a single-item scale of pleasure and arousal." Journal of personality and social psychology **57**(3): 493.

Schafer, J. L. (1999). "Multiple imputation: a primer." Statistical Methods in Medical Research **8**(1): 3-15.

Schafer, J. L. and M. K. Olsen (1998). "Multiple imputation for multivariate missing-data problems: A data analyst's perspective." Multivariate Behavioral Research **33**(4): 545-571.

Schwartz, J. E. and A. A. Stone (1998). "Strategies for analyzing ecological momentary assessment data." Health Psychology **17**(1).

Schwerdtfeger, A. R. and S.-M. Scheel (2012). "Self-esteem fluctuations and cardiac vagal control in everyday life." International Journal of Psychophysiology **83**(3): 328-335.

Selby, E. A., P. Doyle, et al. (2012). "Momentary emotion surrounding bulimic behaviors in women with bulimia nervosa and borderline personality disorder." J Psychiatr Res **46**(11): 1492-1500.

Shadel, W. G., S. C. Martino, et al. (2012). "Momentary effects of exposure to prosmoking media on college students' future smoking risk." Health Psychology **31**(4): 460-466.

Shiffman, S., M. Hufford, et al. (1997). "Remember that? A comparison of real-time versus retrospective recall of smoking lapses." Journal of consulting and clinical psychology **65**(2): 292.

Shiffman, S. and A. A. Stone (1998). "Introduction to the special section: Ecological momentary assessment in health psychology." Health Psychology **17**(1).

Shiffman, S., A. A. Stone, et al. (2008). "Ecological momentary assessment." Annual Review of Clinical Psychology **4**.

Shiyko, M. P., S. T. Lanza, et al. (2012). "Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self confidence on smoking urges: differences between successful quitters and relapsers." Prevention Science **13**(3): 288-299.

Siddiqui, O., H. M. J. Hung, et al. (2009). "MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets." Journal of Biopharmaceutical Statistics **19**(2): 227-246.

Silvia, P. J., T. R. Kwapil, et al. (2013). "Missed Beeps and Missing Data: Dispositional and Situational Predictors of Nonresponse in Experience Sampling Research." Social Science Computer Review: 0894439313479902.

Sivo, S., X. Fan, et al. (2005). "The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates." Structural Equation Modeling **12**(2): 215-231.

Skrondal, A. and S. Rabe-Hesketh (2007). Multilevel and related models for longitudinal data. Handbook of multilevel analysis. J. Leeuw and E. Meijer. New York, Springer**:** (277-301).

SleepCycle. from https://itunes.apple.com/gb/app/sleep-cycle-alarm-clock/id320606217?mt=8.

Snijders, T. A. (2005). "Power and sample size in multilevel linear models." Encyclopedia of statistics in behavioral science.

Snijders, T. A. and R. J. Bosker (1993). "Standard errors and sample sizes for two-level research." Journal of Educational and Behavioral Statistics **18**(3): 237-259.

Snijders, T. A. B. and R. J. Bosker (1999). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, Sage Publications.

Solhan, M. B., T. J. Trull, et al. (2009). "Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall." Psychological Assessment **21**(3): 425-436.

StataCorp (2013). Stata 13 Base Reference Manual. College Station, TX, Stata Press.

Steele, F. (2008). Module 5: Introduction to Multilevel Modelling Concepts LEMMA VLE

Steele, F. (2014). Multilevel Modelling of Repeated Measures Data. LEMMA VLE Module 15**:** 1-62.

Steele, F., J. Rasbash, et al. (2013). "A multilevel simultaneous equations model for within-cluster dynamic effects, with an application to reciprocal parent–child and sibling effects." Psychological Methods **18**(1): 87.

Sterne, J. A. C., I. R. White, et al. (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." British Medical Journal **339**.

Stone, A. A. and S. Shiffman (1994). "Ecological momentary assessment (EMA) in behavorial medicine." Annals of Behavioral Medicine.

Stone, A. A. and S. Shiffman (2002). "Capturing momentary, self-report data: A proposal for reporting guidelines." Annals of Behavioral Medicine **24**(3): 236-243.

Thewissen, V., R. P. Bentall, et al. (2008). "Fluctuations in self-esteem and paranoia in the context of daily life." Journal of Abnormal Psychology **117**(1): 143.

To, M. L., C. D. Fisher, et al. (2012). "Within-person relationships between mood and creativity." Journal of Applied Psychology **97**(3): 599-612.

Tyler, E., S. Jones, et al. (2015). "The Relationship between Bipolar Disorder and Cannabis Use in Daily Life: An Experience Sampling Study." PLoS One **10**(3): e0118916.

Udachina, A., F. Varese, et al. (2012). "Dynamics of self-esteem in "poor-me" and "bad-me" paranoia." Journal of Nervous & Mental Disease **200**(9): 777-783.

uMotif. (2016). "Cloudy with a chance of pain." from https://www.cloudywithachanceofpain.com/.

van Knippenberg, R. J. M., M. E. de Vugt, et al. (2016). "Dealing with daily challenges in dementia (deal-id study): effectiveness of the experience sampling method intervention 'Partner in Sight' for spousal caregivers of people with dementia: design of a randomized controlled trial." BMC psychiatry **16**(1): 1-14.

Verdoux, H., C. Gindre, et al. (2003). "Effects of cannabis and psychosis vulnerability in daily life: an experience sampling test study." Psychological Medicine **33**(01): 23-32.

Wichers, M., C. Lothmann, et al. (2012). "The dynamic interplay between negative and positive emotions in daily life predicts response to treatment in depression: a momentary assessment study." British Journal of Clinical Psychology **51**(2): 206-222.

Wichers, M., F. Peeters, et al. (2012). "A time-lagged momentary assessment study on daily life physical activity and affect." Health Psychology **31**(2): 135-144.

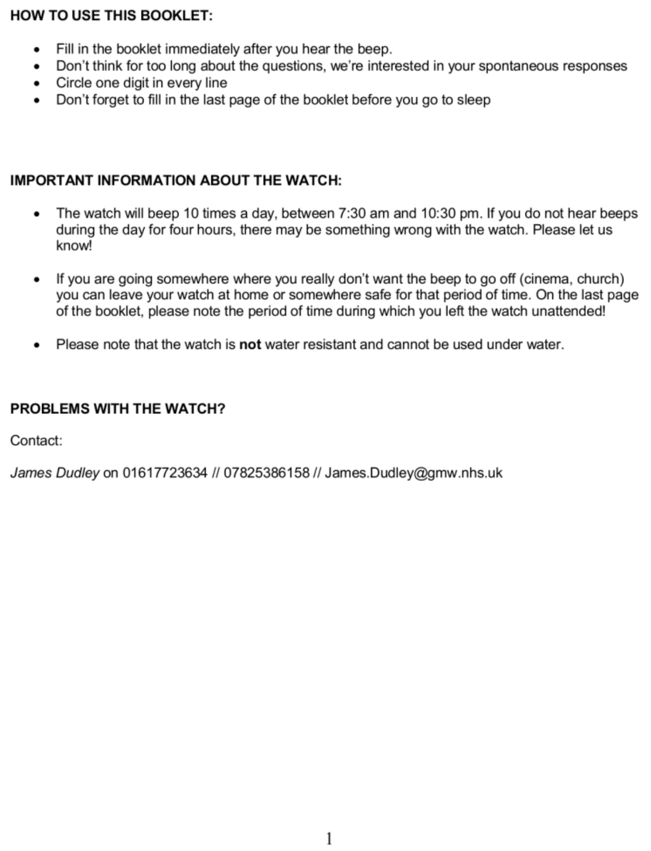
Wilhelm, F. H. and P. Grossman (2010). "Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment." Biological Psychology **84**(3).

Yeh, V. M., D. E. McCarthy, et al. (2012). "An ecological momentary assessment analysis of prequit markers for smoking-cessation failure." Exp Clin Psychopharmacol **20**(6): 479-488.

APPENDIX

APPENDIX 1: MOTIVATING EXAMPLES

A 1. RECOVERY STUDY ESM DIARY SAMPLE



HOW TO USE THIS BOOKLET:

- Fill in the booklet immediately after you hear the beep.
- Don't think for too long about the questions, we're interested in your spontaneous responses
- Circle one digit in every line
- Don't forget to fill in the last page of the booklet before you go to sleep

IMPORTANT INFORMATION ABOUT THE WATCH:

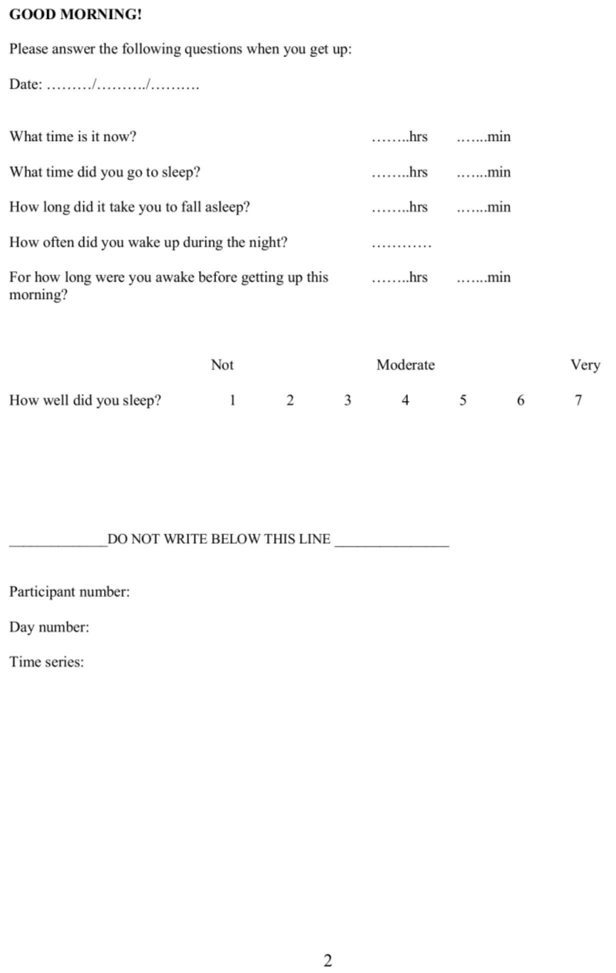
- The watch will beep 10 times a day, between 7:30 am and 10:30 pm. If you do not hear beeps during the day for four hours, there may be something wrong with the watch. Please let us know!
- If you are going somewhere where you really don't want the beep to go off (cinema, church) you can leave your watch at home or somewhere safe for that period of time. On the last page of the booklet, please note the period of time during which you left the watch unattended!
- Please note that the watch is **not** water resistant and cannot be used under water.

PROBLEMS WITH THE WATCH?

Contact:

James Dudley on 01617723634 // 07825386158 // James.Dudley@gmw.nhs.uk

1



GOOD MORNING!

Please answer the following questions when you get up:

Date: ………/………./……….

What time is it now? ……..hrs …….min

What time did you go to sleep? ……..hrs …….min

How long did it take you to fall asleep? ……..hrs …….min

How often did you wake up during the night? …………

For how long were you awake before getting up this morning? ……..hrs …….min

	Not			Moderate			Very
How well did you sleep?	1	2	3	4	5	6	7

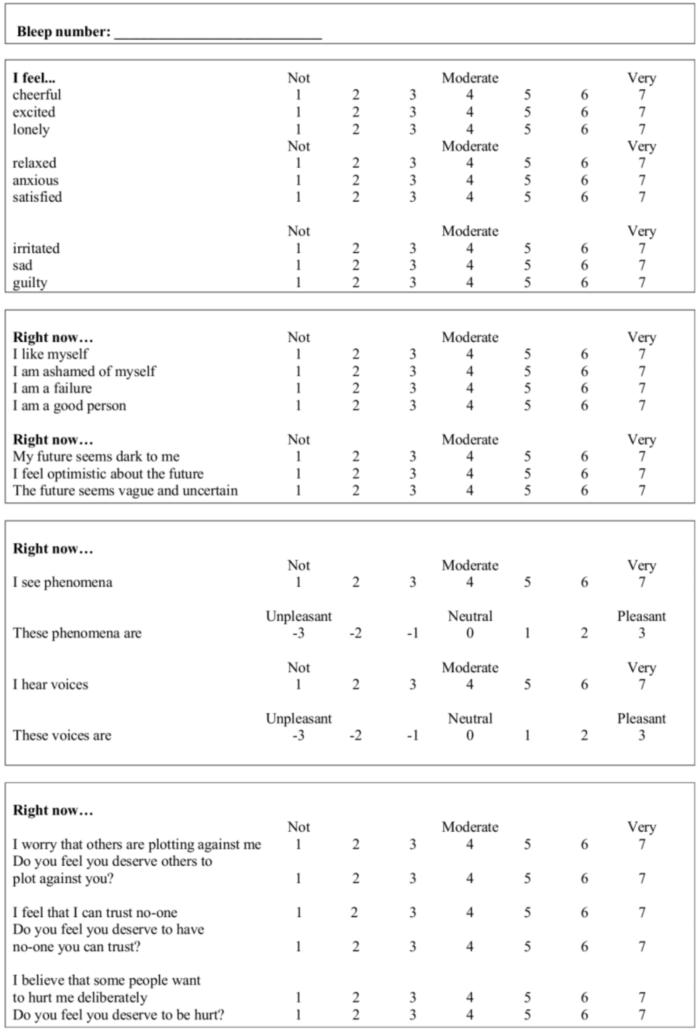
_____________DO NOT WRITE BELOW THIS LINE _______________

Participant number:

Day number:

Time series:

2



Bleep number: ___________________________

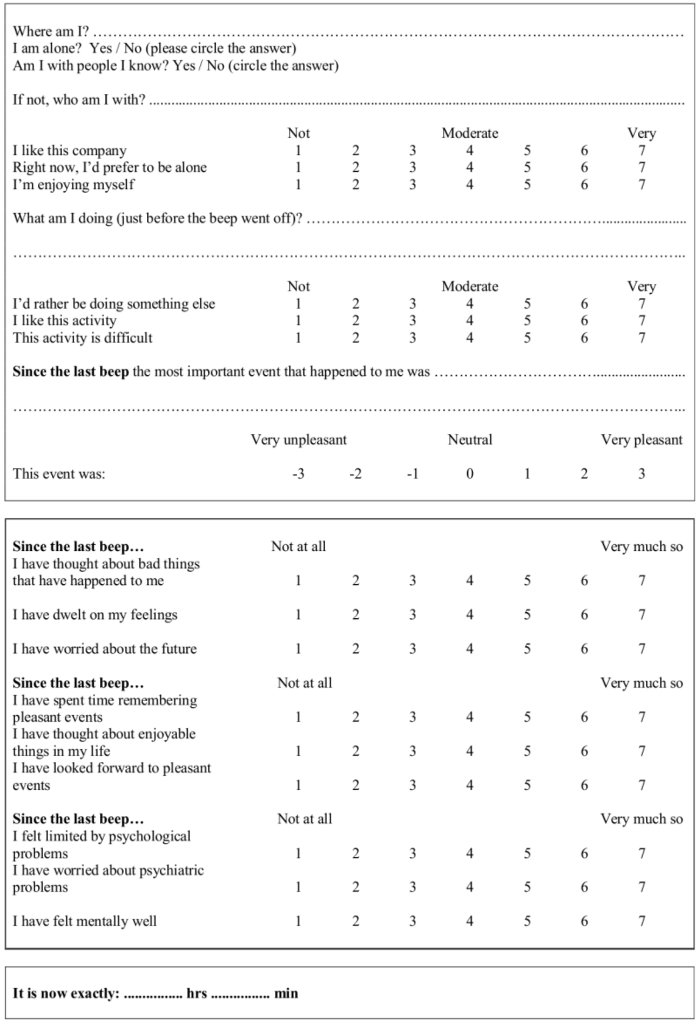
I feel...	Not			Moderate			Very
cheerful	1	2	3	4	5	6	7
excited	1	2	3	4	5	6	7
lonely	1	2	3	4	5	6	7
	Not			Moderate			Very
relaxed	1	2	3	4	5	6	7
anxious	1	2	3	4	5	6	7
satisfied	1	2	3	4	5	6	7
	Not			Moderate			Very
irritated	1	2	3	4	5	6	7
sad	1	2	3	4	5	6	7
guilty	1	2	3	4	5	6	7

Right now…	Not			Moderate			Very
I like myself	1	2	3	4	5	6	7
I am ashamed of myself	1	2	3	4	5	6	7
I am a failure	1	2	3	4	5	6	7
I am a good person	1	2	3	4	5	6	7
Right now…	Not			Moderate			Very
My future seems dark to me	1	2	3	4	5	6	7
I feel optimistic about the future	1	2	3	4	5	6	7
The future seems vague and uncertain	1	2	3	4	5	6	7

Right now…							
	Not			Moderate			Very
I see phenomena	1	2	3	4	5	6	7
	Unpleasant			Neutral			Pleasant
These phenomena are	-3	-2	-1	0	1	2	3
	Not			Moderate			Very
I hear voices	1	2	3	4	5	6	7
	Unpleasant			Neutral			Pleasant
These voices are	-3	-2	-1	0	1	2	3

Right now…							
	Not			Moderate			Very
I worry that others are plotting against me	1	2	3	4	5	6	7
Do you feel you deserve others to plot against you?	1	2	3	4	5	6	7
I feel that I can trust no-one	1	2	3	4	5	6	7
Do you feel you deserve to have no-one you can trust?	1	2	3	4	5	6	7
I believe that some people want to hurt me deliberately	1	2	3	4	5	6	7
Do you feel you deserve to be hurt?	1	2	3	4	5	6	7

3



Where am I? ……………………………………………………………………………………………………

I am alone? Yes / No (please circle the answer)

Am I with people I know? Yes / No (circle the answer)

If not, who am I with? ....

	Not			Moderate			Very
I like this company	1	2	3	4	5	6	7
Right now, I'd prefer to be alone	1	2	3	4	5	6	7
I'm enjoying myself	1	2	3	4	5	6	7

What am I doing (just before the beep went off)? ……………………………………………………......................

……

	Not			Moderate			Very
I'd rather be doing something else	1	2	3	4	5	6	7
I like this activity	1	2	3	4	5	6	7
This activity is difficult	1	2	3	4	5	6	7

Since the last beep the most important event that happened to me was …………………………......................

……

	Very unpleasant			Neutral			Very pleasant
This event was:	-3	-2	-1	0	1	2	3

Since the last beep…	Not at all						Very much so
I have thought about bad things that have happened to me	1	2	3	4	5	6	7
I have dwelt on my feelings	1	2	3	4	5	6	7
I have worried about the future	1	2	3	4	5	6	7
Since the last beep…	Not at all						Very much so
I have spent time remembering pleasant events	1	2	3	4	5	6	7
I have thought about enjoyable things in my life	1	2	3	4	5	6	7
I have looked forward to pleasant events	1	2	3	4	5	6	7
Since the last beep…	Not at all						Very much so
I felt limited by psychological problems	1	2	3	4	5	6	7
I have worried about psychiatric problems	1	2	3	4	5	6	7
I have felt mentally well	1	2	3	4	5	6	7

It is now exactly: .... hrs .... min

4

APPENDIX 2: SYSTEMATIC REVIEW

A 2.1: DATA EXTRACTION FORM

Author	
Title	
Year	

Study Eligibility				**Notes**
Is it a study using ESM?	Yes	Unclear	No Exclude	
Pilot study?	No	Unclear	Yes Exclude	
ESM used to validate measure?	No	Unclear	Yes Exclude	
Used to assess feasibility for future study	No	Unclear	Yes Exclude	
Used to assess recall	No	Unclear	Yes Exclude	
Included	1 X "Exclude" = Exclude 1 X "Unclear" = Unclear"			□ Yes □ No □ Unclear

Background		**Notes**
Research Area	□ Unclear	
Research Area 2	□ Unclear	
Research Question	□ Association □ Response to treatment □ Mediation □ Moderation □ Other ……………………………………	
Randomised Trial	□ Yes □ No □ Unclear	
Control Group	□ Yes □ No □ Unclear	

ESM		**Notes**
ESM or EMA	□ ESM □ EMA □ Unclear	
Method	□ 1 – Paper □ 2 – PDA □ 3 – Text □ 4 – call □ 5 – internet □ 6 – app □ 999 - unclear	
ESM design	□ 1 – Event □ 2 – Interval □ 3 – Random □ 4 – E&R □ 5 – I&R □ 6 – E&I	
Ambulatory assessment	□ Yes □ No □ Unclear	
AA Method	□ N/A □ Unclear	

Compensation given?	□ Yes	□ No	□ Unclear	

Data		Number			Notes
Participants (total)				□ Unclear	
# in control group			□ N/A	□ Unclear	
Measurement	Overall			□ Unclear	
	Days			□ Unclear	
	Beeps			□ Unclear	
Min # to be completed				□ Unclear	
Missing data		□ Yes	□ No	□ Unclear	
Adherence (%)				□ Unclear	
MD technique				□ Unclear	

Analysis			Notes
Analysis		□ Unclear	
Reference		□ Unclear	
Analysis 2		□ Unclear	
Reference		□ Unclear	

If Multilevel…					Notes
Lagged covariates	□ Yes	□ No	□ Not ML	□ Unclear	
Lagged outcome as covariate	□ Yes	□ No	□ Not ML	□ Unclear	
Continuous covariates centred	□ Yes	□ No	□ Not ML	□ Unclear	
Levels			□ Not ML	□ Unclear	
Autocorrelation?	□ Yes	□ No	□ Not ML	□ Unclear	
Covariance structure specified?			□ Not ML	□ Unclear	

Further Notes

A 2.2: Full list of research areas

Area	N	Area	N	Area	N
Affect	10	Enjoyment	1	Physical activity	5
Alcohol use	3	Exercise	2	Pregnancy	2
Anger	1	Eye tracking	1	Productivity	1
Anxiety	5	Flow	3	Psychological demands	1
Appraisal	1	Genotype	1	Psychosis	8
Bipolar disorder	1	Hangovers	1	Rehabilitation	1
Body image	1	Menopause	1	Relapse	1
Borderline personality disorder	2	Mind wandering	1	Self-harm	1
Cardiac vagal control	1	Mindfulness	1	Self-esteem	2
Carotid artery atherosclerosis	1	Mood	2	Smoking	10
Challenge	1	Motivation	1	Social conflict	1
Compulsive buying behaviour	1	Motor control	1	Social functioning	1
Coping	1	Nutrition	1	Social interaction	2
Craving	1	Occupational health	1	Statistical theory	1
Creativity	1	Pain	3	Stress	5
Depression	6	Panic disorder	1	Suicide ideation	2
Desire	2	Paranoia	2	Virtual reality	1
Drug use	4	Parental communication	1	Wellbeing	2
Eating disorder	5	Perfectionist concerns	1		
Emotion	3	Personality	1		

* Two areas of research were allowed for each study so percentages are not given. 74 studies.

Appendix 3: Missing data

A 3. Full table of item nonresponse

Item	1	2	3	4	5	6	7	8	9	10	Total
cheerful	4%	6%	4%	4%	5%	4%	3%	3%	2%	2%	4%
excited	6%	7%	6%	7%	7%	5%	4%	5%	4%	3%	5%
lonely	4%	5%	3%	4%	4%	1%	2%	1%	0%	1%	2%
relaxed	4%	6%	4%	5%	3%	3%	4%	2%	2%	1%	3%
anxious	4%	4%	4%	4%	3%	2%	4%	3%	3%	2%	3%
satisfied	7%	8%	6%	8%	5%	5%	5%	6%	2%	4%	6%
irritated	6%	8%	6%	6%	6%	4%	5%	4%	3%	4%	5%
sad	8%	8%	6%	7%	7%	4%	5%	4%	3%	3%	6%
guilty	6%	9%	6%	7%	5%	3%	5%	4%	3%	3%	5%
Self-esteem 1	4%	4%	3%	3%	2%	2%	1%	1%	2%	1%	2%
Self-esteem 2	4%	6%	4%	5%	4%	2%	2%	3%	1%	1%	3%
Self-esteem 3	4%	5%	4%	4%	3%	3%	4%	3%	3%	2%	3%
Self-esteem 4	4%	6%	5%	5%	4%	3%	4%	5%	2%	3%	4%
belonging	6%	7%	7%	6%	6%	5%	4%	5%	6%	3%	5%
warmth	9%	13%	11%	12%	8%	8%	8%	7%	7%	8%	9%
Future 1	7%	5%	4%	5%	4%	5%	3%	3%	3%	1%	4%
Future 2	8%	6%	5%	5%	4%	6%	3%	3%	4%	1%	4%
Future 3	7%	6%	5%	5%	4%	5%	4%	3%	3%	0%	4%
Hallucination 1	7%	8%	8%	7%	8%	5%	6%	5%	4%	2%	6%
Hallucination 2	86%	85%	87%	85%	86%	85%	85%	84%	86%	87%	86%
Hallucination 3	7%	9%	9%	8%	9%	7%	6%	8%	7%	8%	8%
Hallucination 4	75%	75%	74%	70%	75%	71%	68%	73%	70%	73%	72%
Paranoia 1	4%	3%	3%	3%	1%	2%	1%	1%	1%	1%	2%

Paranoia 2	59%	61%	64%	62%	64%	59%	59%	62%	63%	61%	62%
Paranoia 3	5%	6%	4%	4%	2%	2%	1%	1%	3%	1%	3%
Paranoia4	54%	57%	60%	57%	59%	55%	58%	58%	59%	55%	57%
Paranoia 5	3%	5%	3%	4%	3%	2%	1%	1%	1%	1%	2%
Paranoia 6	64%	60%	62%	62%	61%	58%	60%	61%	63%	60%	61%
Current location	6%	6%	4%	5%	4%	3%	2%	4%	2%	1%	4%
Alone	5%	6%	5%	4%	5%	3%	3%	4%	2%	2%	4%
Company 1	49%	42%	40%	36%	38%	37%	38%	43%	40%	39%	40%
Company 2	34%	39%	45%	42%	42%	45%	44%	42%	44%	45%	42%
Like company	35%	29%	25%	21%	22%	24%	22%	27%	23%	26%	25%
Prefer alone	28%	25%	21%	18%	19%	20%	18%	23%	20%	20%	21%
Enjoying self	24%	21%	20%	16%	18%	19%	19%	23%	21%	20%	20%
Activity 1	7%	8%	9%	6%	8%	7%	4%	7%	4%	5%	6%
Activity 2	6%	6%	5%	4%	4%	3%	4%	4%	2%	2%	4%
Activity 3	8%	5%	5%	6%	5%	4%	3%	4%	3%	2%	4%
Activity 4	8%	6%	6%	6%	6%	4%	4%	6%	3%	3%	5%
Important event 1	22%	18%	17%	17%	17%	16%	18%	15%	11%	11%	16%
Important event 2	27%	27%	25%	27%	25%	22%	24%	23%	19%	23%	24%
Worry 1	7%	4%	3%	3%	1%	2%	1%	0%	0%	0%	2%
Worry 2	8%	3%	4%	5%	2%	3%	2%	2%	1%	1%	3%
Worry 3	8%	5%	5%	4%	2%	4%	2%	3%	1%	2%	4%
Positive 1	8%	3%	3%	4%	1%	2%	1%	1%	0%	0%	2%
Positive 2	8%	4%	5%	4%	3%	3%	2%	3%	1%	2%	3%
Positive 3	9%	4%	5%	5%	3%	3%	4%	4%	2%	2%	4%
Recovery 1	7%	3%	3%	3%	1%	2%	1%	1%	0%	2%	2%
Recovery 2	8%	4%	5%	5%	3%	4%	2%	4%	2%	2%	4%
Recovery 3	9%	5%	5%	6%	3%	4%	3%	4%	2%	2%	4%

APPENDIX 4: PREDICTING MOMENTARY CHANGE

A 4.1 CHANGE MODEL USING CONCURRENT PREDICTOR

Fixed effects

				$\lambda_u = \lambda_v = 1$		
	Naive model			**IC Model S2**		
$i = 1$ All days	Coeff.	SE	N	Coeff.	SE	N
α_0 Intercept			1910	4.886	0.117	2102
α_1Self-esteem (x_1)				0.435	0.047	
$i > 1$ All days						
β_0 Intercept	3.587	0.139		3.865	0.142	
β_2 Self-esteem (x_i)	0.323	0.024		0.333	0.024	
β_1 Lag Recovery (y_{i-1})	0.283	0.021		0.224	0.020	

Random effects

Level		Variance	SE	Variance	SE
Person		0.510	0.108	0.631	0.127
Day		0.058	0.015	0.080	0.016
Residuals		0.444	0.017		
	$(i = 1)$			0.433	0.016
	$(i > 1)$			0.645	0.073

APPENDIX 5: POWER AND SAMPLE SIZE

A 5.1 SIMULATION CODE FOR ASSOCIATION EXAMPLE

The program was used for all effect size and variance scenarios. The program was adapted slightly for the examples with missing data. The simulation command relates to the example where effect size $r = 0.1$, small model variances $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ and small predictor variances $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$.

```
// Simulation program R = 0.1 - MINIMUM VARIANCES //

program define assrec, rclass
      syntax [, n3(integer 1) n2(integer 1) n1(integer 1) ///
      b_0(real 0) r_x(real 0)                          ///
      sigu(real 0) sigv(real 0) sige(real 0)           ///
      x_m(real 0) x_l3v(real 0) x_l2v(real 0) x_l1v(real 0)]

      version 13
      drop _all

      set obs `n3'
      gen id = _n

      * Random intercept for id level
      gen u1=rnormal()
      gen u = u1 * sqrt(`sigu')

      * Level 3 part of predictor
      gen x_3 = rnormal(0, sqrt(`x_l3v'))

      * Expand to level 2 units
      expand `n2'
      bysort id: gen day = _n
      sort id day

      * Random intercept for day
      gen v1 = rnormal()
      gen v = v1 * sqrt(`sigv')

      * Generate day level predictor
      gen x_2 = rnormal(0, sqrt(`x_l2v'))


      * Expand for level 1 units
      expand `n1'
      bysort id day: gen beep = _n
      sort id day beep

      * Residual variance
      gen e1 = rnormal()
      gen e = e1 * sqrt(`sige')

      *Generate total SD for effect size transformation
      gen sigy = sqrt(`sigu' +`sigv' +`sige')
      * Level 1 part of predictor
      gen x_1 = rnormal(`x_m', sqrt(`x_l1v'))

      * Combine parts to make predictor x
      egen x_total = rowtotal(x_1 x_2 x_3)
      * Centre predictor
      sum x_total
      gen x = x_total - r(mean)
      * Define sd x for effect size transformation
      sum x
      * gen beta from r
      gen beta_x = `r_x'*(sigy/r(sd))

      * Generate y - recovery
      gen y = `b_0' + beta_x*x + u + v + e


      * Fit model
      mixed y x || id:, || day:, mle var iterate(1000)

      * Pull out coeffs
      capture local b_x = _b[x]
      return scalar b_x = _b[x]
      capture local se_x = _se[x]
      return scalar se_x = _se[x]

      * Return Logliklihood
      capture local ll = e(ll)
      return scalar ll = e(ll)
end

            // Simulation commands //


* Create empty data set to append to
drop _all
gen n3 =.
gen n2 = .
gen n1 = .
gen rx = .
gen siguve = .
gen x_l321v = .
gen sim = .

gen x_reject = .

save "C:\Users\mbbxjlc2\Y2 simulations\mood\x_empty.dta", replace


* Set drive and loop over l3 and l1 values
local drive = "C:\Users\mbbxjlc2\Y2
simulations\mood\b0.1_suve0.26_x321.26"

local i = 0
local sim =  984              // from power calc for simulation

forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {
      local sigu = 0.26 // lowest level variance
      local sigv = 0.26
      local sige = 0.26
      local b_0 = 3.5         // average (of 7 point scale) for when
      centred x = 0 (so average x)
      local r_x = 0.1         // small effect size
      local x_m = 0.4
      local x_l3v = 0.26      // lowest level variance
      local x_l2v = 0.26
      local x_l1v = 0.26

      local i = `i' + 1

      display `i', c(current_time)

simulate b_x = r(b_x) se_x = r(se_x) ll = r(ll), reps(`sim') seed(112)
///
: assrec, n3(`n3') n2(`n2') n1(`n1') b_0(`b_0') r_x(`r_x') sigu(`sigu')
sigv(`sigv') sige(`sige') x_m(`x_m') x_l3v(`x_l3v') x_l2v(`x_l2v')
x_l1v(`x_l1v')

// Record sample size
gen n3 = `n3'
gen n2 = `n2'
gen n1 = `n1'
gen rx = `r_x'
gen siguve = `sigu'
gen x_l321v = `x_l3v'
gen sim= `sim'


// Create z statistic
gen z_x = b_x/se_x

// Create p values
gen pval_x = 2*normal(-abs(z_x))

// Find the proportion who reject the null
gen x_reject = 0
replace x_reject = 1 if pval_x<0.05
save "`drive'\Full\Full_`n3'_`n2'_`n1'_sim`sim'.dta", replace

collapse (mean) x_reject
gen n3 = `n3'
gen n2 = `n2'
gen n1 = `n1'
gen rx = `r_x'
gen siguve = `sigu'
gen x_l321v = `x_l3v'
gen sim = `sim'

order x_reject, after(n1)

save "`drive'\Collapsed\Collapse_`n3'_`n2'_`n1'_sim`sim'.dta", replace


}     // End of n1 loop
}     // End of n2 loop
}     // End of n3 loop
```

```
* Collect all files together
// Full simulations
use "C:\Users\mbbxjlc2\Y2 simulations\mood\x_empty.dta", clear
forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {
append using "`drive'\Full\Full_`n3'_`n2'_`n1'_sim`sim'.dta"

}
}
}
save "`drive'\allFull_sim`sim'.dta", replace


// Collapsed results
use "C:\Users\mbbxjlc2\Y2 simulations\mood\x_empty.dta", clear
forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {
append using "`drive'\Collapsed\Collapse_`n3'_`n2'_`n1'_sim`sim'.dta"

}
}
}
save "`drive'\allCollapse_sim`sim'.dta", replace
```

A 5.2 Simulation code for group difference example

The program was used for all effect sizes and variances. It was slightly adapted to examine different levels of missing data. The simulation commands relate to the example with effect size 0.8 and small model variances.

```
// Simulation program - Group differences - random intercept model //

program define groupdiff, rclass
      syntax [, n3(integer 1) n2(integer 1) n1(integer 1) ///
      b0(real 0) b_rec_b(real 0)                         ///
      prob_x(real 0)                                  ///
      sigu(real 0) sigv(real 0) sige(real 0)]

      version 13
      drop _all

      set obs `n3'
      gen id = _n

      * Random intercept for id level
      gen u1 = rnormal()
      gen u = u1 * sqrt(`sigu')

      * Generate binary baseline recovery
      gen rec_b = rbinomial(1, `prob_x')


      * Expand to level 2 units
      expand `n2'
      bysort id: gen day = _n
      sort id day

      * Random intercept for day
      gen v1 = rnormal()
      gen v = v1 * sqrt(`sigv')


      * Expand for level 1 units
      expand `n1'
      bysort id day: gen beep = _n
      sort id day beep

      * Residual variance
      gen e1 = rnormal()
      gen e = e1 * sqrt(`sige')

      * Gen coeff based off cohens d
      gen r = `b_rec_b'/(sqrt(`b_rec_b'^2 + (1/(`prob_x'*(1-`prob_x')))))

      * Generate y - recovery
      gen y = `b0' + r*rec_b + u + v + e


      * Fit random intercept model
      mixed y rec_b || id:  || day:, mle var iterate(1000)

      * Pull out coeffs
      ** Group difference

      capture local b_rec_b = _b[rec_b]
      return scalar b_rec_b = _b[rec_b]
      capture local se_rec_b = _se[rec_b]
      return  scalar se_rec_b = _se[rec_b]


end

            // Simulation commands //


* Create empty data set to append to
drop _all
gen n3 =.
gen n2 = .
gen n1 = .
gen rec_b_reject = .

save "C:\Users\mbbxjlc2\Y2 simulations\group diff\empty.dta", replace


* Set drive and loop
local drive = "C:\Users\mbbxjlc2\Y2 simulations\group
diff\b0.8_px0.33_suve0.26"

local i = 0
local sim = 984 // based on power calc

forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {

      local prob_x = 0.33 // From recovery data, prob of recovered
      local b0 =  3.5 // midway on 1-7 scale
      local b_rec_b =  0.8 // Large cohen's d
      local sigu =  0.26
      local sigv =   0.26
      local sige =   0.26

      local i = `i' + 1

      display `i', c(current_time)

simulate b_rec_b = r(b_rec_b) se_rec_b = r(se_rec_b) , reps(`sim')
seed(112)    ///
: groupdiff, n3(`n3') n2(`n2') n1(`n1') prob_x(`prob_x') b0(`b0')
b_rec_b(`b_rec_b') sigu(`sigu') sigv(`sigv') sige(`sige')

// Record sample size
gen n3 = `n3'
gen n2 = `n2'
gen n1 = `n1'

gen sim= `sim'

// Create z statistic

gen z_rec_b = b_rec_b/se_rec_b
// Create p values
gen pval_rec_b = 2*normal(-abs(z_rec_b))

// Find the proportion who reject the null
gen rec_b_reject = 0
replace rec_b_reject = 1 if pval_rec_b<0.05

save "`drive'\Full\gdFull_`n3'_`n2'_`n1'_sim`sim'.dta", replace

collapse (mean) rec_b_reject
gen n3 = `n3'
gen n2 = `n2'
gen n1 = `n1'
gen sim = `sim'

order rec_b_reject, after(n1)

* Need something that counts # of simulations

save "`drive'\Collapsed\gdCollapse_`n3'_`n2'_`n1'_sim`sim'.dta", replace


}
}
}

*save "`drive'\gd_all.dta", replace


* Collect all files together
// Full simulations
use "C:\Users\mbbxjlc2\Y2 simulations\group diff\empty.dta", clear
forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {
append using "`drive'\Full\gdFull_`n3'_`n2'_`n1'_sim`sim'.dta"

}
}
}
save "`drive'\gd_allFull_sim`sim'.dta", replace


// Collapsed results
use "C:\Users\mbbxjlc2\Y2 simulations\group diff\empty.dta", clear
forvalues n3 = 10 (10) 60 {
forvalues n2 = 6 (1) 10 {
forvalues n1 = 4 (2) 10 {
append using "`drive'\Collapsed\gdCollapse_`n3'_`n2'_`n1'_sim`sim'.dta"

}
}
}
save "`drive'\gd_allCollapse_sim`sim'.dta", replace
```

A 5.3 Simulated power table for association example: effect size 0.1 and 0.3, small variance estimates

Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				Effect size $r = 0.3$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$			
n3	n2	n1	Proportion rejected	n3	n2	n1	Proportion rejected
10	6	4	0.362805	10	6	4	0.998984
10	6	6	0.501016	10	6	6	1
10	6	8	0.614837	10	6	8	1
10	6	10	0.72561	10	6	10	1
10	7	4	0.422764	10	7	4	0.997968
10	7	6	0.548781	10	7	6	1
10	7	8	0.689024	10	7	8	1
10	7	10	0.772358	10	7	10	1
10	8	4	0.469512	10	8	4	1
10	8	6	0.629065	10	8	6	1
10	8	8	0.730691	10	8	8	1
10	8	10	0.804878	10	8	10	1
10	9	4	0.515244	10	9	4	1
10	9	6	0.678862	10	9	6	1
10	9	8	0.796748	10	9	8	1
10	9	10	0.884146	10	9	10	1
10	10	4	0.534553	10	10	4	1
10	10	6	0.691057	10	10	6	1
10	10	8	0.830285	10	10	8	1
10	10	10	0.894309	10	10	10	1
20	6	4	0.597561	20	6	4	1
20	6	6	0.776423	20	6	6	1
20	6	8	0.877033	20	6	8	1
20	6	10	0.945122	20	6	10	1
20	7	4	0.652439	20	7	4	1
20	7	6	0.817073	20	7	6	1
20	7	8	0.933943	20	7	8	1
20	7	10	0.969512	20	7	10	1
20	8	4	0.712398	20	8	4	1
20	8	6	0.857724	20	8	6	1
20	8	8	0.953252	20	8	8	1
20	8	10	0.980691	20	8	10	1
20	9	4	0.793699	20	9	4	1
20	9	6	0.916667	20	9	6	1
20	9	8	0.961382	20	9	8	1
20	9	10	0.986789	20	9	10	1
20	10	4	0.816057	20	10	4	1
20	10	6	0.923781	20	10	6	1
20	10	8	0.972561	20	10	8	1
20	10	10	0.997968	20	10	10	1
30	6	4	0.779472	30	6	4	1

30	6	6	0.925813	30	6	6	1
30	6	8	0.962398	30	6	8	1
30	6	10	0.982724	30	6	10	1
30	7	4	0.802846	30	7	4	1
30	7	6	0.957317	30	7	6	1
30	7	8	0.981707	30	7	8	1
30	7	10	0.993902	30	7	10	1
30	8	4	0.865854	30	8	4	1
30	8	6	0.971545	30	8	6	1
30	8	8	0.992886	30	8	8	1
30	8	10	0.998984	30	8	10	1
30	9	4	0.906504	30	9	4	1
30	9	6	0.984756	30	9	6	1
30	9	8	0.993902	30	9	8	1
30	9	10	0.997968	30	9	10	1
30	10	4	0.938008	30	10	4	1
30	10	6	0.982724	30	10	6	1
30	10	8	0.998984	30	10	8	1
30	10	10	1	30	10	10	1
40	6	4	0.85874	40	6	4	1
40	6	6	0.962398	40	6	6	1
40	6	8	0.993902	40	6	8	1
40	6	10	1	40	6	10	1
40	7	4	0.908537	40	7	4	1
40	7	6	0.978659	40	7	6	1
40	7	8	0.993902	40	7	8	1
40	7	10	0.998984	40	7	10	1
40	8	4	0.946138	40	8	4	1
40	8	6	0.994919	40	8	6	1
40	8	8	0.998984	40	8	8	1
40	8	10	1	40	8	10	1
40	9	4	0.973577	40	9	4	1
40	9	6	0.995935	40	9	6	1
40	9	8	0.998984	40	9	8	1
40	9	10	1	40	9	10	1
40	10	4	0.976626	40	10	4	1
40	10	6	0.996951	40	10	6	1
40	10	8	1	40	10	8	1
40	10	10	1	40	10	10	1
50	6	4	0.928862	50	6	4	1
50	6	6	0.986789	50	6	6	1
50	6	8	0.995935	50	6	8	1
50	6	10	1	50	6	10	1
50	7	4	0.962398	50	7	4	1
50	7	6	0.992886	50	7	6	1
50	7	8	0.998984	50	7	8	1
50	7	10	1	50	7	10	1
50	8	4	0.981707	50	8	4	1
50	8	6	0.998984	50	8	6	1
50	8	8	1	50	8	8	1

50	8	10	1	50	8	10	1
50	9	4	0.98374	50	9	4	1
50	9	6	1	50	9	6	1
50	9	8	1	50	9	8	1
50	9	10	1	50	9	10	1
50	10	4	0.992886	50	10	4	1
50	10	6	0.998984	50	10	6	1
50	10	8	1	50	10	8	1
50	10	10	1	50	10	10	1
60	6	4	0.976626	60	6	4	1
60	6	6	0.996951	60	6	6	1
60	6	8	0.998984	60	6	8	1
60	6	10	1	60	6	10	1
60	7	4	0.979675	60	7	4	1
60	7	6	0.997968	60	7	6	1
60	7	8	1	60	7	8	1
60	7	10	1	60	7	10	1
60	8	4	0.993902	60	8	4	1
60	8	6	1	60	8	6	1
60	8	8	1	60	8	8	1
60	8	10	1	60	8	10	1
60	9	4	0.997968	60	9	4	1
60	9	6	1	60	9	6	1
60	9	8	1	60	9	8	1
60	9	10	1	60	9	10	1
60	10	4	0.996951	60	10	4	1
60	10	6	1	60	10	6	1
60	10	8	1	60	10	8	1
60	10	10	1	60	10	10	1

A 5.4 Simulated power table for association example: effect size 0.1; varying model variances; small predictor variances

Effect size $r = 0.1$ $\boldsymbol{\sigma_u^2} = \mathbf{2.34}$, $\sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				Effect size $r = 0.1$ $\sigma_u^2 = 0.26$, $\boldsymbol{\sigma_u^2} = \mathbf{2.34}$; $\sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				Effect size $r = 0.1$ $\sigma_u^2 = \sigma_v^2 = 0.26$; $\boldsymbol{\sigma_e^2} = \mathbf{2.34}$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
10	6	4	0.863821	10	6	4	0.797764	10	6	4	0.293699
10	6	6	0.955285	10	6	6	0.928862	10	6	6	0.369919
10	6	8	0.988821	10	6	8	0.982724	10	6	8	0.435976
10	6	10	0.997968	10	6	10	0.994919	10	6	10	0.511179
10	7	4	0.88313	10	7	4	0.816057	10	7	4	0.325203
10	7	6	0.968496	10	7	6	0.950203	10	7	6	0.403455
10	7	8	0.995935	10	7	8	0.988821	10	7	8	0.5
10	7	10	0.998984	10	7	10	0.998984	10	7	10	0.560976
10	8	4	0.919716	10	8	4	0.86687	10	8	4	0.368902
10	8	6	0.981707	10	8	6	0.97561	10	8	6	0.454268
10	8	8	0.998984	10	8	8	0.994919	10	8	8	0.531504
10	8	10	0.997968	10	8	10	0.997968	10	8	10	0.612805
10	9	4	0.957317	10	9	4	0.909553	10	9	4	0.406504
10	9	6	0.994919	10	9	6	0.987805	10	9	6	0.519309
10	9	8	0.998984	10	9	8	0.998984	10	9	8	0.585366
10	9	10	1	10	9	10	1	10	9	10	0.673781
10	10	4	0.953252	10	10	4	0.922764	10	10	4	0.423781
10	10	6	0.996951	10	10	6	0.990854	10	10	6	0.51626
10	10	8	1	10	10	8	0.998984	10	10	8	0.639228
10	10	10	1	10	10	10	1	10	10	10	0.699187
20	6	4	0.988821	20	6	4	0.965447	20	6	4	0.492886
20	6	6	0.998984	20	6	6	0.997968	20	6	6	0.609756
20	6	8	1	20	6	8	1	20	6	8	0.700203

20	6	10	1	20	6	10	1	20	6	10	0.788618
20	7	4	0.993902	20	7	4	0.986789	20	7	4	0.54065
20	7	6	1	20	7	6	1	20	7	6	0.65752
20	7	8	1	20	7	8	1	20	7	8	0.781504
20	7	10	1	20	7	10	1	20	7	10	0.828252
20	8	4	0.995935	20	8	4	0.990854	20	8	4	0.575203
20	8	6	1	20	8	6	1	20	8	6	0.685976
20	8	8	1	20	8	8	1	20	8	8	0.803862
20	8	10	1	20	8	10	1	20	8	10	0.880081
20	9	4	1	20	9	4	0.993902	20	9	4	0.640244
20	9	6	1	20	9	6	1	20	9	6	0.75813
20	9	8	1	20	9	8	1	20	9	8	0.84248
20	9	10	1	20	9	10	1	20	9	10	0.893293
20	10	4	0.997968	20	10	4	0.995935	20	10	4	0.688008
20	10	6	1	20	10	6	1	20	10	6	0.789634
20	10	8	1	20	10	8	1	20	10	8	0.887195
20	10	10	1	20	10	10	1	20	10	10	0.949187
30	6	4	0.998984	30	6	4	0.99187	30	6	4	0.652439
30	6	6	1	30	6	6	1	30	6	6	0.78252
30	6	8	1	30	6	8	1	30	6	8	0.846545
30	6	10	1	30	6	10	1	30	6	10	0.903455
30	7	4	0.998984	30	7	4	0.994919	30	7	4	0.662602
30	7	6	1	30	7	6	1	30	7	6	0.832317
30	7	8	1	30	7	8	0.998984	30	7	8	0.896341
30	7	10	1	30	7	10	1	30	7	10	0.934959
30	8	4	1	30	8	4	0.998984	30	8	4	0.730691
30	8	6	1	30	8	6	1	30	8	6	0.864837
30	8	8	1	30	8	8	1	30	8	8	0.933943
30	8	10	1	30	8	10	1	30	8	10	0.972561
30	9	4	1	30	9	4	0.998984	30	9	4	0.786585

30	9	6	1	30	9	6	1	30	9	6	0.919716
30	9	8	1	30	9	8	1	30	9	8	0.955285
30	9	10	1	30	9	10	1	30	9	10	0.979675
30	10	4	1	30	10	4	1	30	10	4	0.839431
30	10	6	1	30	10	6	1	30	10	6	0.924797
30	10	8	1	30	10	8	1	30	10	8	0.974594
30	10	10	1	30	10	10	1	30	10	10	0.990854
40	6	4	1	40	6	4	1	40	6	4	0.757114
40	6	6	1	40	6	6	1	40	6	6	0.855691
40	6	8	1	40	6	8	1	40	6	8	0.932927
40	6	10	1	40	6	10	1	40	6	10	0.97561
40	7	4	1	40	7	4	1	40	7	4	0.798781
40	7	6	1	40	7	6	1	40	7	6	0.918699
40	7	8	1	40	7	8	1	40	7	8	0.960366
40	7	10	1	40	7	10	1	40	7	10	0.976626
40	8	4	1	40	8	4	1	40	8	4	0.857724
40	8	6	1	40	8	6	1	40	8	6	0.95122
40	8	8	1	40	8	8	1	40	8	8	0.981707
40	8	10	1	40	8	10	1	40	8	10	0.990854
40	9	4	1	40	9	4	1	40	9	4	0.898374
40	9	6	1	40	9	6	1	40	9	6	0.958333
40	9	8	1	40	9	8	1	40	9	8	0.984756
40	9	10	1	40	9	10	1	40	9	10	0.994919
40	10	4	1	40	10	4	1	40	10	4	0.912602
40	10	6	1	40	10	6	1	40	10	6	0.969512
40	10	8	1	40	10	8	1	40	10	8	0.993902
40	10	10	1	40	10	10	1	40	10	10	0.997968
50	6	4	1	50	6	4	1	50	6	4	0.824187
50	6	6	1	50	6	6	1	50	6	6	0.946138
50	6	8	1	50	6	8	1	50	6	8	0.97561

50	6	10	1	50	6	10	1	50	6	10	0.990854
50	7	4	1	50	7	4	1	50	7	4	0.892276
50	7	6	1	50	7	6	1	50	7	6	0.965447
50	7	8	1	50	7	8	1	50	7	8	0.985772
50	7	10	1	50	7	10	1	50	7	10	0.996951
50	8	4	1	50	8	4	1	50	8	4	0.933943
50	8	6	1	50	8	6	1	50	8	6	0.979675
50	8	8	1	50	8	8	1	50	8	8	0.990854
50	8	10	1	50	8	10	1	50	8	10	0.997968
50	9	4	1	50	9	4	1	50	9	4	0.943089
50	9	6	1	50	9	6	1	50	9	6	0.988821
50	9	8	1	50	9	8	1	50	9	8	0.997968
50	9	10	1	50	9	10	1	50	9	10	1
50	10	4	1	50	10	4	1	50	10	4	0.95935
50	10	6	1	50	10	6	1	50	10	6	0.994919
50	10	8	1	50	10	8	1	50	10	8	0.995935
50	10	10	1	50	10	10	1	50	10	10	1
60	6	4	1	60	6	4	1	60	6	4	0.89126
60	6	6	1	60	6	6	1	60	6	6	0.958333
60	6	8	1	60	6	8	1	60	6	8	0.987805
60	6	10	1	60	6	10	1	60	6	10	0.997968
60	7	4	1	60	7	4	1	60	7	4	0.917683
60	7	6	1	60	7	6	1	60	7	6	0.97561
60	7	8	1	60	7	8	1	60	7	8	0.992886
60	7	10	1	60	7	10	1	60	7	10	0.998984
60	8	4	1	60	8	4	1	60	8	4	0.956301
60	8	6	1	60	8	6	1	60	8	6	0.989837
60	8	8	1	60	8	8	1	60	8	8	0.998984
60	8	10	1	60	8	10	1	60	8	10	1
60	9	4	1	60	9	4	1	60	9	4	0.970529

60	9	6	1	60	9	6	1	60	9	6	0.997968
60	9	8	1	60	9	8	1	60	9	8	1
60	9	10	1	60	9	10	1	60	9	10	1
60	10	4	1	60	10	4	1	60	10	4	0.988821
60	10	6	1	60	10	6	1	60	10	6	0.998984
60	10	8	1	60	10	8	1	60	10	8	1
60	10	10	1	60	10	10	1	60	10	10	1

A 5.5 Simulated power table for association example: effect size 0.1; small model variances; vary predictor variances

Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\boldsymbol{\sigma_{xL3}^2} = \mathbf{2.34}$; $\sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = 0.26$; $\boldsymbol{\sigma_{xL2}^2} = \mathbf{2.34}$; $\sigma_{xL1}^2 = 0.26$				Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = 0.26$; $\boldsymbol{\sigma_{xL1}^2} = \mathbf{2.34}$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
10	6	4	0.211382	10	6	4	0.25813	10	6	4	0.585366
10	6	6	0.259146	10	6	6	0.296748	10	6	6	0.815041
10	6	8	0.284553	10	6	8	0.335366	10	6	8	0.919716
10	6	10	0.314024	10	6	10	0.384146	10	6	10	0.955285
10	7	4	0.218496	10	7	4	0.295732	10	7	4	0.651423
10	7	6	0.263211	10	7	6	0.35874	10	7	6	0.843496
10	7	8	0.32622	10	7	8	0.405488	10	7	8	0.935976
10	7	10	0.360772	10	7	10	0.441057	10	7	10	0.980691
10	8	4	0.25	10	8	4	0.332317	10	8	4	0.723577
10	8	6	0.286585	10	8	6	0.395325	10	8	6	0.903455
10	8	8	0.335366	10	8	8	0.440041	10	8	8	0.964431
10	8	10	0.381098	10	8	10	0.490854	10	8	10	0.981707
10	9	4	0.238821	10	9	4	0.355691	10	9	4	0.793699
10	9	6	0.337398	10	9	6	0.443089	10	9	6	0.935976
10	9	8	0.382114	10	9	8	0.470529	10	9	8	0.978659
10	9	10	0.450203	10	9	10	0.54065	10	9	10	0.996951
10	10	4	0.25813	10	10	4	0.39126	10	10	4	0.798781
10	10	6	0.316057	10	10	6	0.448171	10	10	6	0.936992
10	10	8	0.417683	10	10	8	0.517276	10	10	8	0.990854
10	10	10	0.462398	10	10	10	0.566057	10	10	10	1
20	6	4	0.286585	20	6	4	0.442073	20	6	4	0.880081
20	6	6	0.365854	20	6	6	0.49187	20	6	6	0.973577
20	6	8	0.448171	20	6	8	0.580285	20	6	8	0.996951
20	6	10	0.531504	20	6	10	0.646341	20	6	10	1

20	7	4	0.309959	20	7	4	0.492886	20	7	4	0.920732
20	7	6	0.403455	20	7	6	0.575203	20	7	6	0.990854
20	7	8	0.5	20	7	8	0.64939	20	7	8	1
20	7	10	0.574187	20	7	10	0.713415	20	7	10	1
20	8	4	0.35874	20	8	4	0.544716	20	8	4	0.949187
20	8	6	0.440041	20	8	6	0.618902	20	8	6	0.994919
20	8	8	0.548781	20	8	8	0.716463	20	8	8	0.998984
20	8	10	0.611789	20	8	10	0.762195	20	8	10	1
20	9	4	0.365854	20	9	4	0.610772	20	9	4	0.96748
20	9	6	0.489837	20	9	6	0.70122	20	9	6	0.998984
20	9	8	0.57622	20	9	8	0.748984	20	9	8	1
20	9	10	0.658537	20	9	10	0.819106	20	9	10	1
20	10	4	0.401423	20	10	4	0.655488	20	10	4	0.973577
20	10	6	0.513211	20	10	6	0.720529	20	10	6	1
20	10	8	0.639228	20	10	8	0.792683	20	10	8	1
20	10	10	0.713415	20	10	10	0.877033	20	10	10	1
30	6	4	0.385163	30	6	4	0.58435	30	6	4	0.964431
30	6	6	0.498984	30	6	6	0.694106	30	6	6	0.998984
30	6	8	0.569106	30	6	8	0.726626	30	6	8	1
30	6	10	0.658537	30	6	10	0.796748	30	6	10	1
30	7	4	0.417683	30	7	4	0.613821	30	7	4	0.979675
30	7	6	0.533537	30	7	6	0.765244	30	7	6	0.998984
30	7	8	0.64939	30	7	8	0.821138	30	7	8	0.998984
30	7	10	0.707317	30	7	10	0.848577	30	7	10	1
30	8	4	0.442073	30	8	4	0.726626	30	8	4	0.990854
30	8	6	0.571138	30	8	6	0.802846	30	8	6	1
30	8	8	0.713415	30	8	8	0.865854	30	8	8	1
30	8	10	0.799797	30	8	10	0.914634	30	8	10	1
30	9	4	0.497968	30	9	4	0.762195	30	9	4	0.99187
30	9	6	0.593496	30	9	6	0.853659	30	9	6	1

30	9	8	0.713415	30	9	8	0.897358	30	9	8	1
30	9	10	0.822155	30	9	10	0.943089	30	9	10	1
30	10	4	0.530488	30	10	4	0.819106	30	10	4	0.998984
30	10	6	0.668699	30	10	6	0.886179	30	10	6	1
30	10	8	0.791667	30	10	8	0.927846	30	10	8	1
30	10	10	0.85061	30	10	10	0.966463	30	10	10	1
40	6	4	0.454268	40	6	4	0.707317	40	6	4	0.989837
40	6	6	0.580285	40	6	6	0.787602	40	6	6	1
40	6	8	0.705285	40	6	8	0.847561	40	6	8	1
40	6	10	0.798781	40	6	10	0.90752	40	6	10	1
40	7	4	0.512195	40	7	4	0.765244	40	7	4	0.998984
40	7	6	0.670732	40	7	6	0.86687	40	7	6	1
40	7	8	0.744919	40	7	8	0.902439	40	7	8	1
40	7	10	0.818089	40	7	10	0.943089	40	7	10	1
40	8	4	0.555894	40	8	4	0.833333	40	8	4	1
40	8	6	0.727642	40	8	6	0.920732	40	8	6	1
40	8	8	0.809959	40	8	8	0.952236	40	8	8	1
40	8	10	0.877033	40	8	10	0.966463	40	8	10	1
40	9	4	0.601626	40	9	4	0.876016	40	9	4	1
40	9	6	0.724594	40	9	6	0.924797	40	9	6	1
40	9	8	0.83435	40	9	8	0.954268	40	9	8	1
40	9	10	0.906504	40	9	10	0.978659	40	9	10	1
40	10	4	0.634146	40	10	4	0.906504	40	10	4	1
40	10	6	0.781504	40	10	6	0.952236	40	10	6	1
40	10	8	0.868902	40	10	8	0.974594	40	10	8	1
40	10	10	0.940041	40	10	10	0.987805	40	10	10	1
50	6	4	0.535569	50	6	4	0.784553	50	6	4	1
50	6	6	0.705285	50	6	6	0.889228	50	6	6	1
50	6	8	0.802846	50	6	8	0.931911	50	6	8	1
50	6	10	0.859756	50	6	10	0.957317	50	6	10	1

50	7	4	0.606707	50	7	4	0.868902	50	7	4	1
50	7	6	0.74187	50	7	6	0.923781	50	7	6	1
50	7	8	0.851626	50	7	8	0.958333	50	7	8	1
50	7	10	0.91565	50	7	10	0.980691	50	7	10	1
50	8	4	0.640244	50	8	4	0.920732	50	8	4	1
50	8	6	0.778455	50	8	6	0.946138	50	8	6	1
50	8	8	0.896341	50	8	8	0.973577	50	8	8	1
50	8	10	0.930894	50	8	10	0.985772	50	8	10	1
50	9	4	0.684959	50	9	4	0.926829	50	9	4	1
50	9	6	0.830285	50	9	6	0.97561	50	9	6	1
50	9	8	0.911585	50	9	8	0.985772	50	9	8	1
50	9	10	0.955285	50	9	10	0.993902	50	9	10	1
50	10	4	0.730691	50	10	4	0.947155	50	10	4	1
50	10	6	0.860772	50	10	6	0.984756	50	10	6	1
50	10	8	0.945122	50	10	8	0.988821	50	10	8	1
50	10	10	0.964431	50	10	10	0.998984	50	10	10	1
60	6	4	0.601626	60	6	4	0.859756	60	6	4	1
60	6	6	0.740854	60	6	6	0.924797	60	6	6	1
60	6	8	0.862805	60	6	8	0.971545	60	6	8	1
60	6	10	0.909553	60	6	10	0.974594	60	6	10	1
60	7	4	0.652439	60	7	4	0.909553	60	7	4	0.998984
60	7	6	0.829268	60	7	6	0.95935	60	7	6	1
60	7	8	0.896341	60	7	8	0.978659	60	7	8	1
60	7	10	0.942073	60	7	10	0.995935	60	7	10	1
60	8	4	0.724594	60	8	4	0.930894	60	8	4	1
60	8	6	0.864837	60	8	6	0.976626	60	8	6	1
60	8	8	0.931911	60	8	8	0.988821	60	8	8	1
60	8	10	0.969512	60	8	10	0.998984	60	8	10	1
60	9	4	0.778455	60	9	4	0.965447	60	9	4	1
60	9	6	0.897358	60	9	6	0.989837	60	9	6	1

60	9	8	0.930894	60	9	8	0.996951	60	9	8	1
60	9	10	0.977642	60	9	10	0.996951	60	9	10	1
60	10	4	0.786585	60	10	4	0.980691	60	10	4	1
60	10	6	0.906504	60	10	6	0.995935	60	10	6	1
60	10	8	0.972561	60	10	8	0.995935	60	10	8	1
60	10	10	0.986789	60	10	10	0.998984	60	10	10	1

A 5.6 Simulated power table for association example: Missing data; effect size 0.1; small model variances; small predictor variances

20 % Missing data Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				70 % Missing data Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$				Missing data according to time trend $\boldsymbol{t} + \boldsymbol{t}^{\boldsymbol{2}} + s$ Effect size $r = 0.1$, $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$ $\sigma_{xL3}^2 = \sigma_{xL2}^2 = \sigma_{xL1}^2 = 0.26$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
10	6	4	0.333333	10	6	4	0.183943	10	6	4	0.279472
10	6	6	0.405488	10	6	6	0.229675	10	6	6	0.384146
10	6	8	0.527439	10	6	8	0.27439	10	6	8	0.493902
10	6	10	0.631098	10	6	10	0.328252	10	6	10	0.605691
10	7	4	0.327236	10	7	4	0.160569	10	7	4	0.272358
10	7	6	0.461382	10	7	6	0.235772	10	7	6	0.403455
10	7	8	0.575203	10	7	8	0.316057	10	7	8	0.546748
10	7	10	0.684959	10	7	10	0.367886	10	7	10	0.662602
10	8	4	0.388211	10	8	4	0.191057	10	8	4	0.313008
10	8	6	0.526423	10	8	6	0.255081	10	8	6	0.465447
10	8	8	0.632114	10	8	8	0.302846	10	8	8	0.58435
10	8	10	0.698171	10	8	10	0.368902	10	8	10	0.653455
10	9	4	0.408537	10	9	4	0.210366	10	9	4	0.295732
10	9	6	0.546748	10	9	6	0.275407	10	9	6	0.497968
10	9	8	0.700203	10	9	8	0.310976	10	9	8	0.596545
10	9	10	0.771341	10	9	10	0.39126	10	9	10	0.689024
10	10	4	0.469512	10	10	4	0.246951	10	10	4	0.344512
10	10	6	0.59248	10	10	6	0.269309	10	10	6	0.450203
10	10	8	0.738821	10	10	8	0.367886	10	10	8	0.610772
10	10	10	0.804878	10	10	10	0.445122	10	10	10	0.728659
20	6	4	0.488821	20	6	4	0.260163	20	6	4	0.440041

20	6	6	0.684959	20	6	6	0.300813	20	6	6	0.655488
20	6	8	0.808943	20	6	8	0.404472	20	6	8	0.75813
20	6	10	0.881098	20	6	10	0.506098	20	6	10	0.864837
20	7	4	0.578252	20	7	4	0.286585	20	7	4	0.509146
20	7	6	0.75813	20	7	6	0.373984	20	7	6	0.698171
20	7	8	0.85874	20	7	8	0.473577	20	7	8	0.827236
20	7	10	0.917683	20	7	10	0.531504	20	7	10	0.878049
20	8	4	0.619919	20	8	4	0.314024	20	8	4	0.528455
20	8	6	0.789634	20	8	6	0.420732	20	8	6	0.727642
20	8	8	0.910569	20	8	8	0.536585	20	8	8	0.857724
20	8	10	0.944106	20	8	10	0.594512	20	8	10	0.927846
20	9	4	0.688008	20	9	4	0.346545	20	9	4	0.544716
20	9	6	0.851626	20	9	6	0.452236	20	9	6	0.747968
20	9	8	0.924797	20	9	8	0.572155	20	9	8	0.864837
20	9	10	0.962398	20	9	10	0.63313	20	9	10	0.912602
20	10	4	0.74187	20	10	4	0.363821	20	10	4	0.579268
20	10	6	0.873984	20	10	6	0.50813	20	10	6	0.757114
20	10	8	0.955285	20	10	8	0.610772	20	10	8	0.876016
20	10	10	0.981707	20	10	10	0.710366	20	10	10	0.955285
30	6	4	0.695122	30	6	4	0.35061	30	6	4	0.610772
30	6	6	0.828252	30	6	6	0.438008	30	6	6	0.793699
30	6	8	0.917683	30	6	8	0.549797	30	6	8	0.904472
30	6	10	0.966463	30	6	10	0.654472	30	6	10	0.956301
30	7	4	0.727642	30	7	4	0.329268	30	7	4	0.642276
30	7	6	0.894309	30	7	6	0.492886	30	7	6	0.852642
30	7	8	0.960366	30	7	8	0.613821	30	7	8	0.939024
30	7	10	0.980691	30	7	10	0.737805	30	7	10	0.96748
30	8	4	0.828252	30	8	4	0.416667	30	8	4	0.70935
30	8	6	0.925813	30	8	6	0.543699	30	8	6	0.865854
30	8	8	0.969512	30	8	8	0.667683	30	8	8	0.950203

30	8	10	0.995935	30	8	10	0.775407	30	8	10	0.985772
30	9	4	0.843496	30	9	4	0.460366	30	9	4	0.699187
30	9	6	0.958333	30	9	6	0.596545	30	9	6	0.909553
30	9	8	0.985772	30	9	8	0.74187	30	9	8	0.965447
30	9	10	0.995935	30	9	10	0.794716	30	9	10	0.985772
30	10	4	0.869919	30	10	4	0.495935	30	10	4	0.718496
30	10	6	0.969512	30	10	6	0.652439	30	10	6	0.901423
30	10	8	0.995935	30	10	8	0.789634	30	10	8	0.969512
30	10	10	0.997968	30	10	10	0.832317	30	10	10	0.989837
40	6	4	0.817073	40	6	4	0.406504	40	6	4	0.718496
40	6	6	0.918699	40	6	6	0.553862	40	6	6	0.886179
40	6	8	0.974594	40	6	8	0.650407	40	6	8	0.973577
40	6	10	0.994919	40	6	10	0.768293	40	6	10	0.993902
40	7	4	0.844512	40	7	4	0.449187	40	7	4	0.752033
40	7	6	0.954268	40	7	6	0.615854	40	7	6	0.929878
40	7	8	0.984756	40	7	8	0.769309	40	7	8	0.985772
40	7	10	0.993902	40	7	10	0.830285	40	7	10	0.993902
40	8	4	0.895325	40	8	4	0.510163	40	8	4	0.797764
40	8	6	0.974594	40	8	6	0.676829	40	8	6	0.936992
40	8	8	0.992886	40	8	8	0.797764	40	8	8	0.99187
40	8	10	1	40	8	10	0.881098	40	8	10	0.995935
40	9	4	0.922764	40	9	4	0.544716	40	9	4	0.818089
40	9	6	0.986789	40	9	6	0.726626	40	9	6	0.95935
40	9	8	0.998984	40	9	8	0.838415	40	9	8	0.989837
40	9	10	1	40	9	10	0.917683	40	9	10	0.998984
40	10	4	0.948171	40	10	4	0.593496	40	10	4	0.82622
40	10	6	0.993902	40	10	6	0.765244	40	10	6	0.974594
40	10	8	0.998984	40	10	8	0.85874	40	10	8	0.99187
40	10	10	1	40	10	10	0.921748	40	10	10	1
50	6	4	0.869919	50	6	4	0.48374	50	6	4	0.807927

50	6	6	0.95935	50	6	6	0.645325	50	6	6	0.952236
50	6	8	0.990854	50	6	8	0.752033	50	6	8	0.990854
50	6	10	1	50	6	10	0.871951	50	6	10	0.992886
50	7	4	0.918699	50	7	4	0.558943	50	7	4	0.847561
50	7	6	0.98374	50	7	6	0.736789	50	7	6	0.96748
50	7	8	0.998984	50	7	8	0.832317	50	7	8	0.997968
50	7	10	1	50	7	10	0.91565	50	7	10	0.996951
50	8	4	0.954268	50	8	4	0.596545	50	8	4	0.86687
50	8	6	0.996951	50	8	6	0.768293	50	8	6	0.987805
50	8	8	0.998984	50	8	8	0.86687	50	8	8	0.998984
50	8	10	1	50	8	10	0.916667	50	8	10	1
50	9	4	0.969512	50	9	4	0.66565	50	9	4	0.892276
50	9	6	0.992886	50	9	6	0.819106	50	9	6	0.98374
50	9	8	1	50	9	8	0.91565	50	9	8	1
50	9	10	1	50	9	10	0.953252	50	9	10	1
50	10	4	0.980691	50	10	4	0.692073	50	10	4	0.91565
50	10	6	0.997968	50	10	6	0.871951	50	10	6	0.987805
50	10	8	1	50	10	8	0.935976	50	10	8	0.996951
50	10	10	1	50	10	10	0.977642	50	10	10	1
60	6	4	0.922764	60	6	4	0.586382	60	6	4	0.88313
60	6	6	0.987805	60	6	6	0.734756	60	6	6	0.973577
60	6	8	0.994919	60	6	8	0.85061	60	6	8	0.996951
60	6	10	1	60	6	10	0.910569	60	6	10	1
60	7	4	0.966463	60	7	4	0.607724	60	7	4	0.914634
60	7	6	0.994919	60	7	6	0.792683	60	7	6	0.989837
60	7	8	1	60	7	8	0.890244	60	7	8	1
60	7	10	1	60	7	10	0.947155	60	7	10	1
60	8	4	0.976626	60	8	4	0.690041	60	8	4	0.933943
60	8	6	0.997968	60	8	6	0.83435	60	8	6	0.992886
60	8	8	1	60	8	8	0.914634	60	8	8	1

60	8	10	1	60	8	10	0.963415	60	8	10	1
60	9	4	0.990854	60	9	4	0.716463	60	9	4	0.941057
60	9	6	1	60	9	6	0.882114	60	9	6	0.993902
60	9	8	1	60	9	8	0.943089	60	9	8	0.997968
60	9	10	1	60	9	10	0.980691	60	9	10	1
60	10	4	0.995935	60	10	4	0.772358	60	10	4	0.952236
60	10	6	0.998984	60	10	6	0.911585	60	10	6	0.994919
60	10	8	1	60	10	8	0.972561	60	10	8	1
60	10	10	1	60	10	10	0.992886	60	10	10	1

A 5.7 Simulated power table for group difference example: effect sizes 0.2, 0.5, 0.8; small model variances

Effect size $\boldsymbol{d = 0.2}$ $\sigma_u^2 = \sigma_{\mathrm{v}}^2 = \sigma_e^2 = 0.26$				**Effect size $\boldsymbol{d = 0.5}$** $\sigma_u^2 = \sigma_{\mathrm{v}}^2 = \sigma_e^2 = 0.26$				**Effect size $\boldsymbol{d = 0.8}$** $\sigma_u^2 = \sigma_{\mathrm{v}}^2 = \sigma_e^2 = 0.26$							
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
10	6	4	0.103659	10	6	4	0.140244	10	6	4	0.210366	70	6	4	0.719512
10	6	6	0.137195	10	6	6	0.167683	10	6	6	0.248984	70	6	6	0.70935
10	6	8	0.117886	10	6	8	0.15752	10	6	8	0.23374	70	6	8	0.719512
10	6	10	0.112805	10	6	10	0.164634	10	6	10	0.248984	70	6	10	0.680894
10	7	4	0.117886	10	7	4	0.152439	10	7	4	0.228659	70	7	4	0.723577
10	7	6	0.135163	10	7	6	0.158537	10	7	6	0.214431	70	7	6	0.724594
10	7	8	0.118902	10	7	8	0.16565	10	7	8	0.246951	70	7	8	0.734756
10	7	10	0.126016	10	7	10	0.161585	10	7	10	0.230691	70	7	10	0.726626
10	8	4	0.114837	10	8	4	0.152439	10	8	4	0.205285	70	8	4	0.728659
10	8	6	0.117886	10	8	6	0.167683	10	8	6	0.252033	70	8	6	0.705285
10	8	8	0.127033	10	8	8	0.189024	10	8	8	0.269309	70	8	8	0.713415
10	8	10	0.123984	10	8	10	0.189024	10	8	10	0.243902	70	8	10	0.730691
10	9	4	0.125	10	9	4	0.162602	10	9	4	0.231707	70	9	4	0.70122
10	9	6	0.118902	10	9	6	0.162602	10	9	6	0.235772	70	9	6	0.720529
10	9	8	0.150407	10	9	8	0.20122	10	9	8	0.261179	70	9	8	0.718496
10	9	10	0.113821	10	9	10	0.172764	10	9	10	0.259146	70	9	10	0.743902
10	10	4	0.112805	10	10	4	0.160569	10	10	4	0.234756	70	10	4	0.731707
10	10	6	0.113821	10	10	6	0.161585	10	10	6	0.238821	70	10	6	0.724594
10	10	8	0.121951	10	10	8	0.169715	10	10	8	0.25813	70	10	8	0.753049
10	10	10	0.125	10	10	10	0.181911	10	10	10	0.255081	70	10	10	0.723577
20	6	4	0.098577	20	6	4	0.182927	20	6	4	0.302846	80	6	4	0.751016
20	6	6	0.09248	20	6	6	0.195122	20	6	6	0.331301	80	6	6	0.754065
20	6	8	0.09248	20	6	8	0.162602	20	6	8	0.297764	80	6	8	0.771341
20	6	10	0.104675	20	6	10	0.182927	20	6	10	0.319106	80	6	10	0.779472

20	7	4	0.09248	20	7	4	0.178862	20	7	4	0.315041	80	7	4	0.781504
20	7	6	0.106707	20	7	6	0.175813	20	7	6	0.323171	80	7	6	0.778455
20	7	8	0.098577	20	7	8	0.203252	20	7	8	0.331301	80	7	8	0.762195
20	7	10	0.111789	20	7	10	0.188008	20	7	10	0.317073	80	7	10	0.753049
20	8	4	0.105691	20	8	4	0.189024	20	8	4	0.309959	80	8	4	0.754065
20	8	6	0.093496	20	8	6	0.177846	20	8	6	0.318089	80	8	6	0.770325
20	8	8	0.094512	20	8	8	0.184959	20	8	8	0.294715	80	8	8	0.779472
20	8	10	0.096545	20	8	10	0.194106	20	8	10	0.34248	80	8	10	0.781504
20	9	4	0.096545	20	9	4	0.191057	20	9	4	0.347561	80	9	4	0.775407
20	9	6	0.099594	20	9	6	0.199187	20	9	6	0.340447	80	9	6	0.768293
20	9	8	0.077236	20	9	8	0.155488	20	9	8	0.292683	80	9	8	0.788618
20	9	10	0.090447	20	9	10	0.158537	20	9	10	0.286585	80	9	10	0.745935
20	10	4	0.077236	20	10	4	0.169715	20	10	4	0.298781	80	10	4	0.793699
20	10	6	0.093496	20	10	6	0.183943	20	10	6	0.317073	80	10	6	0.776423
20	10	8	0.106707	20	10	8	0.20935	20	10	8	0.327236	80	10	8	0.799797
20	10	10	0.09248	20	10	10	0.185976	20	10	10	0.329268	80	10	10	0.756098
30	6	4	0.079268	30	6	4	0.20935	30	6	4	0.398374	90	6	4	0.788618
30	6	6	0.075203	30	6	6	0.191057	30	6	6	0.387195	90	6	6	0.776423
30	6	8	0.088415	30	6	8	0.223577	30	6	8	0.393293	90	6	8	0.802846
30	6	10	0.083333	30	6	10	0.224594	30	6	10	0.403455	90	6	10	0.765244
30	7	4	0.08435	30	7	4	0.204268	30	7	4	0.390244	90	7	4	0.804878
30	7	6	0.082317	30	7	6	0.192073	30	7	6	0.392276	90	7	6	0.818089
30	7	8	0.09248	30	7	8	0.224594	30	7	8	0.419715	90	7	8	0.798781
30	7	10	0.10061	30	7	10	0.219512	30	7	10	0.409553	90	7	10	0.828252
30	8	4	0.085366	30	8	4	0.212398	30	8	4	0.418699	90	8	4	0.813008
30	8	6	0.091463	30	8	6	0.22561	30	8	6	0.427846	90	8	6	0.837398
30	8	8	0.090447	30	8	8	0.211382	30	8	8	0.405488	90	8	8	0.827236
30	8	10	0.095529	30	8	10	0.213415	30	8	10	0.404472	90	8	10	0.832317
30	9	4	0.096545	30	9	4	0.206301	30	9	4	0.405488	90	9	4	0.804878
30	9	6	0.087398	30	9	6	0.238821	30	9	6	0.405488	90	9	6	0.813008

30	9	8	0.094512	30	9	8	0.22561	30	9	8	0.445122	90	9	8	0.819106
30	9	10	0.080285	30	9	10	0.228659	30	9	10	0.419715	90	9	10	0.837398
30	10	4	0.096545	30	10	4	0.232724	30	10	4	0.422764	90	10	4	0.821138
30	10	6	0.099594	30	10	6	0.242886	30	10	6	0.41565	90	10	6	0.802846
30	10	8	0.090447	30	10	8	0.210366	30	10	8	0.395325	90	10	8	0.836382
30	10	10	0.08435	30	10	10	0.223577	30	10	10	0.402439	90	10	10	0.804878
40	6	4	0.085366	40	6	4	0.221545	40	6	4	0.463415	100	6	4	0.83435
40	6	6	0.105691	40	6	6	0.239837	40	6	6	0.462398	100	6	6	0.853659
40	6	8	0.10061	40	6	8	0.268293	40	6	8	0.474594	100	6	8	0.833333
40	6	10	0.095529	40	6	10	0.252033	40	6	10	0.47561	100	6	10	0.818089
40	7	4	0.114837	40	7	4	0.270325	40	7	4	0.480691	100	7	4	0.848577
40	7	6	0.10061	40	7	6	0.259146	40	7	6	0.511179	100	7	6	0.853659
40	7	8	0.095529	40	7	8	0.260163	40	7	8	0.493902	100	7	8	0.84248
40	7	10	0.087398	40	7	10	0.246951	40	7	10	0.497968	100	7	10	0.85874
40	8	4	0.091463	40	8	4	0.248984	40	8	4	0.492886	100	8	4	0.839431
40	8	6	0.102642	40	8	6	0.255081	40	8	6	0.517276	100	8	6	0.856707
40	8	8	0.083333	40	8	8	0.240854	40	8	8	0.496951	100	8	8	0.872968
40	8	10	0.088415	40	8	10	0.28252	40	8	10	0.506098	100	8	10	0.861789
40	9	4	0.090447	40	9	4	0.260163	40	9	4	0.489837	100	9	4	0.855691
40	9	6	0.086382	40	9	6	0.25	40	9	6	0.505081	100	9	6	0.847561
40	9	8	0.094512	40	9	8	0.25813	40	9	8	0.501016	100	9	8	0.873984
40	9	10	0.107724	40	9	10	0.264228	40	9	10	0.503049	100	9	10	0.873984
40	10	4	0.093496	40	10	4	0.25813	40	10	4	0.51626	100	10	4	0.846545
40	10	6	0.083333	40	10	6	0.245935	40	10	6	0.477642	100	10	6	0.863821
40	10	8	0.086382	40	10	8	0.264228	40	10	8	0.525407	100	10	8	0.85874
40	10	10	0.089431	40	10	10	0.26626	40	10	10	0.505081	100	10	10	0.869919
50	6	4	0.104675	50	6	4	0.300813	50	6	4	0.557927				
50	6	6	0.106707	50	6	6	0.284553	50	6	6	0.571138				
50	6	8	0.10874	50	6	8	0.284553	50	6	8	0.574187				
50	6	10	0.109756	50	6	10	0.303862	50	6	10	0.560976				

50	7	4	0.098577	50	7	4	0.310976	50	7	4	0.570122
50	7	6	0.111789	50	7	6	0.329268	50	7	6	0.596545
50	7	8	0.093496	50	7	8	0.294715	50	7	8	0.591463
50	7	10	0.106707	50	7	10	0.303862	50	7	10	0.601626
50	8	4	0.10061	50	8	4	0.33435	50	8	4	0.60061
50	8	6	0.09248	50	8	6	0.292683	50	8	6	0.590447
50	8	8	0.107724	50	8	8	0.304878	50	8	8	0.57622
50	8	10	0.106707	50	8	10	0.317073	50	8	10	0.597561
50	9	4	0.103659	50	9	4	0.300813	50	9	4	0.564024
50	9	6	0.091463	50	9	6	0.310976	50	9	6	0.613821
50	9	8	0.113821	50	9	8	0.308943	50	9	8	0.602642
50	9	10	0.095529	50	9	10	0.279472	50	9	10	0.561992
50	10	4	0.086382	50	10	4	0.294715	50	10	4	0.581301
50	10	6	0.109756	50	10	6	0.300813	50	10	6	0.59248
50	10	8	0.102642	50	10	8	0.304878	50	10	8	0.579268
50	10	10	0.106707	50	10	10	0.305894	50	10	10	0.559959
60	6	4	0.10061	60	6	4	0.325203	60	6	4	0.625
60	6	6	0.106707	60	6	6	0.347561	60	6	6	0.645325
60	6	8	0.105691	60	6	8	0.346545	60	6	8	0.663618
60	6	10	0.089431	60	6	10	0.332317	60	6	10	0.651423
60	7	4	0.121951	60	7	4	0.388211	60	7	4	0.662602
60	7	6	0.10874	60	7	6	0.359756	60	7	6	0.674797
60	7	8	0.097561	60	7	8	0.35061	60	7	8	0.654472
60	7	10	0.107724	60	7	10	0.359756	60	7	10	0.647358
60	8	4	0.096545	60	8	4	0.339431	60	8	4	0.646341
60	8	6	0.103659	60	8	6	0.352642	60	8	6	0.655488
60	8	8	0.121951	60	8	8	0.35874	60	8	8	0.65752
60	8	10	0.112805	60	8	10	0.367886	60	8	10	0.664634
60	9	4	0.105691	60	9	4	0.339431	60	9	4	0.645325
60	9	6	0.107724	60	9	6	0.343496	60	9	6	0.652439

60	9	8	0.119919	60	9	8	0.367886	60	9	8	0.674797
60	9	10	0.110772	60	9	10	0.344512	60	9	10	0.669716
60	10	4	0.113821	60	10	4	0.377033	60	10	4	0.671748
60	10	6	0.105691	60	10	6	0.365854	60	10	6	0.658537
60	10	8	0.109756	60	10	8	0.368902	60	10	8	0.672764
60	10	10	0.105691	60	10	10	0.353659	60	10	10	0.700203

A 5.8 Simulated power table for group difference example: effect size 0.8; varying model variances

Effect size $d = 0.8$ $\boldsymbol{\sigma_u^2} = \mathbf{2.34}$, $\sigma_v^2 = \sigma_e^2 = 0.26$				Effect size $d = 0.8$ $\sigma_u^2 = 0.26$, $\boldsymbol{\sigma_u^2} = \mathbf{2.34}$; $\sigma_e^2 = 0.26$				Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = 0.26$; $\boldsymbol{\sigma_e^2} = \mathbf{2.34}$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
10	6	4	0.118902	10	6	4	0.121951	10	6	4	0.190041
10	6	6	0.151423	10	6	6	0.138211	10	6	6	0.226626
10	6	8	0.120935	10	6	8	0.130081	10	6	8	0.207317
10	6	10	0.123984	10	6	10	0.154472	10	6	10	0.219512
10	7	4	0.118902	10	7	4	0.143293	10	7	4	0.190041
10	7	6	0.138211	10	7	6	0.163618	10	7	6	0.191057
10	7	8	0.135163	10	7	8	0.143293	10	7	8	0.227642
10	7	10	0.137195	10	7	10	0.142276	10	7	10	0.21748
10	8	4	0.136179	10	8	4	0.146342	10	8	4	0.196138
10	8	6	0.131098	10	8	6	0.160569	10	8	6	0.230691
10	8	8	0.147358	10	8	8	0.167683	10	8	8	0.253049
10	8	10	0.154472	10	8	10	0.167683	10	8	10	0.232724
10	9	4	0.128049	10	9	4	0.159553	10	9	4	0.212398
10	9	6	0.127033	10	9	6	0.145325	10	9	6	0.223577
10	9	8	0.15752	10	9	8	0.188008	10	9	8	0.254065
10	9	10	0.127033	10	9	10	0.163618	10	9	10	0.253049
10	10	4	0.115854	10	10	4	0.174797	10	10	4	0.204268
10	10	6	0.112805	10	10	6	0.183943	10	10	6	0.220529
10	10	8	0.128049	10	10	8	0.170732	10	10	8	0.247968
10	10	10	0.138211	10	10	10	0.192073	10	10	10	0.25
20	6	4	0.122968	20	6	4	0.170732	20	6	4	0.251016
20	6	6	0.114837	20	6	6	0.186992	20	6	6	0.293699
20	6	8	0.094512	20	6	8	0.179878	20	6	8	0.26626
20	6	10	0.121951	20	6	10	0.179878	20	6	10	0.298781

20	7	4	0.105691	20	7	4	0.190041	20	7	4	0.269309
20	7	6	0.103659	20	7	6	0.203252	20	7	6	0.307927
20	7	8	0.118902	20	7	8	0.199187	20	7	8	0.307927
20	7	10	0.120935	20	7	10	0.194106	20	7	10	0.298781
20	8	4	0.122968	20	8	4	0.199187	20	8	4	0.27439
20	8	6	0.106707	20	8	6	0.200203	20	8	6	0.281504
20	8	8	0.10874	20	8	8	0.185976	20	8	8	0.273374
20	8	10	0.120935	20	8	10	0.211382	20	8	10	0.32622
20	9	4	0.118902	20	9	4	0.226626	20	9	4	0.307927
20	9	6	0.10874	20	9	6	0.221545	20	9	6	0.302846
20	9	8	0.091463	20	9	8	0.191057	20	9	8	0.268293
20	9	10	0.102642	20	9	10	0.189024	20	9	10	0.283537
20	10	4	0.093496	20	10	4	0.193089	20	10	4	0.268293
20	10	6	0.114837	20	10	6	0.222561	20	10	6	0.291667
20	10	8	0.125	20	10	8	0.24187	20	10	8	0.298781
20	10	10	0.119919	20	10	10	0.221545	20	10	10	0.317073
30	6	4	0.10874	30	6	4	0.20935	30	6	4	0.333333
30	6	6	0.095529	30	6	6	0.215447	30	6	6	0.321138
30	6	8	0.104675	30	6	8	0.245935	30	6	8	0.361789
30	6	10	0.10061	30	6	10	0.218496	30	6	10	0.377033
30	7	4	0.112805	30	7	4	0.229675	30	7	4	0.327236
30	7	6	0.110772	30	7	6	0.216463	30	7	6	0.367886
30	7	8	0.106707	30	7	8	0.259146	30	7	8	0.380081
30	7	10	0.11687	30	7	10	0.24187	30	7	10	0.384146
30	8	4	0.105691	30	8	4	0.25813	30	8	4	0.368902
30	8	6	0.123984	30	8	6	0.260163	30	8	6	0.387195
30	8	8	0.10874	30	8	8	0.257114	30	8	8	0.363821
30	8	10	0.115854	30	8	10	0.239837	30	8	10	0.364837
30	9	4	0.10874	30	9	4	0.260163	30	9	4	0.352642
30	9	6	0.102642	30	9	6	0.288618	30	9	6	0.38313

30	9	8	0.11687	30	9	8	0.288618	30	9	8	0.395325
30	9	10	0.105691	30	9	10	0.244919	30	9	10	0.400407
30	10	4	0.11687	30	10	4	0.284553	30	10	4	0.368902
30	10	6	0.115854	30	10	6	0.28252	30	10	6	0.382114
30	10	8	0.115854	30	10	8	0.243902	30	10	8	0.378049
30	10	10	0.094512	30	10	10	0.281504	30	10	10	0.368902
40	6	4	0.109756	40	6	4	0.267276	40	6	4	0.371951
40	6	6	0.11687	40	6	6	0.260163	40	6	6	0.401423
40	6	8	0.130081	40	6	8	0.277439	40	6	8	0.446138
40	6	10	0.120935	40	6	10	0.298781	40	6	10	0.452236
40	7	4	0.132114	40	7	4	0.304878	40	7	4	0.417683
40	7	6	0.129065	40	7	6	0.294715	40	7	6	0.450203
40	7	8	0.117886	40	7	8	0.288618	40	7	8	0.443089
40	7	10	0.117886	40	7	10	0.29065	40	7	10	0.463415
40	8	4	0.118902	40	8	4	0.325203	40	8	4	0.423781
40	8	6	0.125	40	8	6	0.307927	40	8	6	0.453252
40	8	8	0.121951	40	8	8	0.279472	40	8	8	0.460366
40	8	10	0.115854	40	8	10	0.330285	40	8	10	0.481707
40	9	4	0.113821	40	9	4	0.332317	40	9	4	0.441057
40	9	6	0.111789	40	9	6	0.304878	40	9	6	0.449187
40	9	8	0.134146	40	9	8	0.313008	40	9	8	0.460366
40	9	10	0.119919	40	9	10	0.328252	40	9	10	0.477642
40	10	4	0.111789	40	10	4	0.327236	40	10	4	0.460366
40	10	6	0.101626	40	10	6	0.306911	40	10	6	0.441057
40	10	8	0.109756	40	10	8	0.324187	40	10	8	0.504065
40	10	10	0.10874	40	10	10	0.328252	40	10	10	0.482724
50	6	4	0.142276	50	6	4	0.323171	50	6	4	0.478659
50	6	6	0.126016	50	6	6	0.319106	50	6	6	0.47561
50	6	8	0.13313	50	6	8	0.317073	50	6	8	0.519309
50	6	10	0.140244	50	6	10	0.321138	50	6	10	0.520325

50	7	4	0.13313	50	7	4	0.329268	50	7	4	0.490854
50	7	6	0.136179	50	7	6	0.347561	50	7	6	0.531504
50	7	8	0.127033	50	7	8	0.336382	50	7	8	0.544716
50	7	10	0.132114	50	7	10	0.33435	50	7	10	0.563008
50	8	4	0.147358	50	8	4	0.388211	50	8	4	0.52439
50	8	6	0.119919	50	8	6	0.361789	50	8	6	0.529472
50	8	8	0.13313	50	8	8	0.372968	50	8	8	0.544716
50	8	10	0.139228	50	8	10	0.376016	50	8	10	0.571138
50	9	4	0.131098	50	9	4	0.371951	50	9	4	0.501016
50	9	6	0.122968	50	9	6	0.398374	50	9	6	0.563008
50	9	8	0.136179	50	9	8	0.412602	50	9	8	0.561992
50	9	10	0.10874	50	9	10	0.375	50	9	10	0.544716
50	10	4	0.125	50	10	4	0.394309	50	10	4	0.521341
50	10	6	0.13313	50	10	6	0.390244	50	10	6	0.52439
50	10	8	0.130081	50	10	8	0.373984	50	10	8	0.537602
50	10	10	0.126016	50	10	10	0.364837	50	10	10	0.528455
60	6	4	0.144309	60	6	4	0.347561	60	6	4	0.520325
60	6	6	0.152439	60	6	6	0.389228	60	6	6	0.593496
60	6	8	0.143293	60	6	8	0.373984	60	6	8	0.60874
60	6	10	0.136179	60	6	10	0.361789	60	6	10	0.60061
60	7	4	0.168699	60	7	4	0.421748	60	7	4	0.583333
60	7	6	0.151423	60	7	6	0.412602	60	7	6	0.611789
60	7	8	0.132114	60	7	8	0.39939	60	7	8	0.595529
60	7	10	0.146342	60	7	10	0.394309	60	7	10	0.593496
60	8	4	0.136179	60	8	4	0.406504	60	8	4	0.561992
60	8	6	0.151423	60	8	6	0.420732	60	8	6	0.605691
60	8	8	0.163618	60	8	8	0.427846	60	8	8	0.603659
60	8	10	0.160569	60	8	10	0.438008	60	8	10	0.635163
60	9	4	0.144309	60	9	4	0.427846	60	9	4	0.572155
60	9	6	0.137195	60	9	6	0.416667	60	9	6	0.595529

60	9	8	0.151423	60	9	8	0.429878	60	9	8	0.637195
60	9	10	0.14939	60	9	10	0.433943	60	9	10	0.629065
60	10	4	0.15752	60	10	4	0.470529	60	10	4	0.605691
60	10	6	0.150407	60	10	6	0.460366	60	10	6	0.607724
60	10	8	0.156504	60	10	8	0.461382	60	10	8	0.645325
60	10	10	0.137195	60	10	10	0.480691	60	10	10	0.66565

A 5.9 Simulated power table for group difference example: Missing outcome data; effect size 0.8; varying model variances

20% missing data in y Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$				**70% missing data in y** Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$				**Missing data in y according to time trend** $\boldsymbol{t} + \boldsymbol{t}^{\boldsymbol{2}} + s$ Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
70	6	4	0.710366	70	6	4	0.650407	70	6	4	0.699187
70	6	6	0.691057	70	6	6	0.66565	70	6	6	0.692073
70	6	8	0.703252	70	6	8	0.682927	70	6	8	0.707317
70	6	10	0.721545	70	6	10	0.712398	70	6	10	0.728659
70	7	4	0.694106	70	7	4	0.643293	70	7	4	0.690041
70	7	6	0.751016	70	7	6	0.707317	70	7	6	0.736789
70	7	8	0.689024	70	7	8	0.672764	70	7	8	0.685976
70	7	10	0.735772	70	7	10	0.732724	70	7	10	0.744919
70	8	4	0.730691	70	8	4	0.666667	70	8	4	0.71748
70	8	6	0.756098	70	8	6	0.723577	70	8	6	0.738821
70	8	8	0.730691	70	8	8	0.720529	70	8	8	0.731707
70	8	10	0.716463	70	8	10	0.711382	70	8	10	0.714431
70	9	4	0.745935	70	9	4	0.702236	70	9	4	0.734756
70	9	6	0.710366	70	9	6	0.702236	70	9	6	0.716463
70	9	8	0.760163	70	9	8	0.735772	70	9	8	0.742886
70	9	10	0.743902	70	9	10	0.720529	70	9	10	0.748984
70	10	4	0.716463	70	10	4	0.662602	70	10	4	0.689024
70	10	6	0.710366	70	10	6	0.685976	70	10	6	0.702236
70	10	8	0.73374	70	10	8	0.724594	70	10	8	0.730691
70	10	10	0.707317	70	10	10	0.704268	70	10	10	0.70935
80	6	4	0.745935	80	6	4	0.690041	80	6	4	0.739837
80	6	6	0.756098	80	6	6	0.71748	80	6	6	0.740854

80	6	8	0.770325	80	6	8	0.742886	80	6	8	0.769309
80	6	10	0.744919	80	6	10	0.731707	80	6	10	0.746951
80	7	4	0.755081	80	7	4	0.708333	80	7	4	0.751016
80	7	6	0.763211	80	7	6	0.721545	80	7	6	0.76626
80	7	8	0.760163	80	7	8	0.720529	80	7	8	0.740854
80	7	10	0.767276	80	7	10	0.754065	80	7	10	0.770325
80	8	4	0.746951	80	8	4	0.705285	80	8	4	0.745935
80	8	6	0.772358	80	8	6	0.728659	80	8	6	0.762195
80	8	8	0.775407	80	8	8	0.759146	80	8	8	0.768293
80	8	10	0.778455	80	8	10	0.770325	80	8	10	0.777439
80	9	4	0.783537	80	9	4	0.740854	80	9	4	0.777439
80	9	6	0.776423	80	9	6	0.742886	80	9	6	0.773374
80	9	8	0.763211	80	9	8	0.751016	80	9	8	0.764228
80	9	10	0.795732	80	9	10	0.791667	80	9	10	0.79065
80	10	4	0.775407	80	10	4	0.745935	80	10	4	0.765244
80	10	6	0.772358	80	10	6	0.768293	80	10	6	0.765244
80	10	8	0.791667	80	10	8	0.770325	80	10	8	0.789634
80	10	10	0.805894	80	10	10	0.793699	80	10	10	0.805894
90	6	4	0.786585	90	6	4	0.736789	90	6	4	0.770325
90	6	6	0.779472	90	6	6	0.745935	90	6	6	0.786585
90	6	8	0.800813	90	6	8	0.784553	90	6	8	0.805894
90	6	10	0.800813	90	6	10	0.778455	90	6	10	0.788618
90	7	4	0.804878	90	7	4	0.759146	90	7	4	0.799797
90	7	6	0.796748	90	7	6	0.773374	90	7	6	0.803862
90	7	8	0.815041	90	7	8	0.786585	90	7	8	0.802846
90	7	10	0.811992	90	7	10	0.794716	90	7	10	0.802846
90	8	4	0.807927	90	8	4	0.764228	90	8	4	0.794716
90	8	6	0.843496	90	8	6	0.82622	90	8	6	0.844512
90	8	8	0.817073	90	8	8	0.793699	90	8	8	0.818089
90	8	10	0.822155	90	8	10	0.806911	90	8	10	0.813008

90	9	4	0.831301	90	9	4	0.786585	90	9	4	0.816057
90	9	6	0.803862	90	9	6	0.781504	90	9	6	0.803862
90	9	8	0.819106	90	9	8	0.795732	90	9	8	0.811992
90	9	10	0.838415	90	9	10	0.82622	90	9	10	0.840447
90	10	4	0.810976	90	10	4	0.797764	90	10	4	0.806911
90	10	6	0.851626	90	10	6	0.827236	90	10	6	0.841463
90	10	8	0.829268	90	10	8	0.813008	90	10	8	0.821138
90	10	10	0.831301	90	10	10	0.815041	90	10	10	0.828252
100	6	4	0.796748	100	6	4	0.765244	100	6	4	0.795732
100	6	6	0.849594	100	6	6	0.815041	100	6	6	0.844512
100	6	8	0.853659	100	6	8	0.828252	100	6	8	0.849594
100	6	10	0.849594	100	6	10	0.837398	100	6	10	0.848577
100	7	4	0.816057	100	7	4	0.76626	100	7	4	0.799797
100	7	6	0.839431	100	7	6	0.818089	100	7	6	0.833333
100	7	8	0.84248	100	7	8	0.828252	100	7	8	0.840447
100	7	10	0.864837	100	7	10	0.848577	100	7	10	0.856707
100	8	4	0.837398	100	8	4	0.811992	100	8	4	0.833333
100	8	6	0.869919	100	8	6	0.844512	100	8	6	0.871951
100	8	8	0.847561	100	8	8	0.838415	100	8	8	0.846545
100	8	10	0.853659	100	8	10	0.832317	100	8	10	0.85061
100	9	4	0.849594	100	9	4	0.821138	100	9	4	0.85061
100	9	6	0.867886	100	9	6	0.853659	100	9	6	0.86687
100	9	8	0.84248	100	9	8	0.827236	100	9	8	0.838415
100	9	10	0.879065	100	9	10	0.868902	100	9	10	0.878049
100	10	4	0.861789	100	10	4	0.829268	100	10	4	0.847561
100	10	6	0.853659	100	10	6	0.835366	100	10	6	0.843496
100	10	8	0.869919	100	10	8	0.860772	100	10	8	0.881098
100	10	10	0.881098	100	10	10	0.867886	100	10	10	0.879065

A 5.10 Simulated power table for group difference example: Missing outcome and covariate data; effect size 0.8; varying model variances

10% missing data in $\boldsymbol{x}$; 20% missing data in y Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$				**30% missing data in $\boldsymbol{x}$**; 20% missing data in y Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$				**50% missing data in $\boldsymbol{x}$**; 20% missing data in y Effect size $d = 0.8$ $\sigma_u^2 = \sigma_v^2 = \sigma_e^2 = 0.26$			
n3	n2	n1	x_reject	n3	n2	n1	x_reject	n3	n2	n1	x_reject
70	6	4	0.664634	70	6	4	0.561992	70	6	4	0.439024
70	6	6	0.630081	70	6	6	0.530488	70	6	6	0.428862
70	6	8	0.64126	70	6	8	0.547764	70	6	8	0.418699
70	6	10	0.697155	70	6	10	0.571138	70	6	10	0.444106
70	7	4	0.664634	70	7	4	0.529472	70	7	4	0.413618
70	7	6	0.652439	70	7	6	0.543699	70	7	6	0.419715
70	7	8	0.666667	70	7	8	0.544716	70	7	8	0.429878
70	7	10	0.680894	70	7	10	0.571138	70	7	10	0.439024
70	8	4	0.683943	70	8	4	0.58435	70	8	4	0.458333
70	8	6	0.653455	70	8	6	0.551829	70	8	6	0.432927
70	8	8	0.676829	70	8	8	0.560976	70	8	8	0.448171
70	8	10	0.694106	70	8	10	0.580285	70	8	10	0.452236
70	9	4	0.682927	70	9	4	0.568089	70	9	4	0.452236
70	9	6	0.692073	70	9	6	0.577236	70	9	6	0.446138
70	9	8	0.652439	70	9	8	0.556911	70	9	8	0.448171
70	9	10	0.697155	70	9	10	0.582317	70	9	10	0.465447
70	10	4	0.660569	70	10	4	0.554878	70	10	4	0.428862
70	10	6	0.730691	70	10	6	0.587398	70	10	6	0.464431
70	10	8	0.688008	70	10	8	0.589431	70	10	8	0.464431
70	10	10	0.702236	70	10	10	0.596545	70	10	10	0.477642
80	6	4	0.716463	80	6	4	0.612805	80	6	4	0.474594

80	6	6	0.700203	80	6	6	0.604675	80	6	6	0.498984
80	6	8	0.721545	80	6	8	0.593496	80	6	8	0.466463
80	6	10	0.683943	80	6	10	0.57622	80	6	10	0.454268
80	7	4	0.716463	80	7	4	0.621951	80	7	4	0.492886
80	7	6	0.743902	80	7	6	0.622968	80	7	6	0.502033
80	7	8	0.72561	80	7	8	0.618902	80	7	8	0.480691
80	7	10	0.714431	80	7	10	0.61687	80	7	10	0.487805
80	8	4	0.703252	80	8	4	0.619919	80	8	4	0.501016
80	8	6	0.724594	80	8	6	0.621951	80	8	6	0.50813
80	8	8	0.732724	80	8	8	0.618902	80	8	8	0.485772
80	8	10	0.705285	80	8	10	0.607724	80	8	10	0.494919
80	9	4	0.712398	80	9	4	0.617886	80	9	4	0.461382
80	9	6	0.723577	80	9	6	0.628049	80	9	6	0.503049
80	9	8	0.746951	80	9	8	0.642276	80	9	8	0.497968
80	9	10	0.720529	80	9	10	0.614837	80	9	10	0.481707
80	10	4	0.753049	80	10	4	0.644309	80	10	4	0.498984
80	10	6	0.752033	80	10	6	0.637195	80	10	6	0.501016
80	10	8	0.748984	80	10	8	0.647358	80	10	8	0.520325
80	10	10	0.754065	80	10	10	0.661585	80	10	10	0.518293
90	6	4	0.773374	90	6	4	0.660569	90	6	4	0.506098
90	6	6	0.794716	90	6	6	0.676829	90	6	6	0.549797
90	6	8	0.76626	90	6	8	0.654472	90	6	8	0.519309
90	6	10	0.753049	90	6	10	0.646341	90	6	10	0.498984
90	7	4	0.767276	90	7	4	0.663618	90	7	4	0.533537
90	7	6	0.775407	90	7	6	0.664634	90	7	6	0.531504
90	7	8	0.775407	90	7	8	0.661585	90	7	8	0.517276
90	7	10	0.769309	90	7	10	0.659553	90	7	10	0.533537
90	8	4	0.76626	90	8	4	0.668699	90	8	4	0.534553
90	8	6	0.781504	90	8	6	0.666667	90	8	6	0.517276
90	8	8	0.818089	90	8	8	0.722561	90	8	8	0.588415

90	8	10	0.796748	90	8	10	0.691057	90	8	10	0.545732
90	9	4	0.784553	90	9	4	0.671748	90	9	4	0.560976
90	9	6	0.78252	90	9	6	0.678862	90	9	6	0.541667
90	9	8	0.775407	90	9	8	0.684959	90	9	8	0.579268
90	9	10	0.784553	90	9	10	0.677846	90	9	10	0.546748
90	10	4	0.773374	90	10	4	0.677846	90	10	4	0.544716
90	10	6	0.795732	90	10	6	0.677846	90	10	6	0.551829
90	10	8	0.789634	90	10	8	0.686992	90	10	8	0.565041
90	10	10	0.787602	90	10	10	0.697155	90	10	10	0.565041
100	6	4	0.78252	100	6	4	0.676829	100	6	4	0.531504
100	6	6	0.815041	100	6	6	0.715447	100	6	6	0.574187
100	6	8	0.798781	100	6	8	0.690041	100	6	8	0.534553
100	6	10	0.805894	100	6	10	0.720529	100	6	10	0.587398
100	7	4	0.804878	100	7	4	0.730691	100	7	4	0.547764
100	7	6	0.804878	100	7	6	0.71748	100	7	6	0.566057
100	7	8	0.828252	100	7	8	0.719512	100	7	8	0.589431
100	7	10	0.817073	100	7	10	0.711382	100	7	10	0.577236
100	8	4	0.797764	100	8	4	0.714431	100	8	4	0.572155
100	8	6	0.821138	100	8	6	0.727642	100	8	6	0.579268
100	8	8	0.803862	100	8	8	0.718496	100	8	8	0.569106
100	8	10	0.833333	100	8	10	0.724594	100	8	10	0.593496
100	9	4	0.803862	100	9	4	0.706301	100	9	4	0.561992
100	9	6	0.827236	100	9	6	0.715447	100	9	6	0.574187
100	9	8	0.833333	100	9	8	0.713415	100	9	8	0.595529
100	9	10	0.816057	100	9	10	0.724594	100	9	10	0.597561
100	10	4	0.827236	100	10	4	0.719512	100	10	4	0.585366
100	10	6	0.852642	100	10	6	0.736789	100	10	6	0.605691
100	10	8	0.833333	100	10	8	0.740854	100	10	8	0.603659
100	10	10	0.819106	100	10	10	0.720529	100	10	10	0.581301