Rethinking the nature of mental disorder: a latent structure approach to data from three national psychiatric morbidity surveys

Rachel Louise McCrea

Mental Health Sciences Unit
Faculty of Brain Sciences
University College London

A thesis submitted for the degree of Doctor of Philosophy

2013

Declaration

I, Rachel Louise McCrea confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

High levels of comorbidity between the anxiety and depressive disorders have raised questions about whether the diagnostic boundaries between these disorders need to be redrawn, or even whether both should be considered as different facets of a single disease process. Accordingly, latent class analysis has been used in several attempts to find data-driven groupings of individuals based on the symptoms of anxiety and depression. However, the assumption of conditional independence in this approach risks the extraction of spurious 'severity classes', making findings difficult to interpret. Factor mixture analysis relaxes that assumption by incorporating a common factor within each class, thereby overcoming the problem.

This project investigated whether factor mixture analysis can suggest a data-driven classification of individuals based on the symptoms of common mental disorders. The analysis was based on pooled symptom-level data from three national psychiatric morbidity surveys of adults living in Great Britain carried out in 1993, 2000 and 2007. A comparison of the fit from the various latent variable models indicated that factor mixture models provided the best fit to the data, both in terms of model parsimony and goodness-of-fit. However, subsequent investigations suggested that the classes did not represent true groups in the population, but were rather accommodating violations of key assumptions in the standard factor model. Therefore, the results provide little guidance for revising the psychiatric classification.

This is the first study to carry out an in-depth investigation into the interpretation of the extracted classes after applying factor mixture models to investigate the latent structure of mental disorders; its findings highlight the difficulties of interpreting the results of these models. Consequently, the thesis questions whether factor mixture mixture models are actually useful for exploring the true nature of psychiatric disorders, and whether the present heavy use of such models is justified. An investigation of previously published examples suggests that their results may be prone to misinterpretation. The thesis concludes with a set of recommendations for the reporting of these models that may help to minimise the risks of such misinterpretation.

Acknowledgements

Thank you to all who have supported me during my three years of PhD study. Particular thanks must go to my two supervisors: Professor Michael King and Dr Rumana Omar. I thank them especially for giving me the freedom and encouragement to try out unfamiliar methodologies and to follow my own statistical interests, while always pushing me to apply the same rigour that they would expect in more familiar approaches.

Thanks to all the team from the 2010 summer school in psychometrics at the University of Cambridge Psychometrics Centre — this course proved extremely useful, and really did 'jump start' my knowledge of psychometric techniques, giving me the confidence to strike out into the application of even more complex models on my own. Thanks also to the team from the 2011 spring school of the European Educational Research Association in Dortmund, in particular to Katherine Masyn, who opened my eyes to the ambiguities of latent variable mixture models.

Finally, I thank my husband Philip for all of his encouragement and moral support over the past three years. Without his constant calm support, the last three years would have been a much more stressful and difficult period.

This project was supported by a studentship from the Medical Research Council.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Acronyms and abbreviations have been kept to a minimum in the main body of the text, although they are more heavily used in tables, figures and captions. Acronyms and abbreviations are defined where the relevant terms are first introduced and at the bottom of tables, but for ease of reference they are also listed here.

AIC: Akaike information criterion

ADHD: attention-deficit/hyperactivity disorder

BIC: Bayesian information criterion

BLRT: bootstrap likelihood ratio test

c: class

CAPI: computer-assisted personal interviewing

CIDI: World Health Organization Composite International Diagnostic Interview

CIS-R: revised Clinical Interview Schedule

CFA: confirmatory factor analysis

CFI: comparative fit index

DIF: differential item functioning

DSM-IV: Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition)

EAP: expected a posteriori

EFA: exploratory factor analysis

E/CFA: exploratory factor analysis in the confirmatory factor analysis framework

f: factor

FA: factor analysis

FMA: factor mixture analysis

FMM: factor mixture model

GAD: generalised anxiety disorder

ICD-10: International Classification of Diseases 10th Revision

IRT: item response theory

LCA: latent class analysis

LCFA: latent class factor analysis (also known as non-parametric factor analysis)

LL: log-likelihood

LMR aLRT: Lo-Mendell-Rubin adjusted likelihood ratio test

OCD: obsessive compulsive disorder

MADD: mixed anxiety and depressive disorder

MCMC: Markov chain Monte Carlo

MMLE: marginal maximum likelihood estimation

MLR: maximum likelihood estimator with robust standard errors (in Mplus)

MPA: modified parallel analysis

NP-FA: non-parametric factor analysis (also known as latent class factor analysis)

PAPI: paper-and-pencil interviewing

PSU: primary sampling unit

RMSEA: root mean square error of approximation

SE: standard error

SP-FA: semi-parametric factor analysis

SRMR: standardised root mean square residual

TLI: Tucker Lewis index

WLSMV: weighted least squares mean and variance adjusted estimator (in Mplus)

par: number of parameters

Chapter 1

Introduction

1.1 Background

The current classificatory system for psychiatric disorders has grown out of many years of clinical observation and expert opinion, and has resulted in a scheme in which anxiety and depression are placed into separate major classes in both DSM-IV (American Psychiatric Association, 1994) and ICD-10 (World Health Organization, 1993). However, high levels of comorbidity between the anxiety and depressive disorders have led many (e.g., Maj, 2005; Tyrer, 2001; Watson, 2005) to propose that the classificatory system be modified in some way to account for the frequent co-occurrence of anxiety and depression symptoms. As an example of these high levels of comorbidity, among respondents to the US National Comorbidity Survey with a lifetime diagnosis of depression the prevalence of a lifetime anxiety disorder was 58% (Kessler *et al.*, 1996); the 12-month prevalence of anxiety disorders in the same group was 51%. Other studies have reported similar findings: a meta-analysis by Clark (1989) found that depressed patients showed an overall prevalence of 57% for the presence of any anxiety disorder.

This comorbidity has been described as “one of the ‘hot’ topics in psychopathology research” (Mineka *et al.*, 1998, p. 378), and has received particular attention in the lead up to the major revision of the DSM (Goldberg, 2008; Goldberg *et al.*, 2010). While much of this focus has been on the comorbidity between depression and generalised anxiety disorder, studies indicate that there is also substantial comorbidity between depression and the other anxiety disorders in both clinical and community samples that needs to be taken into account (Brown *et al.*, 2001; Kessler *et al.*, 1996). It has been argued that these high levels of comorbidity may be an artefact of a diagnostic system that makes an artificial division between the symptoms of anxiety and depression that does not exist in nature (Maj, 2005). To quote a colourful phrase of Levine *et al.* (2001, p.96), perhaps anxiety and depression are in fact "different parts of the same elephant".

High levels of comorbidity between the anxiety and depressive disorders are not the only indication that the present rigid division of symptoms of anxiety and depression into separate families of disorders may be inappropriate or unhelpful. The same drugs are often effective for treating both conditions: many antidepressants are known to be effective in those suffering from anxiety disorders (Zohar & Westenberg, 2000), while some anxiolytic drugs have also been shown to reduce the symptoms of depression (Levine *et al.*, 2001).

Furthermore, it seems that the same risk factors may predispose an individual to developing anxiety and depressive disorders. For example, King *et al.* (2011) found that the variables predicting which individuals in a cohort of primary care patients went on to develop generalised anxiety and panic syndromes were remarkably similar to those from an earlier investigation focussing on the outcome of major depression. In addition, twin studies suggest that depression and anxiety disorders share the same genetic vulnerability factors (Kendler *et al.*, 1992; Roy *et al.*, 1995; Middeldorp *et al.*, 2005).

Some may question why it matters whether the classificatory system treats anxiety and depression as unrelated syndromes, so long as individuals with symptoms of both can be diagnosed with two or more disorders. However, a correct understanding of the nature of these disorders may be vital for developing effective treatments. Shorter & Tyrer (2003) argue that the failure to adopt "natural disease categories" which acknowledge the close association between anxiety and depression may be responsible for the faltering of new drug discovery. Furthermore, confusion about the true relationship between anxiety and depression may also help to explain why treatments that target the symptoms of depression only are not always effective, as well as the high rate of relapse. A better understanding of the link between depression and anxiety may also suggest new therapeutic targets for the treatment of depression, both pharmacological and psychological.

A second area where an accurate understanding of the nature of these disorders may be important is in the study of their underlying biological mechanisms. In order to identify the biological pathways and genes involved in these disorders, it is vital to be able to identify accurately which individuals share a particular disorder and which do not. Otherwise, real effects may be diluted or missed completely by the presence of affected individuals in the unaffected control group (and vice versa).

The need to correctly delineate these disorders has led some researchers to consider whether multivariate statistical methods that aim to group objects according to common characteristics can be used to form an 'objectively' derived classification of mental disorders (e.g., Leoutsakos *et al.*, 2010) — a classification which is based on observed patterns of symptom co-occurrence, rather than one that is based on "the orthodox conceptualisation of these psychiatric disorders . . . as discrete nosological entities" (Sullivan & Kendler, 1998, p. 318). Such an objectively revised classification might reflect a simplification of the existing system with more emphasis on what the disorders have in common than on the specific differences in their presentations, as Sullivan and Kendler state (1998, p. 318):

> It is nearly axiomatic that nosology is the art of carving nature at the joints — it is less frequently appreciated that all joints are not of equal importance. For these common disorders, nature may have been, in effect, divided at a minor carpal joint instead of at the shoulder.[1]

However, an objective analysis may also reveal evidence for clearly defined and qualitatively distinct sets of symptoms that are not presently considered as 'diagnostic entities' in the classificatory system, but which may make a more useful basis for future research and treatment decisions than the present categories.

In addition to aiding our understanding of symptom co-occurrence in those meeting the existing criteria for diagnosis of depressive or anxiety disorders, a detailed investigation of symptom patterns may also provide a clearer picture of sub-threshold conditions. Sub-threshold conditions have been shown to account for a substantial proportion of the disability associated with mental ill-health in community surveys (Cuijpers *et al.*, 2004; Das-Munshi *et al.*, 2008). Furthermore, it has been shown that patients presenting with sub-threshold conditions in primary care more often present with a mixed pattern of non-specific symptoms of depression and anxiety than with mild forms of either 'pure' depression or 'pure' anxiety (Zinbarg *et al.*, 1994).

While ICD-10 allows for the sub-threshold diagnosis of mixed anxiety and depression (this diagnosis was only included in an appendix of DSM-IV 'for Further Study', along with sub-threshold or 'minor' depression), the lack of any defined criteria for this diagnosis virtually preclude making reliable prevalence estimates (or cross-national comparisons) for the condition across different measurement instruments. Indeed, the definition of mixed anxiety and depressive disorder used by the UK psychiatric morbidity surveys is based on a simple cut-off on a summed score of neurotic symptoms (Office of Population Censuses and Surveys, 1995), and as such

[1] Permission to reproduce this quotation has been granted by the Royal College of Psychiatrists.

requires neither the presence of depressed mood nor any specific symptoms of anxiety. Diagnostic criteria for this condition were proposed for inclusion in the main body of the future DSM-5 (American Psychiatric Association, 2010b). However, early reports from the field trials of the new diagnostic criteria suggest that their reliability is extremely poor (Aldhous, 2012), and at the time of writing this condition had again been scheduled for inclusion in an appendix of DSM-5 'for further study' (American Psychiatric Association, 2012). It seems clear, therefore, that any reconsideration of diagnostic groupings would do well to incorporate a careful investigation of symptoms at the sub-threshold level.

1.2 Project aims

The broad aim of this project is to see what latent variable modelling can tell us about the structure of common mental disorders; it will use data from the three surveys of psychiatric morbidity among adults living in private households in Great Britain in 1993, 2000 and 2007. The project will not limit itself to either dimensional or categorical approaches to the study of latent structure, but will combine the results of both types of method applied to the symptoms of anxiety and depression, as well as new hybrid approaches.

As part of this general aim, the project will have two specific objectives:

1. Recognising that a categorical system of classification is clinically useful, the project will search for a classification of individuals that is based on the co-occurrence of symptoms, rather than preconceived ideas about which syndromes constitute particular disorders.

2. The project will also look for evidence of whether depression frequently occurs as a 'pure' condition without symptoms of anxiety, or whether it might be more appropriate to consider depressed mood as one part of a broader condition.

The project will also explore the appropriateness of the model assumptions for the various types of latent variable model, checking how well such latent variable models are suited to investigating these research objectives.

1.3 Methodology

1.3.1 Latent class analysis and its limitations

Since the main aim of this project is to find a classificatory grouping of individuals that is based on patterns of symptom co-occurrence, the methodology for this project will focus on latent class analysis and its extensions. (The following chapter will give a brief overview of some of the other statistical approaches that have been used to answer this type of question, along with a consideration of their strengths and limitations.) Latent class analysis is a model-based technique that aims to identify homogenous groups of individuals based on a number of observed 'indicator' variables. The indicator variables may be binary, ordered categorical or continuous variables, although when continuous variables are used the analysis is sometimes described as 'latent profile analysis' (Vermunt & Magidson, 2002, p. 89).

However, in many standard latent class analyses, some of the classes that are extracted seem to reflect only severity differences, i.e. the classes differ only in the number of symptoms that are endorsed, rather than showing any qualitative differences in patterns of symptoms. The reason for the extraction of classes differing only by severity is that the latent class model centres on the assumption of conditional independence, i.e., the symptoms may be correlated in the whole population, but should become 'independent' after conditioning on class membership.

Severity differences within a class will violate this assumption of conditional independence. For example, in a group of individuals with mixed anxiety and depression some individuals may have moderate anxiety along with moderate depression, while others may have severe anxiety along with severe depression. Since the severity of depression and anxiety now co-vary in this group, the conditional independence assumption would be violated, and two or more latent classes would have to be extracted to account for the severity differences within the original group. Nevertheless, it might be considered that the different classes all represent the same condition, albeit with differences in severity. In this case the symptoms may be better described as varying along a continuous dimension, rather than forming discrete categories. When the results of a latent class analysis on the symptoms of mental disorders present some classes that appear to differ largely by severity (such as in Das-Munshi *et al.*, 2008), it is not clear how these results should be used to inform a revision of the classificatory system.

Where the underlying construct is truly continuous rather than categorical (or a combination of dimensions and categories), it has been shown that standard latent class analysis can result in the over-extraction of classes to account for the dimensionality of the data (Lubke & Neale, 2006). Ideally, a method should be used that can identify classes that differ qualitatively in the pattern of symptoms that are endorsed, but that can also allow symptoms to vary along a dimension of severity within a single latent class.

1.3.2 Factor mixture analysis

A recent development in latent variable modelling is the use of hybrid models, such as factor mixture analysis, that allow categorical and continuous latent variables to be combined. (For an overview of 'traditional' latent variable models and 'new' hybrids, see Muthen, 2007). Given the difficulty in psychiatry of deciding between categorical and dimensional approaches, models that allow for both dimensional and categorical approaches simultaneously could be of great interest. In addition, they mirror the proposals for DSM-V to incorporate dimensional assessments of severity within categorical diagnoses (American Psychiatric Association, 2010a,c). The factor mixture approach has already been used to examine the latent structure of symptoms of a number of mental disorders, including attention-deficit/hyperactivity disorder (ADHD; Lubke *et al.*, 2007), alcohol dependence (Muthen, 2006) and tobacco dependence (Muthen & Asparouhov, 2006).

An advantage of the factor mixture approach is that both factor analysis and latent class models can be viewed as sub-types of the more general factor mixture model: factor analysis can be seen as a factor mixture model with one or more continuous latent variables but only a single latent class, while latent class analysis can be seen as a type of factor mixture analysis with two or more latent classes but a factor covariance matrix of $\mathbf{0}$ (Clark *et al.*, 2009). This means that candidate models of different types can be usefully compared using goodness-of-fit statistics, such as the Bayesian Information Criterion, or a bootstrapped likelihood ratio test. Consequently, a choice between dimensional, categorical or hybrid approaches need not be made purely on the grounds of interpretability or preconceptions about the nature of the concept in question.

As yet, it appears that no published studies have used hybrid latent variable models to investigate the co-occurrence of symptoms of anxiety and depression. One study has used Rasch mixture analysis (a type of hybrid model that combines a very restrictive factor model with latent classes) to investigate the symptoms of depression alone (Hong & Min, 2007). However, this analysis was based on data from several classes of university students, so it is not clear whether it included the full range of symptom severity that might be seen in a large community survey. Another paper has also applied hybrid latent variable models to the symptoms of depression, this time using data from a large community sample. However, this was a methodology paper that looked at issues such as sample size and power in factor mixture analysis (Lubke & Neale, 2008). As an applied illustration, the authors selected a model with one factor and three classes to describe the symptoms of depression, but they did not present parameter estimates or describe the classes, so their chosen model cannot be interpreted. Therefore, this project aims to provide new insight as one of the first studies to use these hybrid latent variable models to investigate patterns of symptom co-occurrence in the common mental disorders.

1.4 Suitability of data and measurement instrument

As will be demonstrated in the literature review in the next chapter, there is presently a dearth of studies that have used high quality data to examine patterns of symptom co-occurrence in the common mental disorders among adults. This is possibly because of the structure of the measurement instruments that have been used for many of the major mental health surveys around the world; many surveys screen respondents for the likely presence of an anxiety or depressive disorder early on in questioning, and then only ask detailed questions about the presence of particular symptoms where it seems likely that an individual may meet the criteria for the associated diagnosis. As a result, information on the co-occurrence of individual symptoms for the anxiety and depressive disorders is patchy.

However, this is one of the areas where the data from the three UK psychiatric morbidity surveys have particular strengths. The three cross-sectional surveys in 1993, 2000 and 2007 each used the same interview to measure symptoms of common mental disorders, the revised Community Interview Schedule (CIS-R; Lewis *et al.*, 1992). The CIS-R interview is well suited to addressing the aims of the project: a key feature of the CIS-R interview is that it asks *all* respondents about each different type of symptom (for example, fatigue, sleep problems or worry), without first screening for the likely presence of the disorders to which those symptoms are thought to relate (in contrast with instruments such as the Composite International Diagnostic Interview; Kessler & Ustun, 2004). This means that it is possible to examine the co-occurrence of symptoms in all individuals, not just those who have met the disorder screening criteria.

Furthermore, the CIS-R interview focuses on symptoms that have been experienced in the previous week (although questions are asked about symptom duration to help with ICD-10 diagnoses, but they will not be used here). As a result, we can be confident that when a respondent has reported a number of different symptoms, these symptoms actually co-occurred as part of the same syndrome. (This is not the case for instruments that only report 'lifetime' symptoms or symptoms that were experienced in the previous year, where symptoms that are reported together may in fact have been experienced completely independently.) Another benefit of the focus on the past seven days is that respondents' self-reports will be less affected by forgetting or the differing amounts of effort that participants put into formulating their answers to survey questions. The analysis will also be less affected by the fact that some symptoms may be more memorable than others or that those who were diagnosed with a particular mental health condition may have been more likely to recall the symptoms related to that condition.

Despite its many strengths, the use of data from the CIS-R interview does still carry some limitations. As with all latent structure analyses based on a particular measurement instrument, the results will always be dependent (at least to a certain extent) on the number of items that were chosen to measure each symptom and the types of symptom that the instrument covers. In addition, some questions may have been poorly understood by respondents or be poor measures of the symptom they were written to assess. In common with many other measures of mental health, there are also some assumptions that are built into the structure of the questionnaire. This is particularly the case for the section on depressive ideas (such as feeling guilty or hopeless) where the assumption is made that only those who have experienced symptoms of depression in the previous week could have experienced these symptoms — other respondents are taken straight to the next section of the questionnaire. Nonetheless, the CIS-R interview provides data on the symptoms of common mental disorders that are less constrained by assumptions about which symptoms ought to occur together than data from most of the other major mental health surveys.

The three surveys in 1993, 2000 and 2007 each used similar community populations and similar random sampling schemes with the intention that results would be comparable from one survey to the next. Since the samples for the three surveys were drawn independently from a very large population, they can be combined easily to form a larger pooled data set (Korn & Graubard, 1999, p. 278). The result of combining these three data sets is a very large sample size (22,574 adults aged 16-64) that will provide one of the best opportunities yet for detecting symptom patterns that are relatively rare.

Consequently, the pooled UK psychiatric morbidity survey data sets appear to be very well suited to address the research aims of this project and to supplement the meagre research literature on patterns of co-occurrence of symptoms of depression and anxiety in the adult population. While a paper by Das-Munshi *et al.* (2008) has used a subset of these data to carry out a latent class analysis based on binary summaries of the 14 symptom scores, this project will extend that research in a number of ways: first, by the use of a larger data set with more power to investigate rare symptoms; secondly, by use of full symptom scores rather than binary summaries; and thirdly, by the incorporation of dimensional measures into the latent class approach through the use of hybrid latent variable models.

The following chapter will investigate what the existing research using latent class analysis has been able to tell us about patterns of symptom co-occurrence in the common mental disorders, and whether this research has been able to describe any useful groupings of individuals that may inform revisions to the psychiatric classification. It will also consider briefly what other multivariate statistical approaches have been used to investigate this area.

Chapter 2

Literature review

This chapter will consider what multivariate statistical analyses have told us so far about patterns of symptom co-occurrence in anxiety and depression, and whether these results can tell us anything about how the diagnostic classification of common mental disorders might be revised. However, before looking at individual papers, it will take a look at some of the statistical methods that have been used to investigate structure in the symptoms of anxiety and depression, and will consider which of these are the most appropriate for answering such a question

2.1 Multivariate statistical methods that could be used for investigating the structure of symptoms in depression and anxiety

2.1.1 Cluster analysis and latent class analysis

A number of different statistical methods have been used to examine the structure of common mental disorders. However, when the focus is on informing a new classification that will take account of the high levels of comorbidity between disorders, methods are needed that can separate individuals into groups with similar symptom and comorbidity profiles.

Two popular techniques that aim to identify homogeneous groups of individuals are non-hierarchical (e.g. k-means) cluster analysis and latent class analysis (LCA): both of these methods attempt to allocate individuals to a pre-defined number of homogeneous groups, the characteristics of which are not known in advance (Vermunt & Magidson, 2002, p. 89). These methods can both be used in an exploratory fashion, whereby the actual number of these classes (as well as their characteristics) is unknown — the analysis generally proceeds by finding a set of different solutions over a range of numbers of classes, and then trying to identify the solution with the optimum number of classes (based on interpretability or statistical criteria).

The key difference between these two methods is that latent class analysis is based on an underlying statistical model: it is assumed that the observed sample consists of a mixture of probability distributions, each corresponding to one of the specified number of latent classes (Vermunt & Magidson, 2002, p. 90). During model estimation, the unknown parameters of the distributions are estimated (by maximum likelihood or some other method) — these parameter estimates then yield a corresponding set of membership probabilities for each of the derived classes for all possible response patterns. As such, individuals in the sample can subsequently be allocated to the class for which they have the highest probability of membership (although for some individuals there may still be substantial uncertainty). Crucially, as long as the model holds, the estimated class membership probabilities also apply to unsampled members of the population who have a corresponding pattern of symptoms.

Conversely, in cluster analysis no population model is hypothesised — sampled individuals are simply allocated to one of the clusters by a numerical algorithm. Consequently, cluster analysis takes no account of the uncertainty or error that may be present in allocating individuals to clusters. There are a number of algorithms available for cluster analysis — these generally aim to minimise the differences between individuals within each cluster or to maximise the differences between the clusters. Unfortunately, since cluster analysis does not involve the estimation of parameters for the probability of endorsing each symptoms within each cluster, its results are sometimes poorly reported (with little more than a prose description of cluster characteristics) — this can make it very difficult to compare results between studies (Blashfield & Morey, 1979). Furthermore, cluster solutions can vary substantially according to the algorithm that is used (Byrne, 1978), and this lack of stability/replicability limits the usefulness of cluster analysis results for informing psychiatric classification.

2.1.2 Other methods

Most of the other methods that have been used to investigate the structure of common mental disorders are not so well suited to identifying groups of individuals with homogeneous symptom patterns. For example, 'grade of membership' analysis is closely related to latent class analysis and has been used to investigate the symptoms of anxiety and depression (Piccinelli *et al.*, 1999). However, whereas latent class analysis assumes that an individual can belong to only one class (even if there is uncertainty about which class this should be), in grade of membership analysis individuals are allowed to belong to more than one class at the same time. The goal of the analysis is to identify ‘pure types’ of symptoms, so there can be high levels of ‘type comorbidity’ (and since a high level of comorbidity is one of the undesirable features of the present classification system, this approach seems less than ideal as a method to identify a more appropriate set of categories).

Another technique is the taxometric approach, which has been used in several analyses of major depressive disorder (Prisciandaro & Roberts, 2005; Ruscio *et al.*, 2007; Ruscio & Ruscio, 2000). However, the main goal of taxometric studies is to decide whether the symptoms of a particular condition lie upon a continuous dimension with normal experience, or whether there are discrete groups of those with the condition and those without it — the method looks for evidence of a ‘discontinuity’ along a single dimension of symptoms. This type of analysis is not well suited to situations where there may be multiple groups described by several different types of symptom and will not be considered here.

The other major approach that has been used to investigate the structure of common mental disorders is the use of factor analysis — this method assumes the presence of underlying continuous dimensions of symptoms, and as such, the results from factor analysis studies suggest no clear way to divide individuals into groups for research or treatment purposes. In fact, standard factor analysis models assume that the population is homogeneous and that the same underlying factor structure applies to all individuals in the population. If the population is actually comprised of one or more distinct groups of individuals, this may result in the extraction of additional ‘factors’ to account for the differences between groups (Lubke & Neale, 2006), even though a categorical division may be more appropriate. Consequently, although factor analytic studies have provided useful evidence that anxious and depressive symptoms tend to be closely linked (e.g., Watson, 2005; Markon, 2010), they can provide little guidance about whether or not particular diagnostic boundaries are needed, or where best to place them.

When the research aim is to inform the classification of mental disorders, a method that can identify meaningful categories (if there are any) is needed — latent class analysis seems to be the most suitable of the widely used methods for this purpose. The following sections review what evidence has been provided so far by latent class analysis about patterns of symptom co-occurrence in the common mental disorders.

2.2 Research findings

2.2.1 Papers looking at depression only

Many of the papers that have used latent class analysis to investigate common mental disorders have only been interested in searching for/validating sub-types of depressive disorders. Table 2.1 lists a number of studies using latent class analysis that have limited themselves to the study of depression, considering only symptoms that are part of the DSM III or DSM-IV criteria for major depression. They can tell us little about the co-occurrence of symptoms of anxiety and depression, because none included symptoms of anxiety in their analysis. Nonetheless, they may provide some information about how frequently the 'non-specific' symptoms of depression (such as sleep problems, fatigue and concentrations difficulties) occur in the absence of depressed mood.

Table 2.1 illustrates that some latent class studies have been limited to clinical samples of patients — these studies are not well suited to examining natural patterns of symptom co-occurrence, since some combinations of symptoms may make individuals more likely to seek medical help than others. Moreover, many of the studies that used population samples limited their analysis to individuals who had experienced symptoms of depression — these studies are of limited use for examining anything beyond the depressive syndrome itself, since it is not possible to see how often the symptoms under investigation occur together in the absence of depressed mood.

Table 2.1: Papers that have tried to investigate patterns of symptom co-occurrence in depression with latent class analysis

Paper	Sample type
Mora *et al.* (2012)	Population: sample of older adults (nearly all over 65)
Carragher *et al.* (2009)	Population: those with lifetime symptoms of depression only
Hybels *et al.* (2009)	Clinical: patients with major depression
Prisciandaro & Roberts (2009)	Population: those with lifetime symptoms of depression only
Sullivan *et al.* (2002)	Population: those with at least 1 symptom of major depression in past year
Chen *et al.* (2000)	Population
Parker *et al.* (1999)	Clinical: inpatients and outpatients with depression
Sullivan *et al.* (1998)	Population: those with a lifetime episode of depression
Kendler *et al.* (1996)	Population
Parker *et al.* (1995)	Clinical: patients with major depression
Keller & Kempf (1993)	Clinical: inpatients with depression
Eaton *et al.* (1989a)	Population

Table 2.1 indicates that four latent class studies have investigated the symptoms of depression in population samples without filtering their analysis to exclude those without depression. The earliest of these used data from the Epidemiological Catchment Area Study (Eaton *et al.*, 1989a) — this was a high quality study because, as well as using a large representative sample, it focussed only on symptoms that occurred in the month prior to interview (in contrast to many of the more recent studies that have looked at lifetime or past year symptoms). As such, the results are unlikely to be biased by selective recall of symptoms or to misrepresent symptoms as occurring together when in fact they were experienced months or years apart. Eaton *et al.* selected a three class model in which one class appeared to represent major depression, while a second larger class reported some of the non-specific symptoms but with a low probability of dysphoria (illustrated in Figure 2.1). However, it should be noted that Eaton *et al.* constrained the symptom of dysphoria to be present in class III in order to be consistent with the DSM III structure of depression — without this constraint, the most severe class included many individuals who endorsed four or more of the other symptoms but who did not appear to be suffering from DSM III major depression.

The Kendler *et al.* (1996) study also used a large unfiltered population sample and tried to focus on symptoms that co-occurred as part of a single syndrome. However, their latent class analysis was marred by a serious methodological flaw: several mutually exclusive pairs of symptoms (such as decreased and increased appetite) were included in the analysis together. Therefore, their latent class analysis is likely to have been dominated by the need to separate individuals reporting mutually exclusive symptoms into different classes. (This separation will occur because the conditional independence assumption of latent class analysis does not allow such pairs of mutually exclusive items to remain negatively correlated within any one class.) Since the results of this study may have been completely distorted by the inclusion of these item pairs, the results are not reported here.

A third study (Chen *et al.*, 2000) used data from an Epidemiological Catchment Area follow-up study. The authors chose a model with five classes, which they took to indicate the existence of four depressive sub-types (illustrated in Figure 2.2 on page 32); this is in contrast to the results of Eaton *et al.* described above. However, their analysis used 'lifetime' data focussing on 'reported symptoms during the worst episode', so it is not clear to what extent these differences may result from incomplete recall of depressive symptoms from mood episodes that may have occurred many years previously.

The most recent of the four papers (Mora *et al.*, 2012) focussed on older adults

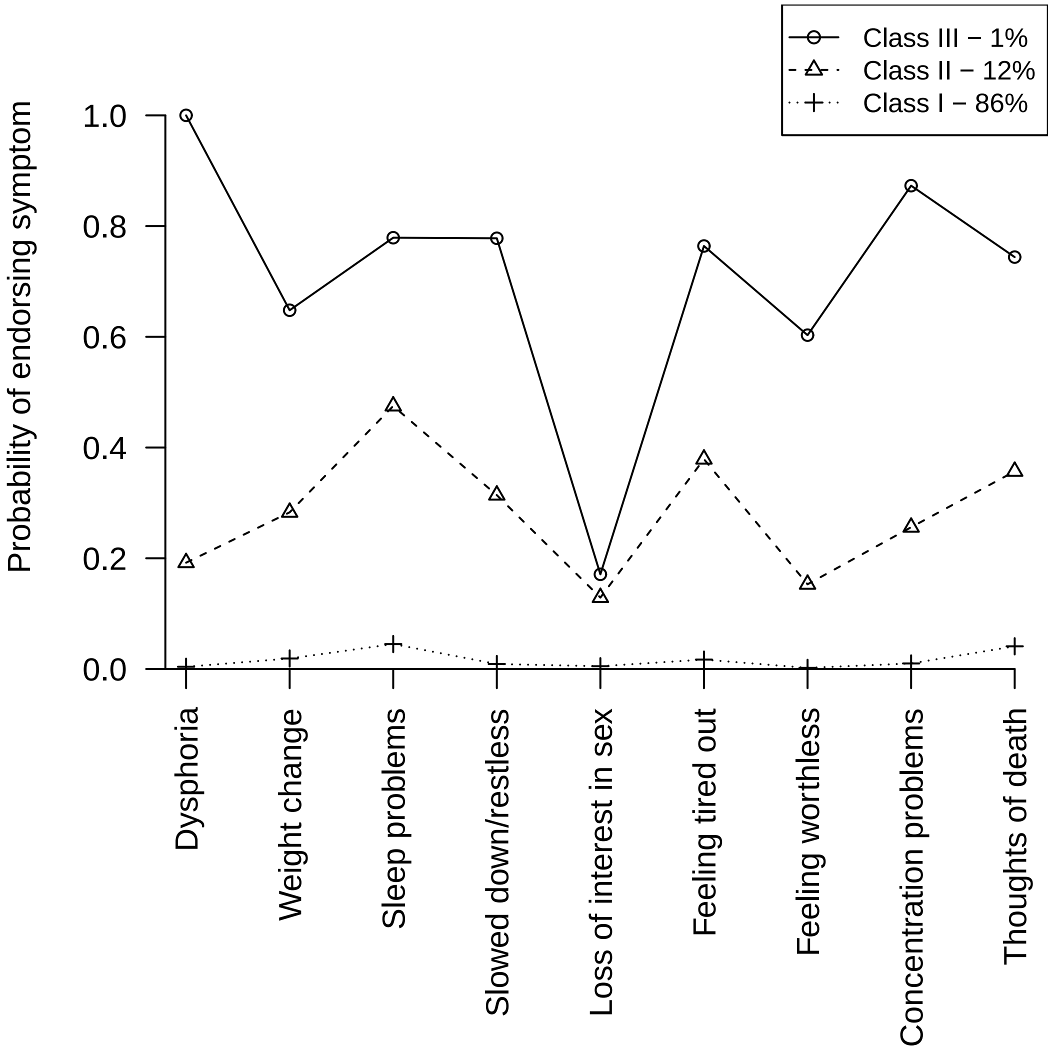

Figure 2.1: Latent classes in the symptoms of depression from the Eaton *et al.* (1989a) study

(virtually all of whom were over 65). As a result, their findings may not generalise directly to adults in the age range of 16-64 that are of interest in the current project. More problematically, the reporting of their preferred model is extremely limited — they do not actually report the item probabilities (or mean item scores) in each class for the 20 items of the Center for Epidemiologic Studies–Depression Scale (CES-D) on which their analysis is based, so the information about the symptom profiles of each class is rather limited. Nonetheless, they do report the mean scores on the four CES-D subscales for each class, and these are shown in Figure 2.3 on page 33.

Figure 2.3 suggests that the four classes in the model selected by Mora *et al.* differ primarily in terms of overall symptom severity, although one of the two 'intermediate' classes is marked out by a high mean score on the 'negative interpersonal

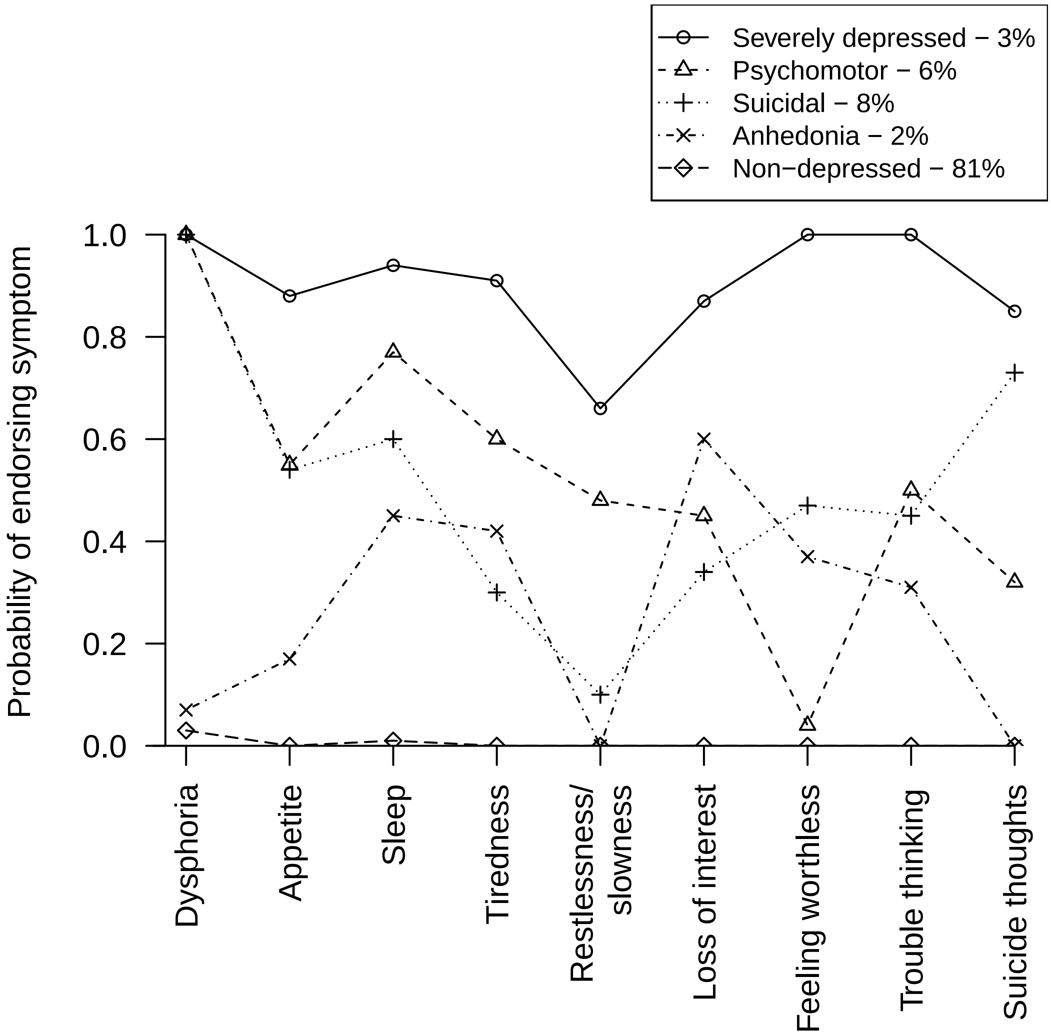

Figure 2.2: Latent classes in the symptoms of depression from the Chen *et al.* (2000) study

feelings' scale (which comprises only two items 'people were unfriendly' and 'felt people disliked me'). However, it is possible that the classes are further distinguished by differences in individual symptoms that have been obscured by the way the results were reported. These four papers that considered only symptoms of depression highlight some of the limitations present in many such studies, which limit their usefulness for informing psychiatric nosology.

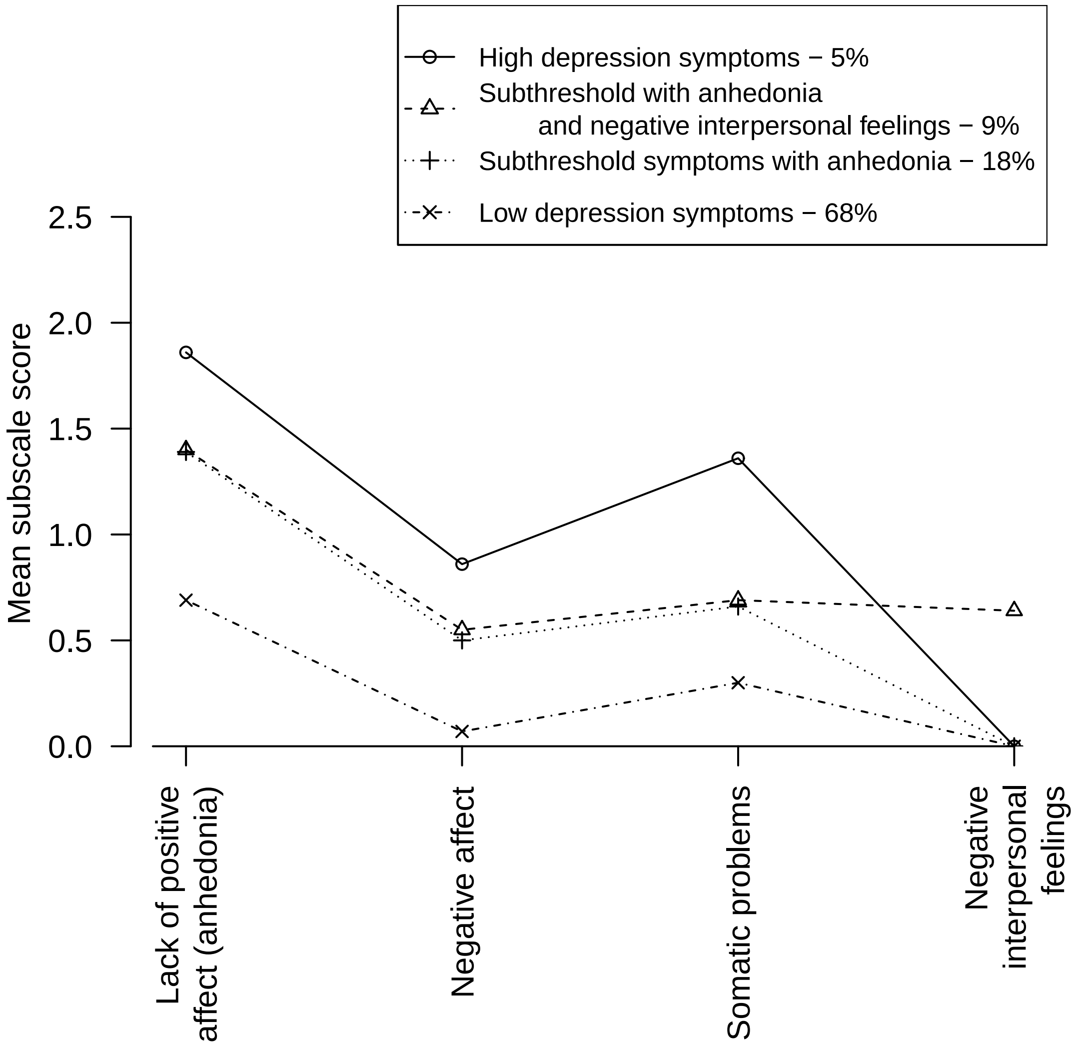

Figure 2.3: Latent classes in the symptoms of depression from the Mora *et al.* (2012) study

2.2.2 Papers looking at depression and anxiety

There are very few studies that have used latent class analysis to look at the symptoms of depression and anxiety together. A few papers, such as Sullivan & Kendler (1998) and Weich *et al.* (2011), have used latent class analysis to examine patterns of comorbidity among DSM-IV *diagnoses* of anxiety and depressive disorders (rather than looking at the actual symptoms contributing to these diagnoses). However, these papers ignore any symptoms that fall short of a full diagnosis and will be heavily influenced by the existing classificatory structure, to the extent that the natural groupings of symptoms may be completely obscured.

Papers focussed on children and adolescents

Of the papers that have looked at actual symptoms, three have focussed solely on children and adolescents (Wadsworth *et al.*, 2001; Ferdinand *et al.*, 2005; van Lang *et al.*, 2006), where models may apply that are very different from those that describe the adult population. Nonetheless, given the dearth of papers that have carried out similar investigations in adults, it is worth looking briefly at the results of their analyses.

Wadsworth *et al.* (2001) analysed two samples of children and adolescents aged 4-18 — one sample consisted of those who had been referred to outpatient mental health services, while the other sample comprised non-referred young people who were selected to be representative of the US population. In both samples and in both the younger and the older age groups, Wadsworth *et al.* (2001) found no evidence of separate anxiety or depression classes; all of the symptomatic classes represented a mixture of both anxious and depressive symptoms, and the classes tended to differ only in terms of severity.

Meanwhile, Ferdinand *et al.* (2005) looked at young people aged 11-18 who had been referred to outpatient mental health services. They estimated separate categorical latent variables for anxiety symptoms and for depression symptoms (so they obtained two sets of latent classes, one relating to the symptoms of depression and the other relating to the symptoms of anxiety). However, they estimated the two sets of latent classes as part of a single model — this meant that they could estimate the joint probabilities of belonging to each combination of the anxiety and depression classes. They found that there were no individuals in the 'affective problems' classes who were simultaneously in the 'no anxiety problems' class, which they took as evidence that affective problems do not occur separately from anxiety problems.

Finally, van Lang *et al.* (2006) specifically tested a confirmatory latent class model that was set up to allow groups of individuals to have mainly symptoms of depression or mainly symptoms of anxiety (as well as to allow groups of individuals to have both types of symptoms or few symptoms of either). For their analysis they looked at a population sample of Dutch young people aged 10-12. They found that very few individuals were allocated to the 'mainly depression' or 'mainly anxiety' classes (0.1% and 0.5% of the sample respectively). In their subsequent exploratory latent class model, all of their symptomatic classes showed raised levels of symptoms for *both* depression and anxiety. These results, along with those of (Wadsworth *et al.*, 2001) and (Ferdinand *et al.*, 2005), suggest that the symptoms of anxiety and depression tend to occur together rather than separately among young people.

Papers focussed on adults

Only three papers could be found that used latent class analysis to examine patterns in the symptoms of depression and anxiety among adults. Of these three papers, one was limited by a number of methodological issues (Unick *et al.*, 2009), including the fact that the analysis was limited to participants who responded positively to the screening questions for *both* depression and generalised anxiety disorder, and the fact that their analysis made use of 'lifetime' symptom data (with no effort made to restrict the analysis to symptoms that co-occurred as part of a single syndrome). This study used data from the US National Comorbidity Survey. The authors selected a model with seven classes — the results are presented here in tabular form (Table 2.2) because a plot of the conditional probabilities for so many classes is difficult to read. The class labels are those provided by Unick *et al.*.

Table 2.2: Conditional probabilities of symptom endorsement from Unick *et al.* (2009, adapted from Table 3)

	Class 1: 11% Low distress	Class 2: 11% Mild psycho-logical depression	Class 3: 19% Mild somatic anxiety	Class 4: 22% Psychological anxious depression	Class 5: 6% Anxious misery	Class 6: 15% Somatic depressed anxiety	Class 7: 16% Restless somatic depression
Depressed mood	0.87	1.00	0.93	0.98	0.88	0.99	1.00
Lack of interest	0.00	0.35	0.09	0.50	0.97	0.69	0.88
Decreased appetite/weight	0.00	0.32	0.14	0.54	0.20	0.87	0.63
Increased appetite/weight	0.00	0.09	0.09	0.18	0.44	0.05	0.57
Decreased sleep (D)	0.06	0.45	0.55	0.89	0.32	0.97	0.98
Increased sleep	0.02	0.18	0.01	0.18	0.52	0.15	0.34
Psychomotor retardation	0.00	0.00	0.04	0.22	0.12	0.24	0.80
Psychomotor agitation	0.00	0.03	0.04	0.28	0.08	0.32	0.48
Lack of energy	0.23	0.71	0.67	0.89	0.93	0.92	1.00
Feelings of worthlessness	0.12	0.86	0.62	0.69	0.94	0.82	0.92
Decreased concentration (D)	0.04	0.50	0.33	0.77	0.79	0.92	0.98
Thoughts of death	0.07	0.39	0.25	0.55	0.85	0.75	0.75
Feeling restless	0.55	0.62	0.98	0.99	1.00	1.00	1.00
Feeling fatigued (A)	0.25	0.26	0.65	0.68	0.96	0.78	0.97
Decreased concentration (A)	0.73	0.59	0.86	0.92	0.96	1.00	1.00
Feeling irritable	0.36	0.58	0.71	0.71	0.90	0.82	0.88
Muscle tension	0.19	0.04	0.53	0.39	0.43	0.72	0.73
Decreased sleep (A)	0.38	0.40	0.76	0.92	0.43	0.95	0.97
Feeling dizzy	0.07	0.03	0.50	0.17	0.51	0.83	0.63
Excessive sweating	0.06	0.03	0.60	0.13	0.59	0.89	0.66
Stomach problems	0.15	0.20	0.61	0.42	0.69	0.96	0.80
Dry mouth	0.12	0.02	0.59	0.11	0.69	0.99	0.44
Racing heart	0.33	0.20	0.86	0.63	0.75	0.96	0.90

Abbreviations: D = item came from depression section; A = item came from the generalised anxiety section of the interview.
Permission to reproduce this table has been granted by Wolters Kluwer Lippincott Williams & Wilkins.

Several aspects of these results appear confusing or unhelpful, to the extent that the classes are very difficult to interpret. Some respondents (see Class 7 in particular) appear to have reported both decreased and increased appetite/weight or both decreased and increased sleep — this is presumably since the interview asked about lifetime symptoms, so the opposing symptoms may have been experienced many years apart. Furthermore, where some symptoms, such as sleep and concentration problems, were included twice in the questionnaire (in both the depression and generalised anxiety disorder sub-sections), these duplicate symptoms were both entered into the analysis as if they were different symptoms (distinguished in Table 2.2 by D for depression and A for generalised anxiety symptoms). Such details mean that the results of this study provide limited information about the actual co-occurrence of symptoms of depression and anxiety.

A far more useful paper comes from a second analysis of data from the Epidemiological Catchment Area study (Eaton *et al.*, 1989b). This study did not filter respondents by their answers to any screening questions, and only considered symptoms that had occurred in the month prior to interview. Since the number of symptoms that could be included in the latent class analysis was limited at the time that the analysis was carried out by the available computing power, the authors were forced to group symptoms into content clusters for analysis. Their 'somatic depression' cluster included appetite problems, sleep, fatigue and concentration problems; the 'dysphoria' cluster included thoughts of death, feeling worthless, lost interest in sex, and feeling sad or hopeless; the 'somatic anxiety' cluster included fainting, breathing hard, heart beating hard, feeling dizzy or feeling weak; and their 'phobia' cluster included any symptoms of unreasonable fear related to agoraphobia, simple phobias or social phobias.

Figure 2.4 shows their unrestricted three class model applied to the Baltimore data from the Epidemiological Catchment Area study. While one class appears to resemble major depression, some symptoms of anxiety (such as phobia symptoms and being a 'nervous person') are also common in this category. The larger 'anxiety' class is characterised by the absence of dysphoric symptoms, however, the profile of other symptoms is similar to that of the depression class. There is no evidence of a 'pure' depression class, and the authors noted that the somatic symptoms of depression were not limited to those with dysphoric symptoms.

A second useful paper looked at data from the 2000 UK psychiatric morbidity survey (Das-Munshi *et al.*, 2008). After dichotomising each of the 14 symptom scores from the CIS-R interview, the authors chose a latent class model with five classes — they labelled four of these 'comorbid fear', 'distress', 'fatigued worry' and 'fatigue',

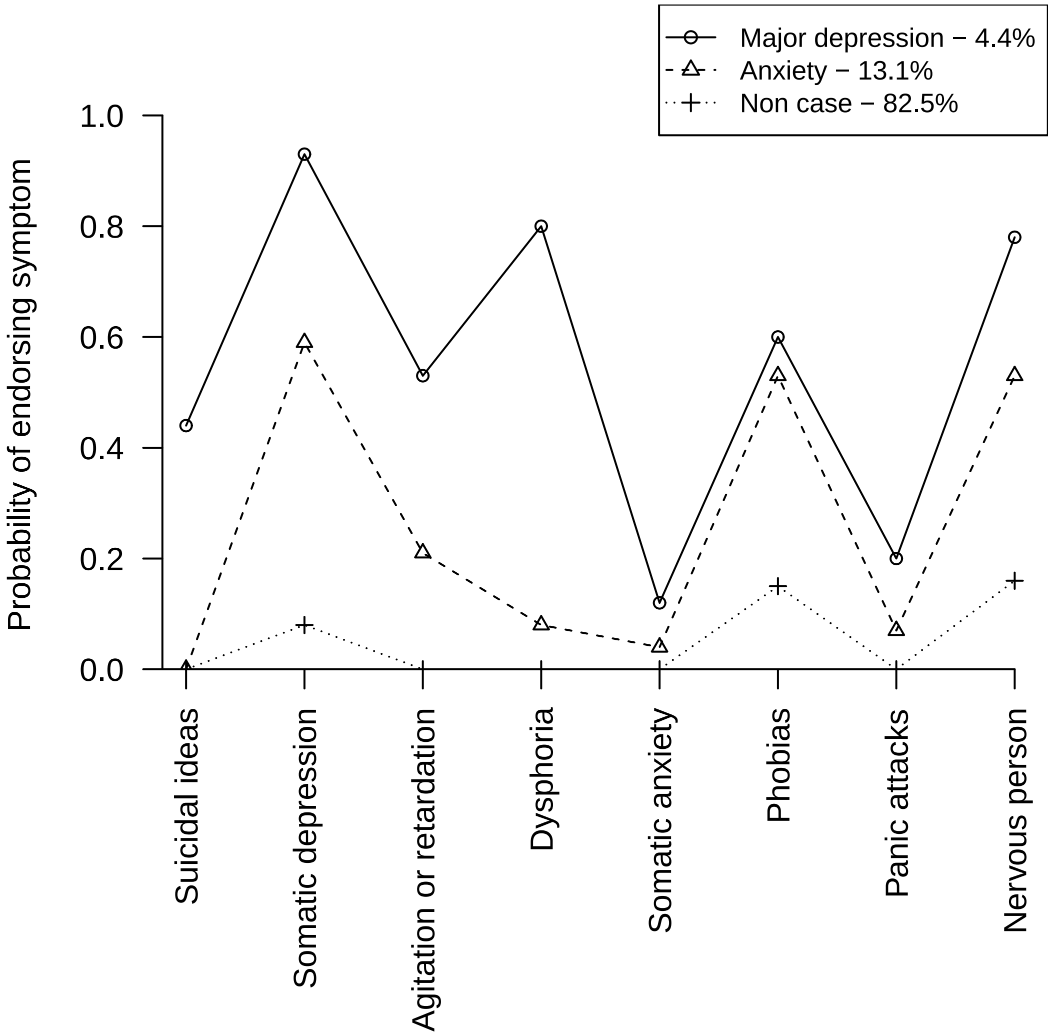

Figure 2.4: Latent classes in the symptoms of anxiety and depression from the Eaton *et al.* (1989b) study

and they considered the fifth class to be a 'non-case' group (see Figure 2.5). There is no evidence of either distinct depression or anxiety classes. Examination of the estimated symptom probabilities for each of the four symptomatic classes suggests that at least some of the differences between the first three classes might be viewed in terms of severity: the 'comorbid fear' class has the highest probabilities of endorsing each of the 14 symptoms, the 'distress' class forms an intermediate category, and the 'fatigued worry' class shows lower probabilities across most symptoms apart from worry. Only the fourth 'fatigue' class stands out as having a clearly different pattern of symptoms — here, fatigue and sleep symptoms have high probabilities, while the probabilities for most other symptoms are low.

This latent class analysis suggests that, if there are distinct types of common mental

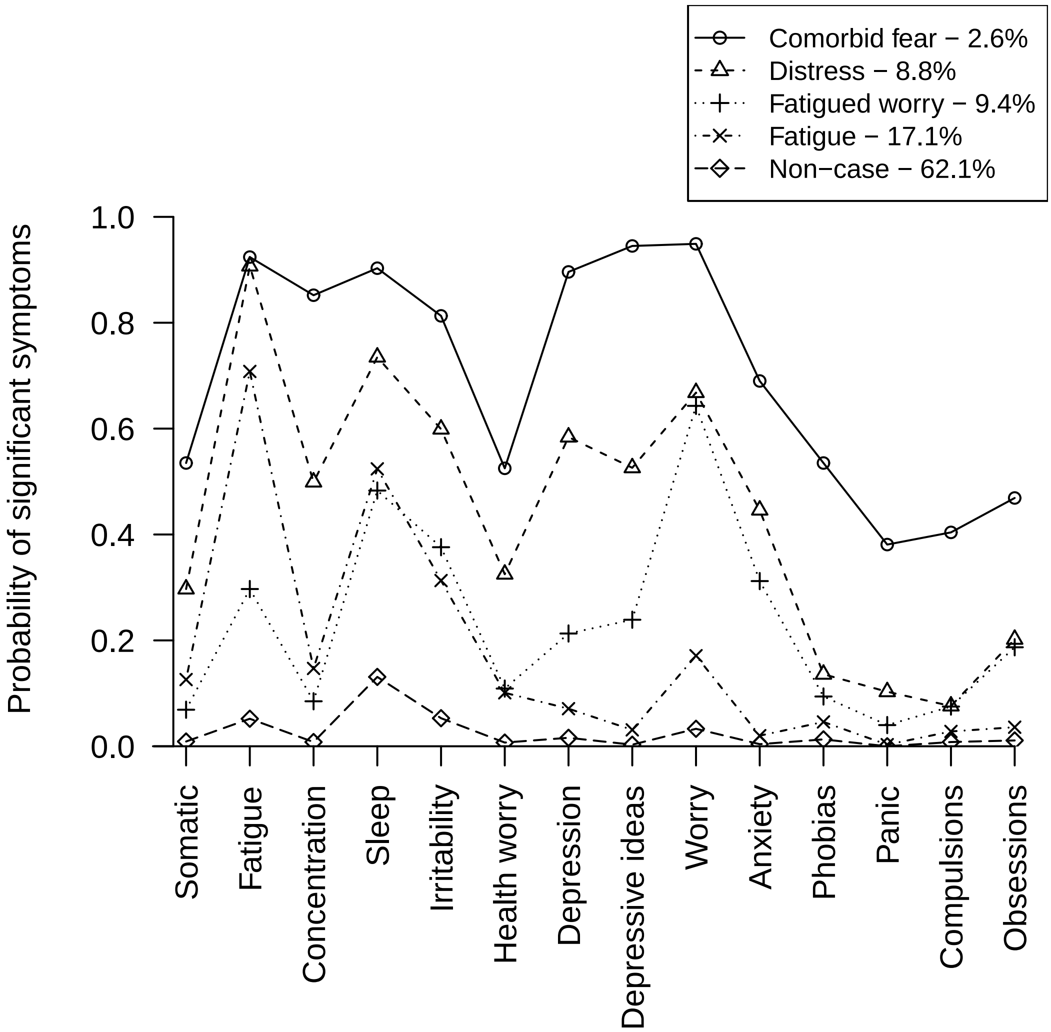

Figure 2.5: Latent classes in the symptoms of anxiety and depression from the Das-Munshi *et al.* (2008) study

disorder, these don't seem to conform to the traditional diagnostic distinctions between depression and anxiety disorders. However, it is less clear from this study whether the extracted symptomatic classes represent qualitatively different categories of anxious/depressive disorders, or whether these classes could be better viewed as part of a continuous dimension representing the severity of mental distress. Given that the data set used for this paper contributes to the pooled data set used in the current project, the analysis to come in later chapters will hopefully be able to answer this question.

2.3 Conclusions from the literature review

There are few papers that have examined patterns of symptom co-occurrence in anxiety and depression using symptom-level data, relevant statistical methods and representative population samples; of the few papers that have, some have suffered from major methodological limitations. Hence, there is a clear need for further high quality research into this area. However, the limited good material that is available suggests that depression and anxiety cannot be easily separated (particularly in young people), and provides little evidence to support the idea that depression frequently occurs without accompanying symptoms of anxiety.

Chapter 3

Data and methods

This chapter will start by describing the data to be used in this project, the characteristics of the sample and the nature of the measurement instrument used to measure the symptoms of common mental disorders. It will then go on to describe details of the statistical analysis that are relevant to the broad range of latent variable models used in this project. However, detailed descriptions of the model specifications for each of the various families of latent variable model will be saved for the chapters presenting the results of those models: this is to allow careful consideration of how parameter specification influences the interpretation of the results.

3.1 Data

3.1.1 The UK psychiatric morbidity surveys

The data to be used in this project come from three cross-sectional surveys of psychiatric morbidity in Great Britain carried out in 1993, 2000 and 2007. These surveys aimed to provide estimates of the prevalence of psychiatric disorders in the household population and to track these over time, as well as investigating other key areas such as levels of service use and the prevalence of risk factors and social disability arising from mental illness (Jenkins *et al.*, 2009). The 1993 survey covered the household population of England, Scotland and Wales; it was carried out by the Office of Population Censuses and Surveys, Social Survey Division (1996), and was funded by the Department of Health, the Scottish Office and the Welsh Office. The 2000 survey also covered the household populations of England, Scotland and Wales; it was carried out by the Office for National Statistics (2003), and was funded by the Department of Health, the Health Department of the Scottish Executive and the

National Assembly for Wales. The 2007 survey covered the household population of England only; it was carried out by the National Centre for Social Research & University of Leicester (2010), and was funded by the Information Centre for Health and Social Care. Anonymised data sets for all three surveys were obtained from the UK Data Archive.

The sample designs for the three surveys were very similar, although the sampling for each period was carried out independently. Each survey used a multi-stage stratified probability sample with proportionate sampling within strata; the strata were defined by region and the socio-economic characteristics of the postcode sectors that formed the primary sampling units (PSUs). Within each stratum, postcode sectors were selected with probability proportionate to the number of addresses they contained, which means that every address had roughly the same probability of selection within a survey (although there was some slight oversampling of Scotland in 2000 to ensure that some primary sampling units would be selected in the Highlands and Islands region). For 2007 only, where addresses contained multiple households, a single household was selected at random within each address. (In 1993, interviewers tried to interview every household in multi-household addresses. In 2000, interviewers selected a maximum of three households at each address, if necessary using random selection methods to select the appropriate number of households.) For all three surveys, a single respondent was selected at random within each eligible household. Sample sizes and numbers of primary sampling units for each period are shown in Table 2.

Table 3.1: Key sample information for each of the three survey periods

Year	Sample size	Response rate %	Number of PSUs	Addresses sampled in each PSU
1993	9,837	79	200	90
2000	8,580	69	438	36
2007	7,403	57	519	28

Abbreviations: PSU, primary sampling unit (postcode sector)

Although the target populations for all three surveys were similar (apart from variations in geographic coverage), the age ranges of individuals included in the survey varied. In 1993 the age range covered by the survey was 16-64, in 2000 it was 16-74, while in 2007 it included all adults aged 16 or over. Since mental disorders may be more difficult to assess in elderly individuals, and since elderly individuals may also differ in their likelihood of reporting key symptoms (Gallo *et al.*, 1994), the analyses in this project are limited to individuals aged 16-64.

3.1.2 The CIS-R interview

The symptoms of common mental disorders were measured in the three surveys using the revised version of the Community Interview Schedule (CIS-R; Lewis *et al.*, 1992). This is a standardised psychiatric interview that has been designed for administration by interviewers without any psychiatric training. The CIS-R interview covers the symptoms of a range of ICD-10 common mental disorders, and is structured around 14 key symptom areas. For each key symptom, screening questions are asked in order to determine whether or not a respondent has experienced that particular symptom in the previous month; if the respondent has experienced that symptom, they are then asked a series of more detailed questions relating to their experiences of that symptom *in the previous seven days*.

Based on their responses to the questions in each of the 14 sections, respondents are given a score indicating the severity of that symptom ranging from 0-4 (or 0-5 for the depressive ideas 'symptom'): scores of 2 or more may be seen as indicating significant symptoms. (Individuals who did not experience a particular symptom in the previous 7 days are given a 0 score for that symptom.) The scores for the 14 symptom areas can then be summed to give an overall CIS-R score that ranges from 0-57 — this may be seen as giving an indication of the severity of that individual's mental distress. Scores of 12 or more are traditionally seen as indicating a psychiatric case (Lewis *et al.*, 1992).

Alternatively, a set of algorithms may be applied to the CIS-R responses that allow for the diagnosis of a range of ICD-10 disorders (including depressive episode, generalised anxiety disorder, obsessive compulsive disorder and phobias). Descriptions of these algorithms and the disorders they cover are provided as appendices to the published reports for each of the three surveys (Office of Population Censuses and Surveys, 1995; Office for National Statistics, 2001; National Centre for Social Research, 2009b). The algorithms themselves can be examined as SPSS syntax in the data documentation supplied with the archived 2007 dataset (National Centre for Social Research, 2010).

The fact that the CIS-R interview focusses only on symptoms from the week prior to interview makes it particularly useful for studies examining patterns of symptom co-occurrence: recall bias is minimised, and there is no difficulty discerning which symptoms occurred together and which occurred weeks, months or years apart (as there may be in interviews focussing on symptoms experienced at any point in the lifetime or in the previous year.) Another strength of the CIS-R interview is that all respondents are asked about *each* symptom area: there is no screening to ensure that only those who are likely to meet the criteria for a particular disorder are asked about the relevant symptoms, as happens in some other standardised psychiatric interviews such as the Composite International Diagnostic Interview (CIDI). Nonetheless, there is one section of the CIS-R interview that has a slightly different structure, and where many respondents are *not* asked about a particular symptom area: this is the section on depressive ideas. Respondents are only asked about the presence of various symptoms grouped together as 'depressive ideas' if they received a score of 1 or more on the preceding 'depression' section.

A brief description of each of the 14 sections of the CIS-R interview is given in Table 3.2. A full list of the items which contribute to the section scores and overall CIS-R score is given in Appendix B. The complete question wording for the CIS-R interview can be found in the appendices to any of the three published survey reports.

3.1.3 Pooling the three data sets

Given the similarity of the sampling schemes for the three surveys in 1993, 2000 and 2007, and the fact that the CIS-R interview was used to measure the symptoms of common mental disorders in all three surveys, it is relatively straightforward to pool the CIS-R data from the three surveys. (There would be much more difficulty in pooling other parts of the surveys where there have been major changes to the measurement instruments used, but that is not a concern here since this project limits itself to the use of the CIS-R data and some very basic demographic information.) Pooling the data yields a much larger sample size than using any of the three surveys alone: this gives more power to detect potentially small 'disorder' groups, but it also makes it feasible to split the data to check for overfitting (as will be discussed in more detail in Section 3.2.3).

While combining the three CIS-R data sets is largely straightforward, there are still a few small issues relating to slight modifications in the wording of some CIS-R questions over the course of the three surveys. Since these changes are relatively minor, it was felt that they were unlikely to have any significant impact on the latent variable modelling in this project, and they have therefore been ignored. However, the changes are described in detail in Section A.1 of Appendix A for any reader who may wish to make up their own mind on this matter.

Combining the three data sets yields a total sample size of 25,820 individuals, although this reduces to 22,574 when only individuals aged 16-64 are included. The act of pooling the datasets makes the implicit assumption that the latent structure

Table 3.2: The 14 sections of the CIS-R interview

Somatic symptoms	Aches, pains or any sort of discomfort that was brought on or made worse because you were feeling low, anxious or stressed.
Fatigue	Feeling tired or lacking in energy for any reason other than physical exercise.
Concentration/ forgetfulness	Problems with concentrating on what you were doing or forgetting things.
Sleep	Problems with trying to get to sleep or with getting back to sleep, or sleeping more than usual.
Irritability	Feeling irritable or short tempered with those around you (over things that seem trivial looking back on them).
Worry about physical health	Worrying about your own physical health (all respondents) or worrying that you might have a serious physical illness (only respondents who didn't report a long-standing illness, disability or infirmity).
Depression	Feeling sad, miserable or depressed, or not being able to enjoy or take an interest in things as much as usual.
Depressive ideas	Feeling guilty, feeling hopeless, feeling not as good as others and thoughts of suicide (only respondents who scored 1 or more in the previous Depression section.)
Worry	Worrying about anything other than your own physical health.
Anxiety	Feeling anxious or nervous, or finding your muscles tense or that you couldn't relax.
Phobias	Feeling anxious, nervous or tense about any specific things or situations when there was no real danger, or avoiding any situation or thing because it would have made you feel nervous or anxious, even though there was no real danger.
Panic	Anxiety or tension getting so bad that you got in a panic (for example, feeling that you might collapse or lose control unless you did something about it).
Compulsions	Finding that you kept on doing things again and again when you knew you had already done them (for example, checking things like taps or washing yourself when you had already done so).
Obsessions	Having thoughts or ideas over and over again that you found unpleasant and that you would have preferred not to think about, that still kept on coming into your mind.

in the symptoms of common mental disorders was the same across the three survey periods. This seems to be a reasonable assumption: if a certain number of discrete disorders existed in 1993, it seems reasonable to assume that the same disorders existed in 2007. Likewise, if a particular continuous dimension of severity existed in 2000, it seems reasonable to assume that the same dimension existed in the other survey periods.

However, pooling the three data sets does *not* make the assumption that the prevalence of disorders remained constant over time, or that the distributions of symptoms along any continuous dimensions remained identical. If such variations did occur, any model estimated prevalences of 'disorder' classes (or of the means and standard deviations of the continuous latent variable distributions) will simply represent averages of those quantities from the three survey periods. Nonetheless, comparisons of the estimated prevalence rates for ICD-10 common mental disorders and the distributions of symptom scores from the three surveys suggest that any such changes were small (e.g., National Centre for Social Research, 2009a, Table 2.4).

3.1.4 Characteristics of the pooled sample

The mean age of the pooled sample aged 16-64 is 41.0 years (standard deviation: 13.2 years), and 45.2% of the pooled sample are male. The distributions of the 14 CIS-R symptom scores are shown in Figure 3.1. This indicates that the symptom score distributions are strongly skewed, with most respondents reporting that they had not experienced a particular symptom in the previous seven days. Fatigue, sleep problems, irritability and worry are the most prevalent symptoms, although the highest symptom scores are uncommon even for these symptoms. Symptoms of panic appear to be particularly rare.

The polychoric correlations between the 14 CIS-R symptom scores are shown in Table 3.3. The extremely high correlation coefficient of 0.87 between depression and depressive ideas may have been inflated by the fact that those with depression scores of 0 were not asked the depressive ideas section (and were given a default score of 0 for this section). Excluding this, the correlations range from 0.29 (for compulsions and sleep problems) to 0.69 (for worry and depressive ideas), indicating that all pairs of symptoms are at least weakly correlated.

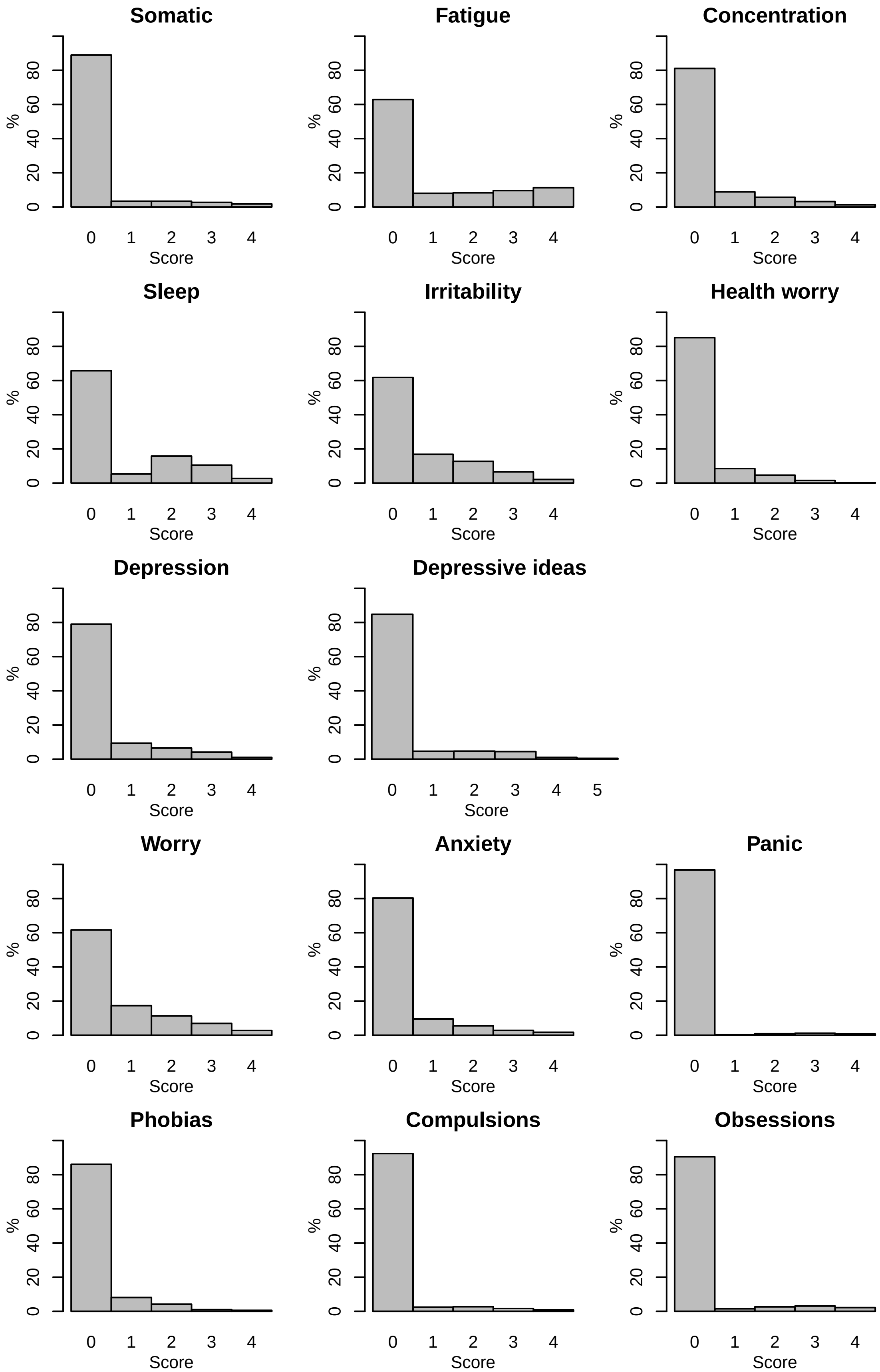

Figure 3.1: Distributions of the 14 CIS-R symptom scores among adults aged 16-64 in the pooled data set

Table 3.3: Polychoric correlations between the 14 CIS-R scores for individuals aged 16-64 ($n = 22,574$)

	Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Depressive ideas	Worry	Anxiety	Phobias	Panic	Compulsions	Obsessions
Somatic	1.00													
Fatigue	0.55	1.00												
Concentration	0.51	0.67	1.00											
Sleep	0.42	0.54	0.50	1.00										
Irritability	0.43	0.53	0.53	0.41	1.00									
Health worry	0.48	0.53	0.51	0.45	0.40	1.00								
Depression	0.49	0.59	0.61	0.51	0.53	0.50	1.00							
Depressive ideas	0.51	0.60	0.64	0.53	0.58	0.53	0.87	1.00						
Worry	0.49	0.52	0.56	0.46	0.54	0.47	0.62	0.69	1.00					
Anxiety	0.49	0.54	0.57	0.46	0.50	0.46	0.61	0.64	0.68	1.00				
Phobias	0.35	0.39	0.47	0.36	0.39	0.39	0.43	0.49	0.45	0.39	1.00			
Panic	0.49	0.51	0.58	0.46	0.49	0.49	0.59	0.63	0.59	0.62	0.62	1.00		
Compulsions	0.30	0.31	0.38	0.29	0.34	0.30	0.35	0.41	0.37	0.37	0.43	0.45	1.00	
Obsessions	0.35	0.38	0.43	0.37	0.38	0.38	0.49	0.53	0.52	0.47	0.42	0.44	0.38	1.00

Correlations greater than 0.6 have been highlighted in dark grey for ease of comparison, while correlations greater than 0.5 have been highlighted in a lighter grey.

16.9% of the pooled sample (3,819 individuals) have CIS-R scores of 12 or more (the standard cut-off used for identifying psychiatric cases; Lewis *et al.*, 1992). Table 3.4 indicates the prevalences and numbers of cases for a range of ICD-10 common mental disorders in the pooled data, as measured by the CIS-R interview. These prevalences are unweighted: they are presented to give an indication of the numbers of individuals experiencing diagnosable levels of common mental distress in the week prior to interview, rather than as estimates of prevalences in the population. The diagnosis of mixed anxiety and depressive disorder has no specific diagnostic criteria, but was given in the UK psychiatric morbidity surveys to individuals with a CIS-R score of 12 or more who did not meet the criteria for any other common mental disorder.

Table 3.4: Unweighted prevalences of ICD-10 common mental disorders in the pooled sample aged 16-64

Disorder	Prevalence %	Cases n
Depressive episode	3.3	749
Generalised anxiety disorder	5.3	1,194
Panic disorder	1.0	229
Obsessive compulsive disorder	1.7	373
Phobias	2.3	528
Mixed anxiety and depressive disorder	9.0	2,033
Any common mental disorder	18.4	4,160

Notes: The phobias category comprises agoraphobia (with and without panic disorder), social phobia and specific (isolated) phobia. The diagnostic algorithms have been harmonised to eliminate some minor discrepancies between disorder coding in the three archived data sets — for further details see Section A.3 of Appendix A.

3.2 Methods

Full descriptions and specifications of the different latent variable models to be used in this project will be given when each model is first introduced in the subsequent latent variable modelling chapters. However, there are a number of technical and methodological details that are common to many or all of the models used in this project; these issues will be documented in the following section in order to reduce the need for repetition in subsequent chapters.

3.2.1 Software

Since the three survey data sets were obtained from the UK Data Archive in Stata format, Stata version 10.1 (StataCorp, 2007) was used for extracting the relevant CIS-R and demographic variables, harmonising the variable names and coding of the three data sets before pooling, and also for providing the descriptive statistics reported in Section 3.1.4. The pooled data file was then converted to a format suitable for analysis in Mplus using the 'Stata2Mplus' program (UCLA: Statistical Consulting Group, no date).

Almost all of the latent variable models described in this project were estimated using Mplus version 6.1 for Windows with the 'Mixture' add-on (Muthen & Muthen, 2010b) — the only exception to this is a single model described towards the end of Chapter 7, where the use of the OpenBUGS software (Lunn *et al.*, 2009) is clearly noted. The 14 CIS-R symptom scores used as indicator variables in all models were declared to Mplus as ordered categorical variables. Unless otherwise indicated, all models were estimated using the 'MLR' estimator (a robust maximum likelihood estimator). All models requiring numerical integration for estimation used the Mplus default of adaptive Gaussian quadrature with 15 integration points per dimension.

The majority of models were estimated using a desktop PC with a powerful Intel Core i7 2.93 GHz processor, using the 64-bit version of Mplus and instructing Mplus to make use of 6 out of the available 8 cores with the `Processors = 6` command; the use of multiple cores in parallel reduced computation times dramatically. Running times for the models varied substantially according to the complexity of the model, with some models requiring only a few seconds while others ran for several weeks. A small number of the simpler models were estimated using the 32-bit version of Mplus 6.1 on less powerful dual core PCs. Examples of Mplus code that can be used to estimate each of the types of latent variable model described in subsequent chapters are given in Appendix D.

The graphs in this thesis were created in the R software environment (R Core Team, 2012) using the Mplus ".gh5" graphics files requested with the `Type = Plot3` plotting command. The graphics files were loaded into R using the 'HDF5' package (Daniels, 2009). (While the 'HDF5' package is no longer available from the CRAN package archive for installation into recent versions of R for Windows, it can be installed into older 32-bit versions of R such as version 2.11.0, and then manually copied into the package library associated with newer versions of 32-bit R.) For some charts, estimated model parameters were read into R directly from the Mplus output file using the 'MplusAutomation' package (Hallquist, 2012).

3.2.2 Use of random starts

One of the difficulties with estimating models with latent classes is that the results may be sensitive to the choice of parameter starting values. While the goal of maximum likelihood estimation is to identify the set of parameter values associated with the greatest likelihood of yielding the observed data (known as the *global* maximum likelihood solution), the estimation algorithm may 'get stuck' on inferior solutions (known as *local* maxima). For this reason, it is important to repeat the model estimation using a wide range of different starting values.

Mplus facilitates the use of multiple sets of starting values by allowing the user to specify the number of sets of 'random starts' to be used — the model estimation is then repeated on each set of starting values, and the results are presented from the solution associated with the highest model log-likelihood. Unfortunately, it is not possible to tell whether the best solution is itself merely a local maximum. However, if several different sets of starting values each converge to the same log-likelihood, this increases our confidence that we have in fact found the true global maximum.

In order to provide the best chance of locating the global maximum, models would ideally be estimated using a very large number of widely dispersed starting values. (The number and dispersion of starting values can be specified in Mplus using the 'Starts' and 'Stscale' commands respectively.) However, since model estimation must be repeated with each set of starting values, computation time will increase dramatically as the number of random starts are increased. Furthermore, when the starting values are widely dispersed it becomes more likely that some sets of starting values will contain implausible combinations of parameter values, and model estimation may become infeasibly slow in such cases.

Therefore, in this project a pragmatic approach was taken to the use of random starts. For models that ran relatively quickly (such as latent class models and factor mixture models with a single latent dimension) a large number of random starts were specified; for example, 2,000 sets of starting values might be run for 10 iterations, and then the 500 sets of starting values with the highest log-likelihoods at this point would be allowed to run to convergence. However, for the slowest models this would have taken months for each model. In this case, a much smaller number of starting values had to be used (such as 50 sets run for an initial 10 iterations, with 30 run to full convergence) — in some cases, model estimation still took two weeks.

In order to increase confidence in the solutions of such models, a number of steps were taken. For the initial run, the user-specified starting values were based on final parameter estimates from a simpler but related model, such as a model with the same number of factors but fewer latent classes. (The user-specified starting values provide the 'unperturbed starting value set' that is randomly perturbed by Mplus to give the sets of random starts.) Following completion of this run with the specified number of random starts, the parameter estimates from the solution with the highest log-likelihood were then supplied as the user-specified starting values for a second run, and the model estimation was repeated.

If the log-likelihoods for the best models changed substantially between the first and second runs, this process was again repeated. However, if the improvement was small or negligible, the model results were examined more closely. Where the best log-likelihood was replicated using several different sets of starting values, this was taken as support for a given model solution. Furthermore, the model solutions from each run were compared graphically — where the implied model was very similar despite small changes in the log-likelihood, this was taken as evidence that the general form of the model was fairly robust. Finally, the process of fitting models separately to split halves of the data (described in the next section) provided another opportunity to check whether the best solution was in fact a local maximum.

3.2.3 Splitting the data to check for overfitting

Since the modelling approach employed here is purposely exploratory and includes models with large numbers of parameters, it therefore carries a high risk of 'overfitting'. Overfitting involves fitting models that are so closely tailored to the observed data that they actually model the random fluctuations in the data arising from sampling variation, in addition to the 'true' population relationships that the models are intended to capture. Given this high risk, it was decided to split the pooled data in half so that the second half of the data could be used to check for overfitting. If a model developed on the first half of the data fails to replicate when applied to the second half of the data, this provides a clear indication that overfitting may have occurred.

All model comparisons within the next three chapters are based on the same random sample of approximately 50% of the available cases, comprising 11,230 individuals. (The split was obtained by generating a random number between 0 and 1 for each individual in the sample, and taking those with numbers less than or equal to 0.5 for the first split half.) This means that, for any particular model, a model with the same number of factors and/or classes was fitted to the second half of the data in order to see whether the same pattern of factors and/or classes emerged.

Given the complexity of some of the models involved, and the fact that in some cases quite different estimates for a few parameters may still imply very similar models, any such comparison will be largely subjective. Therefore, the comparison procedure focussed on comparing graphical model summaries and class prevalence estimates, rather than on a detailed comparison of individual parameters. However, such a checking procedure should identify any major discrepancies between models from the two halves, and should therefore prevent misleading substantive conclusions being drawn from artefacts of overfitting. Nonetheless, this procedure does not remove the desirability of some kind of external validation for the results of the modelling process, since any threats to external validity relating to the measurement instrument or survey sample will be shared by both subsets of the data.

In order to obtain comparison models from the second half of the data that were as independent as possible from those for the first half, user-specified starting values were *not* copied between models for different halves of the data. This means that some of the more complex models were put through the same time-consuming iterative procedure to identify the global maximum likelihood solution in the second half of the data that was described in the previous section for the first half of the data, even though copying over starting values would have saved much time.

The exception to this process of 'independent validation' was when a model from the first half completely failed to replicate in the second half of the data. In this case, two final attempts at replication were made by specifying the model parameter estimates from the first half of the data as user-specified starting values in the model for the second half of the data, and vice versa. Where a model is reported in the text as failing to replicate in the second half of the data, it will have still yielded divergent model solutions in the two halves of the data during both of these final replication attempts.

3.2.4 Excluding the depressive ideas score

All of the models in the next four chapters are based on only 13 of the 14 CIS-R symptom scores — the depressive ideas score has been excluded from all latent variable models. (The depressive ideas score covers feelings of worthlessness, guilt, and inferiority, as well as feeling that life is not worth living and thoughts of suicide.) The reason for this is that only individuals with a score of one or more for depression were asked this section, so it is impossible to include this score in any latent variable analysis *without making some form of assumption* about whether or not those symptoms should be present in other individuals.

A sensitivity analysis (described in Appendix C) suggests that the results of a key latent class model are highly sensitive to the particular assumption made about the unknown scores. Furthermore, it was felt that this sensitivity to the choice of assumption could seriously cloud the interpretation of the results of the latent variable modelling. For this reason, the depressive ideas score has been excluded from all latent variable models discussed in the next three chapters, even though this reduces the range of symptoms covered by the analysis.

3.2.5 Missing data

There is very little missing data in the CIS-R interviews for the 2000 and 2007 surveys, although there is a little more for the 1993 survey. This is likely to be related to the fact that the 1993 survey was administered using paper and pencil interviewing, whereas computer-assisted interviewing was adopted for the later surveys (so there was less scope for interviewer errors); this is discussed in more detail in Section A.2 on page 261 of Appendix A.

In the archived data sets, it is not always clear whether there is any missing data in the CIR-R interview. This is because the default assumption in the archived data is that missing items indicate an absence of the symptom being asked about; where this item contributes to one of the 14 symptom scores, the missing item contributes zero to the score for that symptom.

This same convention has been adopted during the latent variable modelling in the current project; no individuals have any missing symptom scores during analysis, since missing items are simply treated as contributing zero to the relevant symptom score. However, if some of these missing items correspond to symptoms that were not in fact absent, this may slightly distort the results. In order to check that the conclusions of this project are not sensitive to the way that the missing items were handled, several key models were re-estimated using only the data from individuals who had no missing items (complete case analysis). The results of these models were nearly identical to those from the full data set, indicating that the project conclusions are not sensitive to the method used to handle the missing data.

3.2.6 Acknowledging the complex survey design

The three surveys that have been combined in this project to create the pooled data set each used a multi-stage probability sample involving unequal probabilities of selection and cluster sampling. For many types of analysis it is important to use a method that accommodates these features of the sampling scheme (see for example, Heeringa *et al.*, 2010), and Mplus offers features for adjusting the model estimation to account for a complex survey design. However, these features have not been used for the models in the current project for two main reasons:

- There are two main concerns when ignoring the sample design: that standard errors may be underestimated (because of the clustering); and that certain types of individual will be under- or over-represented in the sample (because only one person was interviewed in each household, meaning that those living alone had a better chance of being included than those living in large households). These two concerns would be very important when trying to estimate the prevalence of a particular disorder in the population. However, they are less important for the current project.

  While the standard errors of parameter estimates can be important in latent variable models, the main focus of this project is on the structure and fit of the models rather than their actual parameter values. The conclusions of the project do not depend on the accuracy of the standard errors, and given that many of the models contain hundreds of parameters, few standard errors are even reported. Furthermore, any underestimation of standard errors due to clustering is not likely to be severe, since clustering occurred only by postcode sector — there was no clustering within households, which would tend to result in individuals within a cluster being much more similar (and therefore to greater inflation of the true standard errors).

  The over-representation of individuals living alone (and the under-representation of individuals from large households) should also not affect the validity of this study, since there is no reason to believe that the structure of mental disorders would be different in people with differing household sizes. However, the disproportionate sampling might lead to inaccuracies in class prevalence estimates *if the classes correspond to underlying groups in the population* — this should be borne in mind when interpreting such class prevalence estimates.

- The complex survey methods that are implemented in Mplus were developed with simpler regression models in mind, and there is no indication that they have yet been validated for use in hybrid latent variable models. Indeed, there has been little enough detailed investigation of the properties of factor mixture models and the methods for comparing them with other models when using the standard maximum likelihood estimators; there have been no papers to demonstrate that these properties still hold when pseudo-likelihood survey estimators are applied. Given that estimation can already be problematic for such complex latent variable mixture models, it may be wise to avoid adding a further layer of complexity unless there is a clearer indication that it is needed for a particular analysis.

Nonetheless, in order to check that the conclusions of this project are not sensitive to the decision not to use the Mplus survey commands, several key models were repeated specifying the inclusion of survey weights, the presence of cluster identifiers, and the restriction of the analysis to a subpopulation (those aged 16-64 only). The fitted models were very similar to those ignoring the survey design, although the estimated prevalences of the latent classes did differ slightly: the classes containing the individuals with fewest symptoms tended to be larger by about 4 percentage points. This confirms that the project conclusions are not affected by the decision to ignore the survey design during latent variable modelling.

3.2.7 Comparing models

Once a range of different latent variable models have been estimated, these models need to be compared in some way in order to decide which model or models are best suited for the objectives of the analysis. A popular approach when using factor mixture models is to use a range of fit statistics to try to identify the 'true' data-generating model that describes the true nature of the phenomenon under study. Several simulation studies have assessed the suitability of various fit statistics and significance tests for identifying the 'true' model. For example, Lubke & Neale (2006) focussed on factor models, latent class models and factor mixture models for continuous data; they found that it was possible to correctly identify the true data-generating model in many cases, although this did depend to some extent on the sample size for any real groups and their similarity to other groups that were present. In a later study, Lubke & Neale (2008) extended these results to applications with categorical data; they found a broadly similar result. However, the task of identifying the 'true' model appears to be somewhat more difficult with categorical as opposed to continuous data — it seems that very large sample sizes and quite distinct classes are needed to make such a judgement reliably.

This project will not follow the approach of comparing fit statistics in order to identify the 'true' model. This is for two main reasons:

- As Lubke & Neale (2008) reported, for categorical data (such as the CIS-R data) there may be low power to detect real classes. This is because of the large number of parameters that are needed for each latent class in the model; the exact number will vary according to how the model is specified, but in most cases more than 50 extra parameters are needed for each additional class. Given that the latent class corresponding to a mental disorder with low prevalence may be small, models that incorporate such a class may be rejected inappropriately even with the large sample sizes available in this project. It would be highly undesirable to dismiss such a model (which might provide useful insight into the structure of mental disorders) purely because this model fails to meet some pre-specified statistical criterion.

- The approach of searching for the 'true' model assumes that latent classes should reflect the presence of real underlying groups. However, there are other reasons why a model including latent classes may describe the data much better than a model without classes (as will be discussed in detail in Section 5.1.2 and subsequent chapters). While these situations are generally not considered in simulation studies such as Lubke & Neale (2008), it is possible that these 'untrue but important' classes may occur frequently in applied analyses, and such models may still be informative. Furthermore, it may be difficult to identify whether or not the classes are 'real' in any particular analysis.

For these reasons, it seems best to avoid focussing too much on the concept of a 'true' model; the focus instead will be on trying to identify the model or models that provide the best description of the data. If there are several models that appear to describe the data similarly well, the results of all of these models can be examined and compared — they may all be useful in their own way.

In order to provide a good description of the data, a model needs to fit the data well. However, it is also important to avoid adopting a model with too many parameters that is overfitting the data (and therefore modelling random noise). There are several fit statistics and tests that may help to identify whether one model provides a better description of the data than other. These fit statistics and tests have different emphases: some reward parsimony (favouring simpler models unless there is clear evidence that a more complex model is needed), while others focus purely on model-data fit. Furthermore, some of the tests require certain assumptions to give accurate results (such as that the form of the model has been correctly specified), while others do not. For this reason, different tests will frequently point to different models. However, when taken together, they will help to identify which models fit the data well without including unnecessary or misleading parameters. The rest of this section will describe a range of statistics that can be used to compare the results of the latent variable models reported in later chapters.

Bayesian information criterion

The Bayesian information criterion is widely used for comparing models in studies that apply factor mixture models to categorical data; this may result in part from Lubke & Neale (2008) identifying it as the best performing measure for identifying the 'true' model out of all the measures that they tested. The Bayesian information criterion is based on the model log-likelihood. The log-likelihood reflects how closely the model fits the data; higher values are better. However, the model log-likelihood can always be improved by adding parameters to the model, even though these parameters may simply be modelling the random noise in the data (overfitting). The Bayesian information criterion applies a penalty to the model log-likelihood to 'correct for' the number of parameters in the model. This penalty also takes the sample size of the model into account; the penalty becomes more and more stringent as the sample size increases.

The Bayesian information criterion is given by the following formula

$$\mathrm{BIC} = -2\ln(L) + p\ln(n), \tag{3.1}$$

where $\ln(L)$ is the log-likelihood value for the estimated model, p is the number of parameters estimated in the model and n is the sample size. The Bayesian information can be used to compare both nested and non-nested models — a lower value indicates the preferred model according to this criterion. A difference between models of 10 or more is said to be very strong evidence that the model with the lower Bayesian information criterion is the more appropriate model (Raftery, 1995).

Akaike information criterion

An alternative and widely used information criterion is the Akaike information criterion; this is given by

$$\mathrm{AIC} = -2\ln(L) + 2p, \tag{3.2}$$

where $\ln(L)$ is again the log-likelihood value for the estimated model and p is the number of parameters estimated in the model. Again, Akaike's information criterion can be used to compared both nested and non-nested models, and a lower value indicates the preferred model.

As can be seen from Equation 3.2, the Akaike information criterion makes no adjustment for sample size, and as a result will tend to favour more and more complex models as the sample size increases. This may lead to selecting an overfitted model. For this reason, it will not be reported for the models in this thesis — given the very large sample sizes used in this project, the Akaike information criterion virtually always favours the model with the largest number of parameters, even when the additional parameters appear to make little difference to the model fit (according to the other measures considered) or when this model does not replicate in the second half of the data.

Lo-Mendell-Rubin adjusted likelihood ratio test

The likelihood ratio chi-square test statistic (calculated as 2 times the difference between the log-likelihoods for the two models) is widely used to compare the fit of nested models. Since latent variable models that differ only in the number of latent classes are nested models, some form of likelihood ratio test can be used to test whether the inclusion of the additional class makes a significant improvement to the fit of the model. However, the standard likelihood ratio statistic cannot be used to compare models with k and $k-1$ classes. This is because such a comparison is in effect constraining one class prevalence parameter to be 0 in the smaller model; since 0 is on the very boundary of the possible values for the prevalence parameter, the test statistic no longer follows the standard chi-square distribution (McLachlan & Peel, 2000, p. 185). However, a couple of modified versions of the likelihood ratio test have been proposed specifically for comparing models with k and $k-1$ classes.

One such test is known as the Lo-Mendell-Rubin adjusted likelihood ratio test (Lo *et al.*, 2001); it is available in Mplus via the TECH11 command. This uses an alternative likelihood ratio test procedure that can accommodate a prevalence parameter being constrained to 0. As with other likelihood ratio tests, the test statistic is associated with a p-value; if this is smaller than a pre-chosen criterion (e.g., $p = 0.05$) we reject the null hypothesis that the two models fit equally well and conclude that the model with the additional class provides a better description of the data. This test does not require that either of the nested models to be tested are correctly specified. However, there is some uncertainty over whether the mathematical result underpinning this test is valid for its intended application (Jeffries, 2003).

Bootstrapped likelihood ratio test

An alternative modified version of the likelihood ratio test is known as the bootstrapped likelihood ratio test (McLachlan & Peel, 2000); it is available in Mplus via the TECH14 command. This uses the model with $k-1$ classes to simulate a number of bootstrapped data sets — these simulated data sets can be used to derive the empirical distribution of the likelihood ratio test statistic when the null hypothesis holds. This allows a p value to be calculated for the observed value of the likelihood ratio test statistic. In a simulation study comparing different methods for identifying the number of classes in latent variable mixture models, Nylund *et al.* (2007) found this test to be the most effective. However, unlike the Lo-Mendell-Rubin adjusted likelihood ratio test, the bootstrapped test assumes that the $k-1$ model is correctly specified when applied to the real data (which requires that all the assumptions of the relevant model hold).

A potential problem with the bootstrapped likelihood ratio test is that it can be extremely time-consuming to perform. This is because it requires the model with k classes to be estimated on each simulated data set; since these models all include latent classes, many different sets of starting values must be used for each simulated data set in order to avoid local maxima (as described above in Section 3.2.2). Since Mplus uses a sequential stopping rule to determine the number of simulated data sets to use (Nylund *et al.*, 2007), it is not even possible to estimate in advance how long the test will take to run — for many of the models used in this project, this could be many weeks. Therefore, the bootstrapped likelihood ratio test is only presented in this thesis for standard latent class models (those that do not incorporate a continuous latent dimension), since these are relatively quick to run.

Relative entropy

Relative entropy is a measure of the degree of certainty with which individuals in the sample can be classified into the latent classes of a model. While relative entropy is not an indicator of model quality, it can be used to help choose between models with different numbers of classes (particularly when a researcher intends to allocate individuals to their most likely latent class for subsequent analyses, in which case classification uncertainty may confuse the results of such analyses).

Relative entropy varies between 0 and 1 — a model in which individuals can be classified with certainty into one class or another would have a relative entropy of 1. Models in which many individuals have similar probabilities of belonging to two or more classes will have a much lower relative entropy. There are no defined cut-off points for what constitutes a good value for relative entropy, but it has been found that a value of around 0.8 or higher is desirable where a researcher wishes to use latent class allocations as a variable in further analyses (Clark & Muthen, 2009) — values of 0.6 may be adequate if the researcher uses a method that takes into account the uncertainty in the class allocations.

Bivariate goodness-of-fit

In addition to examining fit indices and significance tests based on the model log-likelihood, it is also possible to consider fit statistics which compare the data that would be expected under a particular model with the actual observed data. In particular, we can compare the distributions of the symptom scores 0–4 on each item that would be expected under the model with the distributions that are actually observed in the data. This can be carried out relatively easily on a univariate or a bivariate basis using the TECH10 command in Mplus. (In the bivariate case, this looks at how well the cross-tabulations of scores for each pairing of two items are recreated by the model.)

It is also possible to do this in terms of complete response patterns (looking at how well the frequencies of each observed combination of the 13 symptom scores match the frequencies that would be expected under the model). However, the sheer number of available response patterns for a dataset with 13 five-category items means that many of the observed response patterns have very small expected frequencies. This means that goodness-of-fit statistics and residuals based on complete response patterns are of limited use for a project such as this, so they also will not be considered in this report.

In general, the latent variable models employed in this study match the univariate proportions without difficulty, so the univariate goodness-of-fit statistics and residuals will not be reported here. However, an examination of the bivariate goodness-of-fit statistics and residuals can provide useful information on how well a model is able to replicate the relationships between pairs of symptoms that are seen in the data; they can also provide information about which pairs of symptoms are causing the biggest problems, if the overall fit appears to be poor.

The overall bivariate Pearson chi-square statistics are calculated by Mplus as the sum of the 78 bivariate Pearson chi-squares for each possible item pairing; these statistics are purely descriptive — a lower value indicates better overall model fit. However, since they cannot be compared against any known reference distribution, it is difficult to know how much of a cause for concern any large chi-square statistics would be without some indication of the range of values that could occur by chance for a correctly specified model. Nevertheless, they are useful as a quick means of comparing the fit of different models within a single model family, or as a quick means of comparing the different families of models that will be considered in the following chapters.

More useful information about exactly where the model misfit occurs comes from the bivariate Pearson chi-square statistics for each particular item pair — these chi-square statistics come from each of the contingency tables formed by cross-tabulating all possible pairs of items. Since the 78 bivariate contingency tables associated with each model share the total number of model parameters between them, there is no clear choice for the number of model parameters to specify when testing the goodness-of-fit of the observed frequencies versus the frequencies predicted by the model in each separate bivariate table. (The usual formula for the number of degrees of freedom when comparing observed and expected values in a contingency table would be the total number of cells in the contingency table minus the number of parameters in the model minus 1, but there are many more parameters in even the most basic factor model when applied to these data than the 25 cells in each contingency table.)

However, Bartholomew *et al.* (2008) suggest simply treating the model parameters as known — the degrees of freedom for the bivariate cross-tabulations would therefore be the number of cells in each contingency table minus 1 (df = 24). Bartholomew *et al.* (2008) also suggest using a p value of 0.01 for identifying item pairs with poor fit — this corresponds to a critical chi-square value of 43.0. The assumption that model parameters are known rather than estimated will make the cut-off slightly less stringent in identifying poorly fitting item pairs than the chosen p value implies. However, item pairs that are identified as fitting poorly would be associated with smaller p values were the correct degrees of freedom known.

Bivariate residuals

The measures described above summarise bivariate fit at the level of the model or the variable pair. However, at the most detailed level of scrutiny a bivariate standardised Pearson residual can be calculated for every cell in every bivariate cross-tabulation of scores (1,950 cells in total for each model) — these are also reported by Mplus using the TECH10 command. Standardised Pearson residuals greater than 3 can be considered to be ‘large’ (Agresti, 2010) — the numbers of these may be compared for each model, and so will be reported for each model used in this project. In a well-fitting model the number of such large residuals should be small, although a few values this large would be expected to appear by chance, given the large number of residuals for these models.

3.3 Overview of the latent variable modelling chapters

The main focus of this project is on the use of factor mixture models for exploring patterns of symptoms co-occurrence in the common mental disorders. However, given that these models combine elements of factor models and latent class models, it makes sense to start off by considering what these two types of model can tell us before moving on to the more complex models; this also makes it possible to see whether the complex hybrid models do indeed offer significant benefits in terms of model fit compared to the more traditional and simpler models. Nevertheless, since factor models and latent class models have been widely applied and thoroughly described in numerous papers and books, the focus in this thesis will be largely limited to drawing out the properties and implications of these models so that they can be contrasted with the hybrid models that follow; detailed technical description will be reserved for the factor mixture models, which are much less widely known.

The following three chapters will make use of each of the three different families of latent variable models in turn to explore the CIS-R data. The first chapter will consider the family of 'standard' factor analysis models, which assume that the latent structure is purely dimensional. The second chapter will consider the family of 'standard' latent class models that make the opposite assumption, namely that the latent structure is purely categorical. Following this, the third chapter will move on to the main focus of this project: the family of 'hybrid' latent variable models (factor mixture models), which allow a more flexible representation of the data by combining both dimensional and categorical latent structures within the same model.

For each of the three families of model, the chapters will consider what that particular family of models can tell us about the structure of the symptoms of common mental disorders, and which model within that family provides the most useful or best fitting summary of the data. Finally, a fourth chapter will focus on the complicated issue of interpreting the results of factor mixture models, before we move on to consider what these models can actually tell us about the structure of common mental disorders in the Discussion.

Chapter 4

Latent variable modelling I: Factor models

4.1 Introduction to the factor analysis family of models

The first broad family of models to be considered is the family of factor models. Factor models attempt to describe the relationships between the items in the data set in terms of one or more continuous latent dimensions. These models are frequently described as 'confirmatory factor analysis' (CFA) or 'exploratory factor analysis' (EFA) models; however, that terminology will be used sparingly in this thesis, since the latent variable modelling in this project is all exploratory in nature. The majority of models used in this project are of the type that are traditionally described as 'confirmatory', but this term will only be used to contrast these models with the small number of models that use the more flexible exploratory factor analysis approach.

Factor models for binary or ordinal data are sometimes known as item response theory (IRT) models; while this project does use ordinal observed data, the 'factor model' description will be preferred here, since this emphasises that similar considerations apply whether the observed data are continuous or categorical. In the chapters to come, factor models will frequently be described as 'standard' factor models: this is to contrast them with the more complex factor mixture models, which may be considered as extensions of the factor models used in this chapter. Detailed technical presentations of the factor model family are provided in many texts, including Bartholomew *et al.* (2011), Brown (2006) and de Ayala (2009).

Factor models tend to require fewer parameters than models involving latent classes in order to describe the latent structure within a particular data set. The main reason for this is that this is a 'parametric' family of models: the task of reproducing the observed relationships in the data is simplified by making a number of important assumptions about the distribution of the latent variables in the population, and about the relationships between latent variables and the observed symptom scores. These assumptions are described below.

In factor models, the latent variables (or factors) are continuous, and are assumed to follow a (multivariate) normal distribution: individuals' factor scores are assumed to be normally distributed, so that they can be easily summarised by only two parameters (the mean and variance of the factor score distribution). Each factor is assumed to be linearly related to any other factors, so the relationships between the factors can be easily summarised by a single parameter (the covariance or correlation coefficient). Furthermore, since the standard factor model for binary or ordinal categorical data takes the form of a generalised linear model, individuals' positions on the underlying latent dimension are assumed to be *linearly* related to the probability of an individual endorsing a particular symptom (after applying a relevant link function such as the logit or probit link).

In addition, the standard factor model treats all individuals in the sample as if they come from a single homogeneous population, and assumes that the CIS-R symptoms perform equally well as measures of the latent dimensions in all individuals. The result of all of these assumptions is that complex relationships between symptoms can be summarised by estimating a relatively small number of parameters. However, when some or all of these assumptions are inappropriate, any chosen factor model may fit the data poorly, and may suggest misleading conclusions about the nature of the latent structure in the data.

In this chapter, factor models will be used to examine the structure of the symptoms of common mental disorders under the assumption that these symptoms follow a dimensional structure. While the primary reason for carrying out this analysis is to allow for comparison with the latent class and hybrid factor mixture models in later chapters, the dimensional analysis will also be of interest in itself, particularly since such models are so frequently applied to data of this type. The analysis will provide insight into how many dimensions are needed to adequately describe the data from the CIS-R interview; it will allow us to assess the appropriateness of approaches that assume unidimensionality of the data, such as combining the items into a single summed CIS-R score; and finally, it will also allow us to examine which sections from the CIS-R interview tell us most about the dimensions being measured, and whether any symptoms are related only weakly to the key latent dimensions.

4.2 The single factor model

4.2.1 Justification

Implicit in the widespread use of the CIS-R total score as a summary of individuals' experience of common mental distress is the assumption that these symptoms are largely unidimensional; that is, that all of the symptoms are tapping aspects of a single underlying dimension of mental ill-health (although there may still be some modest residual correlations between similar symptoms that are not fully accounted for by a unidimensional model). On the other hand, if some of these symptoms measured one aspect of mental distress while other symptoms measured a quite different second aspect, it would be much more informative to provide two separate summary scores — these might reveal important differences in the characteristics and consequences of the two types of mental distress.

Given this assumption of unidimensionality for the CIS-R symptoms, a useful starting point will be to fit a single factor model, and to see how well this fits the data. Robust maximum likelihood estimates of the loadings from a single factor model are shown in Table 4.1. (Sample Mplus code for the single factor model is shown in Section D.1 of Appendix D.) In this model, the standardised loadings range from 0.55 for compulsions up to 0.82 for panic, suggesting that all 13 symptoms relate at least somewhat to a single underlying latent dimension — this dimension may be considered to reflect the overall severity of common mental distress. Of all the symptoms, compulsions have the weakest relationship to the single underlying dimension; nevertheless, the corresponding standardised loading of 0.55 is still non-trivial. Since the factor loadings presented in Table 4.1 are standardised to share the same metric as those from exploratory factor analysis, they can be interpreted as correlation coefficients between the continuous variables implicitly assumed to underlie the ordinal symptom scores and the factor; the loading of 0.55 implies that 30% of the variability in compulsions scores can be explained by the underlying latent dimension (compared to 67% of the variability in panic scores).

4.2.2 Plausibility of the normality assumption

It is useful here to consider the plausibility of one of the key assumptions underlying the factor family of models — namely, that any latent dimensions are normally

Table 4.1: Standardised factor loadings from the single factor model

Symptom	Loading	SE
Somatic symptoms	0.68	0.01
Fatigue	0.74	0.01
Concentration/forgetfulness	0.78	0.01
Sleep	0.63	0.01
Irritability	0.67	0.01
Worry over physical health	0.64	0.01
Depression	0.78	0.01
Worry	0.76	0.01
Anxiety	0.77	0.01
Phobias	0.58	0.01
Panic	0.82	0.01
Compulsions	0.55	0.02
Obsessions	0.63	0.01

Notes: n = 11,230. The factor loadings have been standardised post-estimation so that they share the same metric as loadings from exploratory factor analysis.

distributed. (While little can be done in the standard factor model setting if this assumption turns out to be implausible, it may become important later on during interpretation of the results of the factor mixture analysis; for example, a model with a mixture of two or more normally distributed factors may appear to describe the data much better than a standard factor model simply because the underlying latent factor is not normally distributed.) A histogram of estimated factor scores from the single factor for the 11,230 individuals in the sample is shown in Figure 4.1. This suggests that the distribution of factor scores is not normal — there is a strong positive skew, and there is a discontinuity between the large bar containing those with the very lowest factor scores (the individuals with 0 scores on all CIS-R symptoms) and the bars containing other very low factor scores.

However, it is important to note here that this apparent skew and discontinuity may be artefacts of the method used to estimate the factor scores, and the fact that factor scores are undefined for individuals who scored 0 on every CIS-R symptom without the provision of some external information about the 'true' shape of the factor distribution in the population. The factor scores shown in Figure 4.1 have been estimated by the 'expected a posteriori' (EAP) method. This takes place as a separate process after marginal maximum likelihood estimation (MMLE) of the factor model has been completed, and utilises a Bayesian normal prior distribution.

For those individuals whose factor scores are defined (those who scored one or more for at least one symptom), the choice of prior may have some modest impact on their

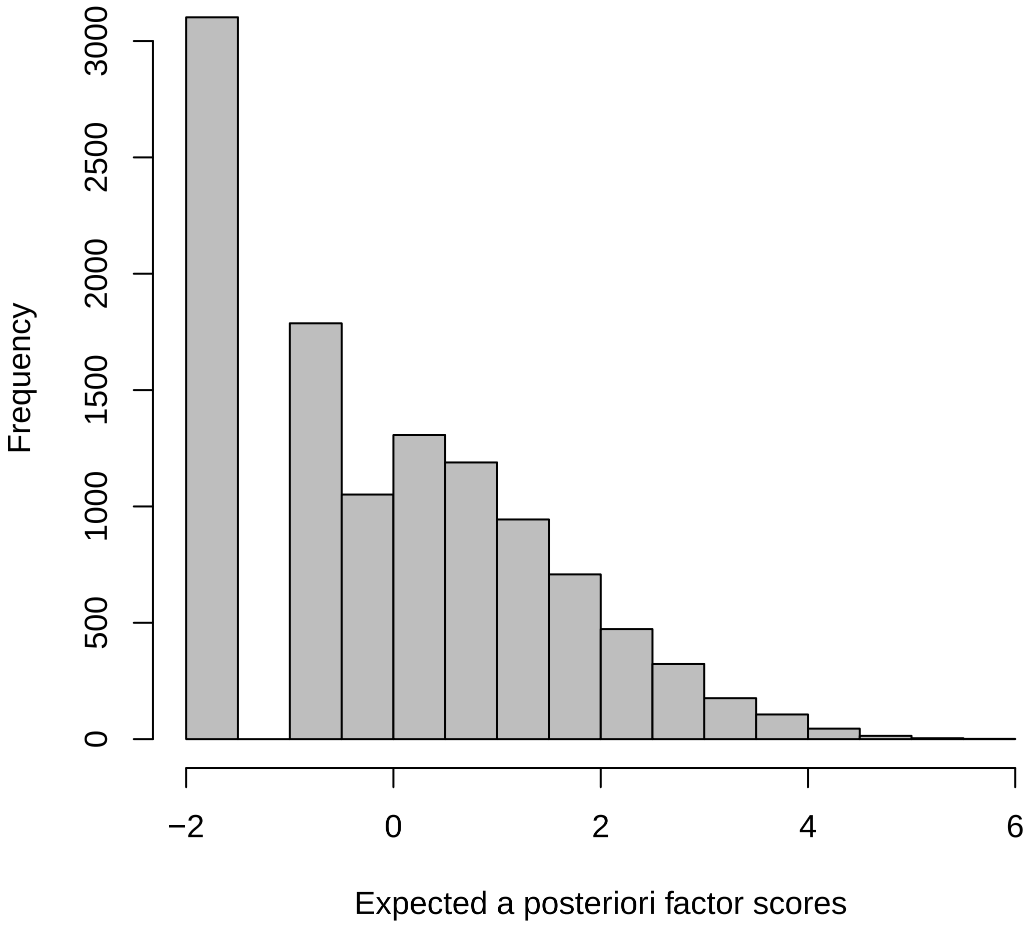

Figure 4.1: Expected a posteriori factor scores from the single factor confirmatory factor model

estimated factor scores — under a normal prior, those with high total scores (severe overall symptoms) may be expected to be affected by 'shrinkage' towards the 0 mean of the prior distribution. However, the 'undefined' factor scores of individuals with a 0 score on all symptoms will be *entirely determined* by the shape of the prior distribution (which cannot be altered in the Mplus software). Since all individuals with the 'no symptom' response pattern will be given the same estimated factor score, and since the overall factor mean is fixed to be 0, the symmetric normal prior here results in a block of factor score estimates for the undefined scores at the point along the latent dimension that 'balances' the well-measured upper end of the factor score distribution. This results in the discontinuity that is seen in Figure 4.1.

Although the distribution in Figure 4.1 is far from normal, this does not necessarily mean that the assumption of a normal underlying latent dimension is inappropriate. Indeed, the upper half of the factor score distribution seems to approximate the upper half of a normal distribution reasonably well. The observed distribution of factor scores may reflect the presence of a floor effect in the CIS-R data: the 'true' latent dimension in the population may actually include a range of negative factor scores that tail off to form the lower half of a symmetrical distribution, but these negative scores cannot be measured by the CIS-R instrument because it was not designed to measure the lower end of this dimension.

With the allowance for a floor effect, the assumption of a normally distributed factor appears to be feasible. However, this does not necessarily mean that a symmetric latent variable is conceptually appropriate. For example, it is not obvious what the interpretation of large negative factor scores should be, nor how to measure that end of the dimension; do these large negative scores represent well-being, resilience, or simply the complete absence of anything that might be considered a 'symptom'? Furthermore, it is still possible that some form of non-normal latent distribution would provide a better fit to the data for individuals located above the 'floor' of the latent dimension.

4.2.3 Plausibility of the unidimensionality assumption

Another key assumption of all unidimensional models is that the data really are unidimensional — if the data would in fact be better described by a two or three factor model, a single factor model will leave residual correlations between particular groups of items. Such residual correlations are inappropriate because maximum likelihood estimation assumes that individuals' responses to items are independent and uncorrelated, once their scores on the latent variable are taken into account. Stated another way, the item responses are assumed to be conditionally independent — all correlations between the items should be explained by the factors included in the model. Again, the plausibility of any assumption of unidimensionality may become important later when trying to interpret the results of factor mixture analysis — a mixture model may appear to fit the data much better than a standard factor model when the mixture is simply addressing the misspecification of the number of factors in the original factor model.

One way to assess whether a single factor is sufficient to account for the correlations in the data is to examine a range of fit indices that are widely used to assess whether or not a factor model 'fits' the data, and which are provided in Mplus when the limited information 'WLSMV' estimator is used. For this reason, the single factor model was re-estimated using this alternative estimator. (The parameter estimates were virtually identical.) Unfortunately, some of these fit indices are of limited use here because of the very large sample size. In particular, the chi-square goodness of fit statistic (which compares the observed and expected covariance matrices) may be misleading; with large sample sizes, this test may detect even small deviations from an exact fit (for simulation studies reporting similar effects with the Hosmer-Lemeshow test, see Kramer & Zimmerman, 2007 and Zhu *et al.*, 1996). For the single factor model, the chi-square goodness of fit test shows evidence of poor fit that is significant with $p < 0.0001$; however, it is not possible to tell from this test whether the lack of fit is of practical as well as statistical significance.

More useful guides to the fit of a particular model when the sample size is large are given by the Comparative Fit Index (CFI) or Tucker Lewis Index (TLI), both of which are based on comparisons of the chi-squared statistic for the hypothesised and the 'null'/independence models. The values of these fit indices range from 0 to 1, where values close to 1 indicate a very good fit — values greater than 0.95 are generally taken to be 'good' (Hu & Bentler, 1999). Other fit indices provided by Mplus are the Root Mean Square Error of Approximation (RMSEA) and the Standardised Root Mean Square Residual (SRMR). Values of the RMSEA smaller than 0.05 and values of the SRMR less than 0.08 are generally taken to indicate a close fit (Bartholomew *et al.*, 2008). The single factor model surpassed all four of these cut-off criteria, with a CFI value of 0.983, a TLI value of 0.979, an RMSEA value of 0.037 (90% CI = 0.035, 0.039) and an SRMR value of 0.040. Therefore, according to these criteria, the single factor model fits the data well.

A method that is commonly used to investigate how many dimensions are needed to describe the data is to look at the eigenvalues of the sample correlation matrix, which for these data are shown in Table 4.2. While there are no hard and fast rules on how to identify the correct number of factors, a widely-used suggestion known as the Kaiser-Guttman rule is to take the same number of factors as there are eigenvalues greater than 1 (see, for example, Brown, 2006). Since there is only one eigenvalue greater than 1 for the CIS-R data, this supports the use of a single factor model for these data. A justification for this approach is that any factor associated with an eigenvalue smaller than 1 explains less of the total variance than one of the individual variables on which the factor analysis is carried out. However, this argument is based on the assumption that a factor should represent a 'broad, fundamental dimension' (Horn, 1965, citing Dickman, 1960). Nevertheless, even where the addition of one or more factors adds only a small amount to the total variance explained, these extra factors may play a useful role in accounting for problematic residual correlations that violate the conditional independence assumption of the factor model.

Table 4.2: Eigenvalues of the sample correlation matrix for the CIS-R data

#	Eigenvalue
1	6.71
2	0.94
3	0.73
4	0.64
5	0.61
6	0.59
7	0.52
8	0.50
9	0.47
10	0.37
11	0.35
12	0.31
13	0.27

Parallel analysis (Horn, 1965) is another method that is widely used to investigate the dimensionality of a dataset. However, in its standard form, this is simply a more sophisticated version of the Kaiser-Guttman rule, and it therefore offers little here. The premise of parallel analysis is that the eigenvalues of the sample correlation matrix are only estimates of the true population eigenvalues, and therefore subject to sampling error; in a dataset where the variables are truly uncorrelated (and hence, all the population eigenvalues are actually 1), half of the eigenvalues would be expected to be greater than 1 by chance, simply because of sampling error. Parallel analysis creates simulated *uncorrelated* data sets (of the same size as the original sample); these are used to establish how large each of the ranked eigenvalues needs to be to suggest that a value greater than 1 is the result of more than just chance. Parallel analysis offers little in the present situation for two reasons:

1. As mentioned above, we are not only interested in factors with eigenvalues greater than 1, if there are minor factors that correspond to important residual correlations between symptoms in the single factor model.

2. Since the second observed eigenvalue in Table 4.2 is less than 1 anyway, parallel analysis would never suggest that it should be retained. (Parallel analysis is only concerned with eigenvalues that appear *greater than* 1 because of sampling error, not with eigenvalues that are inappropriately dismissed as being *less than* 1 because of sampling error.)

However, an alternative approach known as *modified parallel analysis* (MPA) may be more informative (Drasgow & Lissak, 1983; Finch & Monahan, 2008) — this has been used in the context of testing the unidimensionality assumption of item response theory/latent trait models. Rather than simulating data in which the variables are uncorrelated, modified parallel analysis simulates data that are truly unidimensional; the mean second eigenvalue of the simulated data sets can then give some idea of how big the second sample eigenvalue would be expected to be if the observed data were truly unidimensional. If the second eigenvalue of the observed data is significantly greater than the mean second eigenvalue of the simulated datasets, this is taken as evidence for the violation of the unidimensionality assumption.

Since no software could be readily found that would automatically perform modified parallel analysis for ordinal (as opposed to binary) items, the Mplus MONTECARLO procedure was used to generate 1,000 synthetic datasets of size 11,230, assuming that the single factor model estimated earlier in this chapter was the true population model. (The MONTECARLO procedure allows the user to define their own imaginary population where a specified model is known to fit perfectly; 'samples' of a specified size can then be drawn repeatedly from this imaginary population, mimicking the effects of sampling error during the 'real-life' sampling process.) The generated datasets were then imported to R, where the 'polycor' package (Fox, 2010) was used to calculate the polychoric correlation matrix of each data set, from which the eigenvalues were calculated for each data set.

The mean and 95th centile of the simulated eigenvalues are shown in Table 4.3. The second observed eigenvalue of the CIS-R data (0.94) is considerably larger than the 95th centile of the simulated second eigenvalues (0.47), implying that the unidimensionality assumption of the single factor model is violated to some degree. Since misspecification of the number of dimensions could lead to the extraction of spurious classes during subsequent factor mixture modelling, it would be helpful to investigate here whether there are any additional stable dimensions in the CIS-R data that could be incorporated into the factor model in order to address the violation of this assumption.

4.3 Exploratory factor analysis

4.3.1 Standard approach

Exploratory factor analysis (EFA) provides a way to explore whether there are any groups of items that are more closely related to each other than to the other items being modelled, and which could therefore be considered as factors themselves (and which may at least account for problematic residual correlations, even if the

Table 4.3: Mean and 95th centile of the eigenvalues from the 1000 simulated unidimensional datasets, along with eigenvalues observed from the CIS-R data

#	Observed CIS-R eigenvalues	Mean of simulated eigenvalues	95th centile of simulated eigenvalues
1	6.71	9.96	10.08
2	0.94	0.43	0.47
3	0.73	0.37	0.39
4	0.64	0.33	0.35
5	0.61	0.31	0.32
6	0.59	0.29	0.30
7	0.52	0.26	0.28
8	0.50	0.23	0.25
9	0.47	0.19	0.20
10	0.37	0.18	0.19
11	0.35	0.17	0.18
12	0.31	0.16	0.17
13	0.27	0.12	0.15

Notes: Sample size of observed and simulated data sets = 11,230.

amount they contribute to the total variance explained is small). With this aim in mind, an exploratory factor analysis was carried out on the covariance matrix of the CIS-R data with Mplus, using the 'WLSMV' (limited information) estimator and the default Geomin (oblique) rotation. Exploratory factor analysis solutions were requested with one, two, three and four factors.

The standardised factor loadings from the two and three factor solutions are shown in Table 4.4 and Table 4.5. In the two factor solution, the first factor appears to be dominated by fatigue, with some smaller loadings for other bodily symptoms, such as sleep and somatic symptoms, as well as for some more cognitive symptoms such as problems with concentration or forgetfulness and worry about physical health. On the second factor, the strongest estimated loadings are for anxiety-related symptoms such as worry, feelings of anxiety, panic and obsessional thoughts. However, there are also moderately strong loadings for irritability, depressed mood, phobias and compulsions. Several symptoms have a moderate loading on one factor and a small cross-loading on the other factor. The correlation between the two factors is 0.75. These two factors seem to be contrasting the general somatic symptoms of common mental disorders against the more cognitive worry and anxiety symptoms.

In the three factor solution shown in Table 4.5, the first factor again appears to represent general somatic symptoms, although with higher estimated loadings for irritability and depression. Similarly, the second factor still appears to represent

Table 4.4: Standardised loadings from the 2 factor exploratory factor analysis

Symptom	Factor 1	Factor 2
Somatic symptoms	0.39	0.31
Fatigue	**0.88**	0.00
Concentration/forgetfulness	**0.52**	0.33
Sleep	0.38	0.31
Irritability	0.26	**0.45**
Worry over physical health	0.38	0.30
Depression	0.27	**0.56**
Worry	-0.04	**0.84**
Anxiety	0.04	**0.76**
Phobias	0.02	**0.59**
Panic	0.02	**0.78**
Compulsions	-0.06	**0.59**
Obsessions	-0.12	**0.73**

Note: Estimated loadings greater than 0.4 have been highlighted in bold for ease of identification.

symptoms of worry and anxiety, although the estimated loading for panic has become much weaker (0.34 as opposed to 0.78), and some of the moderate loadings on the worry and anxiety factor from the two factor model have weakened (irritability, depression and compulsions) or disappeared (phobias). The third factor is dominated by phobic anxiety, although it also has moderate loadings for panic and compulsions. Again, there are several symptoms (for example, depression) with non-trivial cross-loadings. There is a strong correlation between the first two factors ($r = 0.82$), and there are also substantial correlations between factor 1 and factor 3 ($r = 0.61$) and between factor 2 and factor 3 ($r = 0.65$). The four factor solution yielded a factor on which only one item (sleep) had a loading greater than 0.2 — even this loading was small (0.33), so the four factor solution will not be considered further here.

The fit indices described earlier from the exploratory factor solutions with one, two and three factors are shown in Table 4.6. While all three chi-squared statistics indicate some degree of lack of fit (whether or not this is of practical importance), the models all appear to fit well according to the other fit indices — all three models exceed the recommended cut-off for each of the four measures. Nonetheless, the substantial improvement in chi-squared statistics for the two and three factor solutions (as well as smaller improvements in the other fit indices) suggests that the one factor model is ignoring some aspects of the relationships between the variables. This implies that there is some violation of the conditional independence assumption in the unidimensional factor model, relating to the presence of additional factors.

Table 4.5: Standardised loadings from the 3 factor exploratory factor analysis

Symptom	Factor 1	Factor 2	Factor 3
Somatic symptoms	**0.54**	0.13	0.02
Fatigue	**1.08**	-0.27	-0.01
Concentration/forgetfulness	**0.69**	0.00	0.17
Sleep	**0.52**	0.08	0.08
Irritability	0.38	0.28	0.06
Worry over physical health	**0.52**	0.03	0.14
Depression	**0.42**	0.38	0.02
Worry	0.02	**0.83**	-0.02
Anxiety	0.15	**0.70**	-0.04
Phobias	0.00	0.00	**0.81**
Panic	0.13	0.34	**0.45**
Compulsions	0.00	0.23	**0.41**
Obsessions	-0.06	**0.55**	0.19

Note: Estimated loadings greater than 0.4 have been highlighted in bold for ease of identification.

Table 4.6: Fit statistics for the 1, 2 and 3 factor exploratory factor analysis solutions

Fit statistic	1 factor	2 factors	3 factors
Chi-squared	1085	466	160
Degrees of freedom	65	53	42
P value	0.0000	0.0000	0.0000
CFI	0.983	0.993	0.998
TLI	0.979	0.990	0.996
RMSEA	0.037	0.026	0.016
SRMR	0.040	0.029	0.015

Abbreviations: CFI, comparative fit index; TLI, Tucker Lewis index; RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual.

This finding may become important when trying to interpret the results of factor mixture models with a single latent dimension, but several latent classes.

4.3.2 EFA in the CFA framework

While the results of the preceding section suggest that a single factor model may be adequate for the CIS-R data, they also suggest the existence of one or two additional factors that, while not of major importance, may lead to violations of the assumption of conditional independence if ignored. For this reason, it would be useful during factor mixture modelling to be able to compare factor mixture models with one, two and three dimensions, in order to see whether the desired number of classes changes when the residual correlations are taken into account. However, the exploratory factor analysis framework has substantial limitations for mixture modelling. In particular, since the exploratory factor analysis standardises all observed variables and ignores the mean/threshold structure of the data, it is not possible to investigate the different configurations of the factor mixture model that will be discussed in Chapter 6. Furthermore, the fact that it is not possible to save estimated factor scores or class membership probabilities makes it very difficult to investigate the meaning of any classes that are extracted. In addition, the opportunities for residual analysis and model checking are also greatly reduced.

For that reason, it would be desirable to be able to use confirmatory factor models to investigate mixtures involving two or three factors. Unfortunately, Table 4.4 and Table 4.5 indicate that the two and three factor solutions do not have a 'simple structure' (in which each item loads highly onto one factor only, while other loadings are close to zero) — there are many small but non-trivial loadings, and some symptoms (such as irritability and depression) cannot be assigned easily to one factor or another. For this reason, a confirmatory factor model specified using the standard approach (fixing loadings to 0 for most items on all factors apart from the primary factor for that symptom) would be likely to result in poor model fit, and would not meet the goal of reducing residual correlations. Therefore, as in the factor mixture analysis paper of Muthen & Asparouhov (2006), a more flexible approach will be applied here, specifying only the minimum number of parameter restrictions on the factor model that are needed to identify the model (as in the approach of exploratory factor analysis).

In general, m^2 parameter restrictions are necessary in order identify a model with m dimensions (Joreskog, 1969), so while only a single parameter must be fixed in a one factor model, 4 parameters must be fixed for the two factor model to be identified, and 9 for the three factor model. The parameters to be fixed can be chosen by identifying the item in the results of an exploratory factor analysis that has the highest loading on each factor (the anchor item for that factor); the loadings for this anchor item are then fixed to 0 on each of the other factors (but the loading does not need to be fixed for the factor where the item serves as the anchor item). The choice of anchor items determines the rotation of the factors, and by basing this choice on the results of an exploratory factor analysis, the rotation used by the exploratory factor analysis can be reproduced. The remaining m constraints can be achieved by fixing the variances of each factor to 1, or by fixing the loadings of the anchor items on their primary factor to 1 (these are two widely used but arbitrary choices for the scaling of the latent variable). If anchor items are chosen appropriately, the resulting 'confirmatory' model will closely match the rotation of the original exploratory factor analysis; this keeps the flexibility and good model fit of the exploratory approach, while allowing the more sophisticated examination of the fitted model permitted by the 'confirmatory' specification. This approach has become known as 'exploratory factor analysis in the confirmatory factor analysis framework' or E/CFA (for example, see Brown, 2006).

For the CIS-R data, anchor items were selected after examining Table 4.4 and Table 4.5 as fatigue and worry for the two factor model, and as fatigue, worry and phobias for the three factor model. (An alternative specification was considered in which somatic symptoms were selected as the anchor for the first factor in both of these models, since this allowed the negative loading for fatigue on the second factor in the three factor model to be preserved. However, this was dropped, since it made interpretation of the factors scores extremely difficult.) The standardised loadings for the two and three factor E/CFA models are shown in Table 4.7 and Table 4.8. (Sample Mplus code for the E/CFA model is provided in Section D.2 of Appendix D.)

The estimated loadings from the two factor E/CFA in Table 4.7 are very similar to those from the exploratory two factor solution in Table 4.4 on page 74 — the rotation appears to have been well-preserved, and the cross loadings have been maintained. The estimated correlation between the two factors is 0.74 — virtually the same as the 0.75 correlation estimated in the standard exploratory factor analysis solution.

The estimated loadings from the three factor E/CFA model in Table 4.8 are also similar to those from the standard three factor solution in Table 4.5 on page 75, particularly for the third factor. However, some of the loadings on the first and second factors have been affected by the decision to use fatigue as the anchor item for the first factor (and hence to eliminate the negative cross-loading of this item on

Table 4.7: Standardised loadings for the two factor E/CFA

	Factor 1	Factor 2
Symptom	Loading (SE)	Loading (SE)
Somatic symptoms	**0.41** (0.04)	0.32 (0.04)
Fatigue	**0.86** (0.02)	0*
Concentration/forgetfulness	**0.57** (0.05)	0.28 (0.05)
Sleep	**0.40** (0.04)	0.27 (0.04)
Irritability	0.28 (0.03)	**0.43** (0.03)
Worry over physical health	**0.42** (0.04)	0.28 (0.05)
Depression	0.29 (0.03)	**0.54** (0.03)
Worry	0*	**0.81** (0.01)
Anxiety	0.06 (0.03)	**0.75** (0.03)
Phobias	0.09 (0.06)	**0.52** (0.06)
Panic	0.06 (0.08)	**0.78** (0.07)
Compulsions	0.02 (0.07)	**0.55** (0.06)
Obsessions	-0.10 (0.05)	**0.75** (0.04)

* Anchor item — loadings fixed to 0 on all but its primary factor.
Note: Loadings greater than 0.4 are highlighted in bold for ease of identification.

Table 4.8: Standardised loadings for the three factor E/CFA

	Factor 1	Factor 2	Factor 3
Symptom	Estimate (SE)	Estimate (SE)	Estimate (SE)
Somatic symptoms	**0.42** (0.03)	0.29 (0.04)	0.04 (0.04)
Fatigue	**0.86** (0.01)	0*	0*
Concentration/forgetfulness	**0.57** (0.03)	0.15 (0.04)	0.17 (0.03)
Sleep	**0.41** (0.03)	0.20 (0.04)	0.08 (0.03)
Irritability	0.29 (0.03)	0.38 (0.03)	0.06 (0.03)
Worry over physical health	**0.42** (0.04)	0.16 (0.04)	0.15 (0.04)
Depression	0.31 (0.03)	**0.51** (0.04)	0.02 (0.03)
Worry	0*	**0.83** (0.01)	0*
Anxiety	0.08 (0.04)	**0.78** (0.05)	-0.03 (0.04)
Phobias	0*	0*	**0.80** (0.02)
Panic	0.06 (0.05)	**0.43** (0.07)	**0.47** (0.06)
Compulsions	0.01 (0.04)	0.21 (0.06)	**0.44** (0.06)
Obsessions	-0.09 (0.04)	**0.57** (0.05)	0.23 (0.05)

* Anchor item — loadings fixed to 0 on all but its primary factor.
Note: Loadings greater than 0.4 are highlighted in bold for ease of identification.

the second factor). In general, loadings on the first factor have become smaller, while those on the second factor (apart from worry) have become larger. The correlation between the first two factors has also become smaller (0.73 in the E/CFA, compared to 0.82 in the exploratory factor analysis). The correlations between the first and third factors, and between the second and third factors, are more similar (0.57 and 0.66 in the E/CFA, compared to 0.61 and 0.65). Again, the rotation of the standard exploratory factor analysis seems to have been preserved reasonably well. Therefore, these versions of the two and three factor solutions will be used during factor mixture modelling to consider whether the addition of further dimensions affects the choice of an appropriate number of classes.

4.3.3 Stability in split half

In order to check that the factors extracted during exploratory factor analysis are stable, and that the additional factors are not the result of over-fitting, a three factor exploratory analysis was repeated on the reserved half of the data. While there were some minor differences in the estimates (particularly for the smaller loadings), the pattern of loadings across the three factors was very similar, and the same anchor items were suggested for each factor. This implies that the three dimensions do reflect stable sets of relationships between particular groups of symptoms (as measured by the CIS-R interview). The estimated loadings for the three factor solution in the second half of the data are available for comparison in Appendix E (Table E.1 on page 289).

When a four factor model was fitted to the data, the fourth factor again had only one item with a loading greater than 0.2 — this was a loading of 1.12 for the obsessions item. This is a completely different result from that for the same model in the first half of the data, in which the sole loading over 0.2 was a loading of 0.33 for the sleep item. This finding confirms the absence of a fourth dimension in the data, or of any stable pattern of residual correlations that could be accounted for by such a dimension.

4.3.4 Graphical presentation of the three factor E/CFA

As in the case with the single factor model, it is useful to be able to examine the distributions of the estimated factor scores from the three dimensional factor model. The scatterplot matrix in Figure 4.2 shows histograms illustrating distribution of ‘expected a posteriori’ factor scores for each of the three factors (as measured by the loadings in Table 4.8), as well as scatterplots illustrating the bivariate relationships between each pairing of the factors. In Figure 4.2 the factors are labelled by their anchor items — this is simply for convenience, and is not intended to imply that this symptom is the 'essence' of the trait that is being measured by that dimension. All three factors appear to show the same kind of floor effect as seen in the unidimensional model. The correlations between the factors can also be clearly seen in this figure, particularly between the first and second factors ('fatigue' and 'worry' respectively).

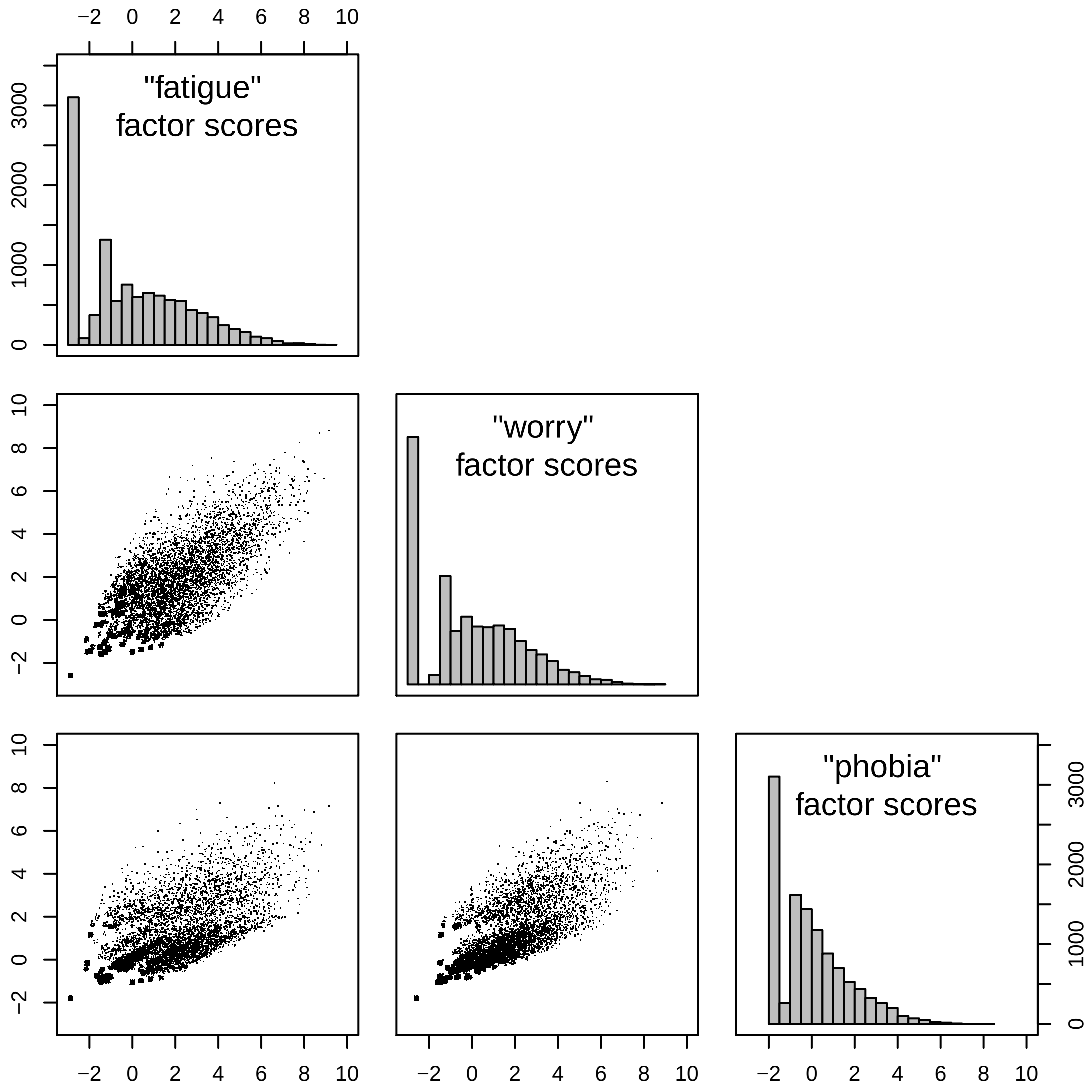

Figure 4.2: Scatterplot matrix of estimated factor scores from the three factor E/CFA. Points have been jittered by a small random value to reveal where the same estimated scores are held by many individuals

4.4 Goodness-of fit and residual analysis

One of the advantages of the E/CFA approach is that it allows us to examine goodness-of-fit statistics and residuals from the two and three dimensional exploratory factor solutions in as much detail as the residuals for the unidimensional factor model. Bivariate fit statistics (introduced on page 61) and residuals (introduced on page 62) are shown in Table 4.9. The overall bivariate Pearson chi-square statistics are large for all three models, although there is no way to assess the importance of such lack of fit from these chi-square statistics. Table 4.9 also indicates that there are many pairs of items showing significant lack of fit. While a few item pairs would be expected to show some degree of poor fit even in a well-fitting model, the values in Table 4.9 indicate widespread discrepancies between the observed and expected counts for all three models.

Although increasing the number of dimensions does result in some reductions in the overall bivariate chi-square and the number of item pairs showing poor fit, this step appears to be insufficient to ensure a well-fitting model. The largest bivariate chi-square statistic for each of the three models (with values of 211, 161 and 165 respectively) occurs for the 'worry-anxiety' item pairing. The second largest bivariate chi-square in each model occurs for the 'depression-anxiety' pairing.

Table 4.9: Bivariate goodness-of-fit statistics and residuals for the factor models

Model	Overall bivariate Pearson chi-square	Pairs with significant* lack of fit (out of 78)	Largest 5 bivariate Pearson chi-square statistics					Bivariate standardised residuals > 3 (out of 1950)
1f	3,974	35	211	132	132	110	106	95
2f E/CFA	3,588	34	161	127	115	103	95	69
3f E/CFA	3,258	27	165	128	103	93	90	55

Abbreviations: f, factor.
* P value used for significance cut-off = 0.01; critical value for χ^2 (24) = 43.0.

According to Table 4.9, the single factor model has 95 bivariate residuals greater than 3, again suggesting considerable problems with the fit of this model. While the addition of one or two further factors in the E/CFA models reduces the number of large residuals, the number of these large residuals still appears to be unacceptably high.

It appears that the factor family of models is limited in its ability to replicate the observed relationships between item pairs, particularly the ‘worry-anxiety’ and ‘depression-anxiety’ pairings. However, this lack of fit cannot be put down to the existence of further dimensions (whether meaningful or not), since exploratory factor analysis with four factors found no evidence of a stable fourth dimension. Furthermore, since E/CFA models allow for the greatest possible flexibility in the specification of the factor loading matrix, the poor fit cannot be ascribed to inappropriately fixing too many factor loadings to zero. It will be interesting to see whether any of the other families of latent variable models to be considered in later chapters can more closely match the observed relationships between each item pair.

4.5 Conclusion

The CIS-R data appear largely unidimensional; this justifies the widespread use of a single CIS-R score (created by summing the scores for each section) to summarise an individual’s experience of the symptoms of common mental distress. Nonetheless, the variations in factor loadings for the different symptoms imply that some symptoms (such as panic) provide more information about the severity of an individual’s distress than others (in particular, the presence of compulsions). However, modified parallel analysis suggests that the assumption of unidimensionality is violated to some degree, and the data could alternatively be viewed as arising from two or three highly correlated dimensions; the E/CFA models with 2 and 3 correlated dimensions had fewer large bivariate residuals and lower overall bivariate chi-square goodness-of-fit statistics. It was noted that ignoring these additional dimensions might have implications later during factor mixture modelling.

Since factor models assume normally distributed factors, these models all imply a strong floor effect in the CIS-R data. However, whether symmetric distributions for each factor are conceptually appropriate is not clear. Finally, goodness-of-fit statistics and residual analysis suggest that the standard factor models are limited in their ability to recreate the relationships between some pairs of symptoms that are observed in the data — an alternative family of models may be able to provide a better description of the data.

Chapter 5

Latent variable modelling II: Latent class analysis

5.1 Introduction

5.1.1 An overview of latent class analysis

After the discussion of the factor analysis class of models in the previous chapter, this chapter moves on to cover the latent class analysis (LCA) family of models. While factor models attempt to describe the relationships in the data through the use of one or more *continuous* latent variables, latent class models attempt to summarise these relationships through the use of a *discrete* latent variable with two or more categories (known as latent classes). While the term latent class analysis is most commonly applied to analyses based on a set of binary items, it can also apply to analysis with ordinal items (as used here); furthermore, latent profile models for continuous indicator variables can also be considered as belonging to the same broad family of models. Detailed technical descriptions of the latent class model are given in Bartholomew *et al.* (2011), Collins & Lanza (2009) and Hagenaars & McCutcheon (2002).

Latent class models are sometimes described by the more general term 'mixture models': rather than assuming that all individuals in a population can be described by a single probability distribution (as factor models do), latent class models describe the population using a *mixture* of distributions — each distribution corresponds to one of the latent classes. However, latent class models have one key property that distinguishes them from many other types of mixture model (such as the factor mixture models to be considered in the next chapter): latent class models do not allow any correlations or covariances between items among individuals in the same latent class, so items are assumed to be independent, *conditional on* an individual's latent class membership. (This is another example of an assumption of conditional independence, similar to that described in the previous chapter for factor models; in that case, items were assumed to be independent conditional on an individual's underlying factor score or scores.)

As a result of the assumption of conditional independence, latent class models do not allow for any underlying severity differences between individuals in the same latent class. Therefore, if there are any underlying severity differences present in the data (which seems quite likely when an ordinal scale has been chosen for the measurement of individual symptoms), the resulting covariances between items must be accounted for by the extraction of additional latent classes.

5.1.2 Direct versus indirect functions of latent classes

It is sometimes reported that latent class models assume the population of interest is heterogeneous and made up of a fixed number of separate and distinctive groups (Kendler *et al.*, 1996); consequently, the results of latent class analysis are often interpreted as implying that the extracted classes represent real and distinct groups in the population. However, as with other types of mixture model, latent classes can play two conceptually quite different roles; these two roles can be described as 'direct' and 'indirect' functions (Bauer & Curran, 2004). The direct function corresponds to the situation described above, where latent classes really do reflect distinct groups of homogeneous individuals in the population. However, in their indirect function, mixture components or latent classes are simply a means of approximating a single more complex distribution that applies to the whole population. The following paragraphs aim to describe this distinction more clearly.

Latent class analysis is actually a special case of finite mixture modelling, a statistical technique that can model many different types of data using a mixture of two or more probability distributions (McLachlan & Peel, 2000). The most obvious application for finite mixture modelling occurs when data really do arise from two distinct populations that each follow different distributions. For example, if blood glucose levels were measured in two populations — a sample of individuals who had been diagnosed with diabetes, and a sample of individuals who were known to be free of diabetes — the resulting combined distribution may well appear to be bimodal and very non-normal. The left-hand panel of Figure 5.1 illustrates a similar hypothetical situation: in this case, the two mixture components are both normally distributed, but the combined distribution would be very poorly approximated by a single normal curve. If the diagnostic codes for the blood glucose data were lost, a finite mixture model could be used to estimate the means and variances of the two original samples, as well as to make a probabilistic assessment of which individuals were likely to belong to each group. This is an example of a 'direct' application of mixture models, in which we are assuming that the two mixture components really do originate from different populations with distinct distributions of glucose levels.

Conversely, in 'indirect' applications there is no assumption that the component distributions arise from different populations — the mixture components are simply being used to approximate difficult univariate or multivariate distributions that cannot be easily modelled with standard tools, for example, a distribution that is skewed. The right-hand panel of Figure 5.1 illustrates a hypothetical example: here, a mixture of three normal distributions are being used to approximate an overall distribution that is skewed. However, it must be noted that the data themselves can never be relied upon to indicate whether or not the mixture components are playing a direct or indirect role. The right-hand panel of Figure 5.1 may also be describing data from three different populations whose distributions overlap considerably, while it is possible that the left-hand panel could be describing data from a single population, perhaps where individuals with extreme values were more likely to be selected for the sample.

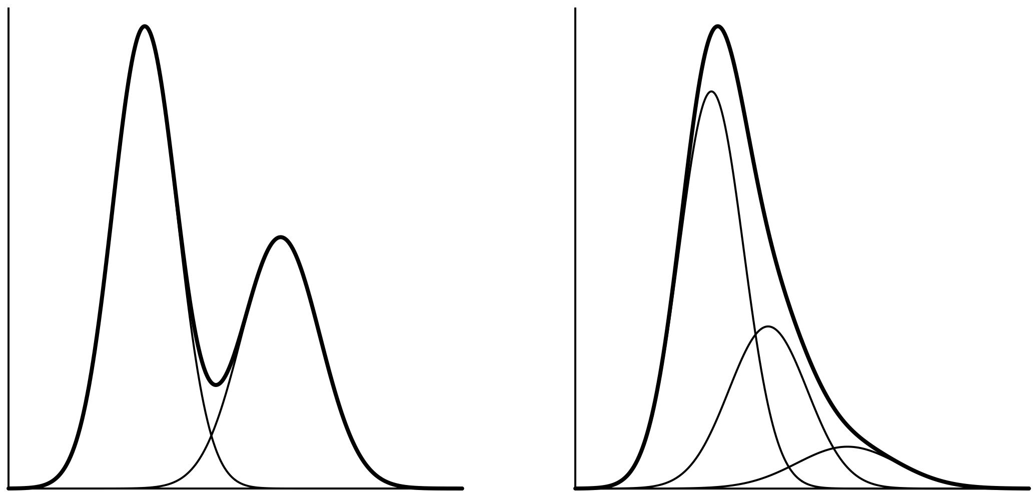

Figure 5.1: Two hypothetical examples of mixtures of normal distributions. The dark curves summarise the overall non-normal population distributions, while the lighter curves describe the normal distributions that combine to create the overall distribution. Although the mixture distribution on the left may appear to be performing a direct function, while the mixture on the right appears to be playing a more indirect role, the data themselves can never 'prove' such a conclusion.

In the case of the latent class models to be used here, classes may also be performing either a direct or indirect function. In their direct role, latent classes represent homogeneous and distinct groups of individuals, perhaps corresponding to particular disorders or clusters of symptoms. However, in their indirect role the classes can approximate a normal, skewed or irregular distribution of one or more latent traits, or they may indeed be representing a dimensional but discrete latent variable (that is, where the latent trait can only take on a limited number of real values). In any of these indirect roles, the latent classes are approximating the underlying dimensions with a series of discrete probability masses, each with 0 variance on the latent trait. Figure 5.2 represents the two possible functions of latent class models. The left-hand panel illustrate a set of classes performing an indirect role by approximating a normally distributed trait, whereas the right-hand panel illustrates a direct application where the data comprise of five distinct groups. (The dotted horizontal axis represents the fact that there is no common dimension being measured by these classes; they may be considered as unordered categories.)

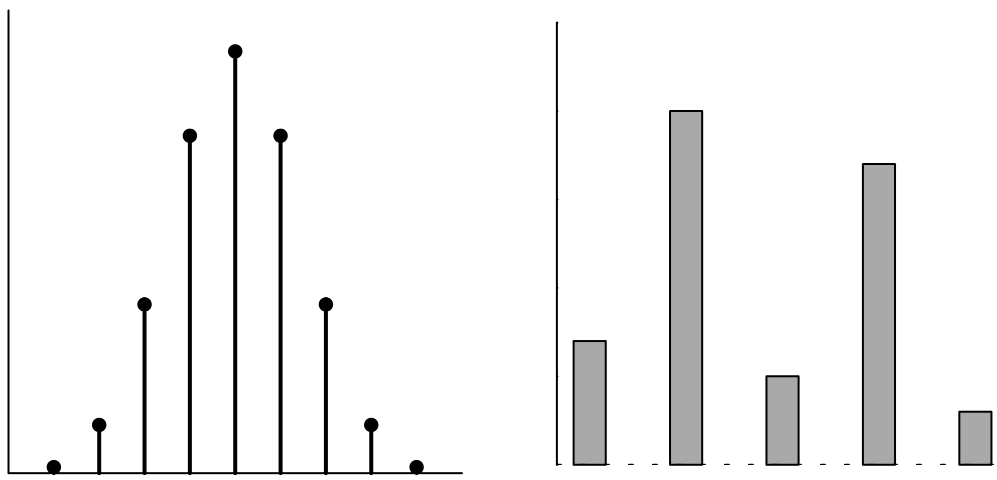

Figure 5.2: Representations of the roles that may be played by latent class models. The left-hand panel represents the indirect function of latent classes — the classes are approximating an underlying dimension. The right-hand panel represents the direct function of latent classes — the classes represent true homogeneous groups in the population, and the lack of x axis reflects the fact that the classes are unordered categories.

Given that latent class models can be used to approximate continuous dimensions, it is worth considering whether factor and latent class models can be used interchangeably. Indeed, in the case of continuous observed variables, it has been demonstrated that a factor model with f continuous factors can reproduce item means and covariances equally well as a latent profile model (a latent class model for continuous observed variables) with $f + 1$ classes (Molenaar & von Eye, 1994). In practice, maximum likelihood estimation tends to be used for both latent profile and latent class analysis — this uses the full information in the data from individuals' complete sets of item responses (rather than just considering the item means and correlations between pairs of items, which is known as a 'limited information' approach); therefore, it may be clear from the comparison of different models and checks of model fit that more than $f + 1$ classes are needed to provide a good description of the data when there truly is a continuous underlying dimension.

In the case of binary variables, it has been shown that that the latent class model with two classes is a special case of the latent trait/categorical factor model, in which the standard normal prior distribution has been replaced by a prior distribution with two discrete probability masses representing the two classes (Bartholomew *et al.*, 2011). A latent class model with more than two classes may be seen as simply approximating the univariate or multivariate distribution of one or more latent traits with additional discrete points. (Note that in the standard normal latent trait model a specific functional form *must* be assumed for the shape of the item response curve that links the probability of endorsing an item to scores on the latent trait — the response function is usually assumed to follow an 'S'-shaped curve described by the logit or probit function, as will be fully described in Section 6.2.1 of the next chapter. However, in a latent class approximation of a latent trait, no assumption about the form of the response function is required — this may be useful where the logit or probit response functions provide a poor fit to the data.)

As a result of this close relationship between the latent trait and latent class models, it may be impossible to distinguish from the data alone whether the true underlying latent dimension is actually continuous or discrete in nature — the expected frequencies predicted by the two models are reported to be often virtually identical (Bartholomew *et al.*, 2008). The problem is exacerbated by the fact that one of the best available methods for performing the complex mathematics required for maximum likelihood estimation of the latent trait model (a method known as Gaussian quadrature) actually simplifies the calculations by approximating the latent trait with a number of discrete points. Given these difficulties, if the choice is between a single dimension or two categories, a knowledge of the subject area will be needed to decide whether the latent construct is truly black and white (for example, pregnant versus not pregnant), or whether it makes more sense to expect the existence of shades of grey. This means that it will be virtually impossible to 'prove' from the data that the classes in a latent class model are performing a direct function, even if this is what a researcher believes.

5.1.3 Choosing between alternative descriptions of the same latent structure

Although it has been shown that *both* factor and latent class models can be used to describe dimensional data, this does not mean that the two types of model are equally desirable for performing this role. In contrast to the relatively parsimonious nature of factor models (resulting from the large number of assumptions made about the distributions of and relationships between variables), latent class models for ordinal items require the estimation of a greater number of parameters in order to account for the same relationships in the data. This is because an item with m ordered response categories requires the estimation of $m - 1$ intercept/threshold parameters for every latent class in the model (a total of 52 intercept parameters per latent class for the 13 symptom scores in the CIS-R data). Conversely, a factor model requires only a single set of intercepts in each model (again, 52 for the CIS-R data), along with a maximum of one factor loading per item for each dimension (13 parameters per dimension in the CIS-R data). Since *at least one more class* than the required number of factors will be needed to model the data adequately (as described in the previous section), the factor model will always be more parsimonious than a latent class description of the same data. However, where some or all of the assumptions of the factor model are badly violated, the latent class model may provide a more accurate description of the data — latent class models are very flexible since they require so few assumptions about the distributions or relationships in the data.

More parsimonious models are often to be preferred in statistical applications, (as long as they describe the data adequately) — this is because they require fewer parameters to be estimated, reported and interpreted, and they provide less scope for overfitting. Even so, the less parsimonious latent class models may provide greater utility when it is necessary to divide individuals into groups based on a latent variable for subsequent analysis. Beyond this, dimensional models can become unwieldy with large numbers of dimensions, and the computational requirements may become impractical for maximum likelihood estimation. Results may also be poorly received by other researchers if they consider there to be too few items loading strongly on each dimension for the factor scores to be measured reliably. Furthermore, the discrete nature of latent classes is often more consistent with the received conceptions of a latent construct — this is particularly so in the field of mental health, where psychiatric disorders are treated by some classification systems (such as DSM-IV) as discrete diagnostic entities that are clearly distinct from the everyday fluctuations of normal experience.

Such considerations may provide a reason for preferring latent class models to other families of latent variable models, even in the absence of other empirical justifications. However, it remains possible that a latent class model may actually provide a better description of the data than other models that are available to a researcher — this is particularly likely to be the case where the assumptions of other models prove untenable. For all these reasons (as well as to enable comparisons with the results from other model families), this project will now employ the latent class family of models to explore the latent structure of the CIS-R data.

5.2 Exploratory latent class analysis

Unless a researcher has strong theoretical grounds to expect a particular number of latent classes in the data, it is usual to carry out latent class analysis in an exploratory fashion. Under this approach, models with a range of numbers of classes are fitted; the models are then compared by a selection of fit statistics, and the researcher then tries to decide which number of classes provides the best description of the data. Some fit indices that can be used to compare nested and non-nested latent variable models were described and discussed in Section 3.2.7 of Chapter 3.

This chapter reports the results of the exploratory latent class analysis carried out on the CIS-R data. Models were fitted with increasing numbers of latent classes, continuing until a model was reached that could no longer be satisfactorily estimated. (Sample Mplus code for estimating a latent class model is provided in Section D.3 of Appendix D.) The rest of this chapter will consider which of the fitted latent class models provides the best description of the CIS-R data, along with what the results of these models suggest about the structure of the symptoms of common mental disorders.

5.2.1 Comparing model fit indices

A selection of fit indices and model comparison statistics for latent class models with increasing numbers of classes are shown in Table 5.1 — each additional class requires an extra 53 model parameters to be estimated (52 ordered thresholds/intercepts, and 1 class prevalence). As would be expected, the log-likelihood improves in each subsequent model as the number of parameters available to describe the data increases. However, models with too many parameters may 'overfit' the data by customising the fit to accommodate random noise, and would therefore generalise poorly to new samples. The Bayesian information criterion (BIC; introduced on page 58) is based on the model log-likelihood, but attempts to penalise models for excessive complexity: lower values on the Bayesian information criterion are better. According to this criterion, the 4 class model is the best of all the fitted models for the CIS-R data.

Table 5.1: Model comparison table for latent class models

Model	# par	LL	BIC	Relative entropy	aLRT p value	BLRT p value	Smallest class size (proportion)
1c	52	-103,735	207,955	-	-	-	11,230 (1.00)
2c	105	-93,045	187,069	0.87	0.0000	0.0000	3,037 (0.27)
3c	158	-90,976	183,426	0.81	0.0000	0.0000	1,134 (0.10)
4c	211	-90,497	**182,961**	0.76	0.0000	0.0000	469 (0.04)
5c	264	-90,258	182,979	0.73	**0.0000**	0.0000	465 (0.04)
6c	317	-90,092	183,141	0.74	0.8025	0.0000	481 (0.04)
7c	370	-89,973	183,397	0.72	0.7603	0.0000	293 (0.03)
8c	423	-89,874	183,694	0.72	0.7818	**0.0000**	224 (0.02)
9c	476	-89,790	184,019	0.73	0.7623	0.6667	227 (0.02)
10c	Not identified						

Abbreviations: # par, number of parameters estimated in the model; LL, log-likelihood; BIC, Bayesian information criterion; aLRT, Lo-Mendell-Rubin adjusted likelihood ratio test; BLRT, bootstrapped likelihood ratio test; c, class.
All models fitted to the same random half of the data. n = 11,230.

There are also a couple of tests that focus on the improvement in fit from adding a class — these compare the log-likelihood of the model with k classes to the log-likelihood of the model with $k-1$ classes, and test whether the difference between the two models is significant. These two tests both suggest keeping more than the 4 classes supported by the Bayesian information criterion — the Lo-Mendell-Rubin adjusted likelihood ratio test (aLRT; see page 59) implies the need for 5 classes, while the bootstrapped likelihood ratio test (BLRT; see page 59) suggests that as many as 8 classes may be important for describing the data. None of these fit statistics or tests support the 9 class model.

Another helpful statistic to consider is relative entropy (introduced on page 60). For the latent class models considered here, relative entropy tends to decrease as the number of classes increases, although there is little difference between the models with 4 or more classes, which all appear to show reasonable but not great levels of classification certainty. This suggests that the relative levels of classification certainty provide little help to decide between the 4 class model suggested by the Bayesian information criterion, the 5 class model suggested by the Lo-Mendell-Rubin adjusted likelihood ratio test, and the 8 class model suggested by the bootstrapped likelihood ratio test. It may therefore be useful to examine the nature of the classes extracted by the different latent class models, and to see what kind of role the latent classes appear to be playing.

5.2.2 Examining the latent class solutions

The easiest way to examine the nature of the latent classes is to look at plots of the estimated probabilities for obtaining a particular symptom score within each class — these 'conditional probabilities' are simply a transformation of the class-specific intercept/threshold parameters estimated during model fitting. Figure 5.3 shows a profile plot with the class-specific probabilities of scoring 2 or more on each symptom from the 4 class model, along with the estimated prevalences of each of the 4 classes. This is the model that was supported by the Bayesian information criterion, which penalises models for too much complexity. What is particularly noticeable from Figure 5.3 is that the 4 classes are clearly ordered, and appear to differ mainly in terms of severity. This type of solution implies the presence of a single underlying latent dimension (which may be continuous or discrete), and may suggest that the latent classes are actually performing the 'indirect' function described in Section 5.1.2, approximating the probability distribution of a single homogeneous population.

Figure 5.4 shows the estimated probabilities from the 5 class model supported by the Lo-Mendell-Rubin adjusted likelihood ratio test. Again there are clear severity differences between the classes. However, this time there appears to be one class (class 4) with a distinct profile that actually crosses the profile of another class (class 3). This suggests that, as well as the four ordered classes described in the 4 class model, there may be a distinct class of individuals with high levels of fatigue and sleep problems, but low levels of most other symptoms. This type of crossing symptom profile is sometimes considered as evidence that latent classes are performing the 'direct' function described in Section 5.1.2, characterising distinct subgroups of individuals (e.g., Silvia *et al.*, 2009). However, Pickles & Angold (2003) caution against jumping to this conclusion, warning that this grouping may simply reflect the presence of more than one underlying latent dimension in the data. In line with this more cautious interpretation, the crossing class profiles of class 3 and class 4 may reflect the attempt of the latent class model to account for the highly correlated 'fatigue' and 'worry' dimensions seen in the exploratory 2 factor model in the previous chapter.

Figure 3 shows the profile plot for the latent class model with 8 classes, as supported

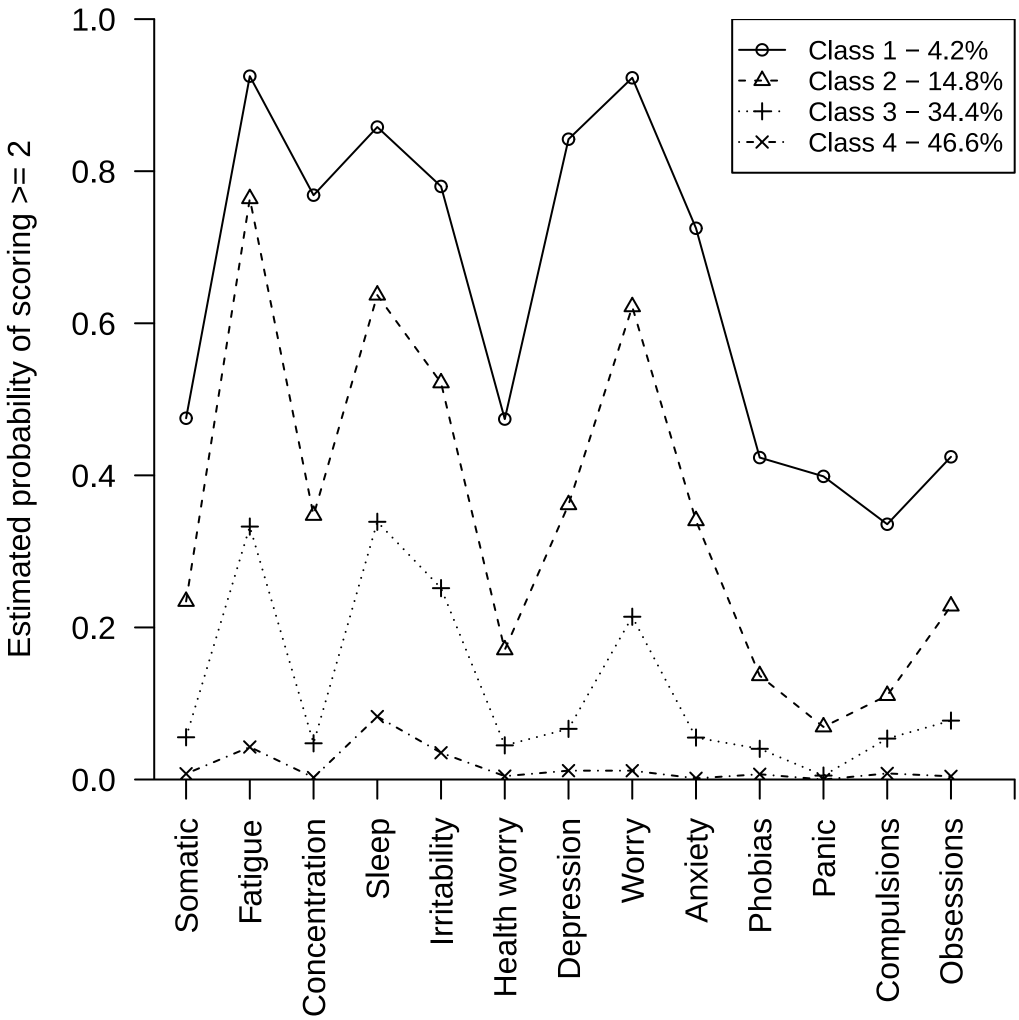

Figure 5.3: Estimated probabilities of scoring 2 or more on each symptom by class for the 4 class latent class model

by the bootstrapped likelihood ratio test. (Since the 8 class profiles are difficult to see in a single plot, they have been divided between two panels to allow the nature of the classes to be seen more clearly.) The panel on the left shows that 5 of the 8 classes appear to be modelling overall differences in severity. Meanwhile, the panel on the right indicates that there are now two ordered classes representing high fatigue and sleep symptoms with relatively low levels of other symptoms, while there is now also a class with relatively low levels of fatigue but high levels of worry. These three classes are again consistent with the 'fatigue' and 'worry' dimensions found in exploratory factor analysis, as discussed for the previous model.

Although the profile plots in Figure 5.3, Figure 5.4 and Figure 5.5 have focussed on the dichotomy of symptoms scores < 2 versus symptom scores ≥ 2 (which simplifies

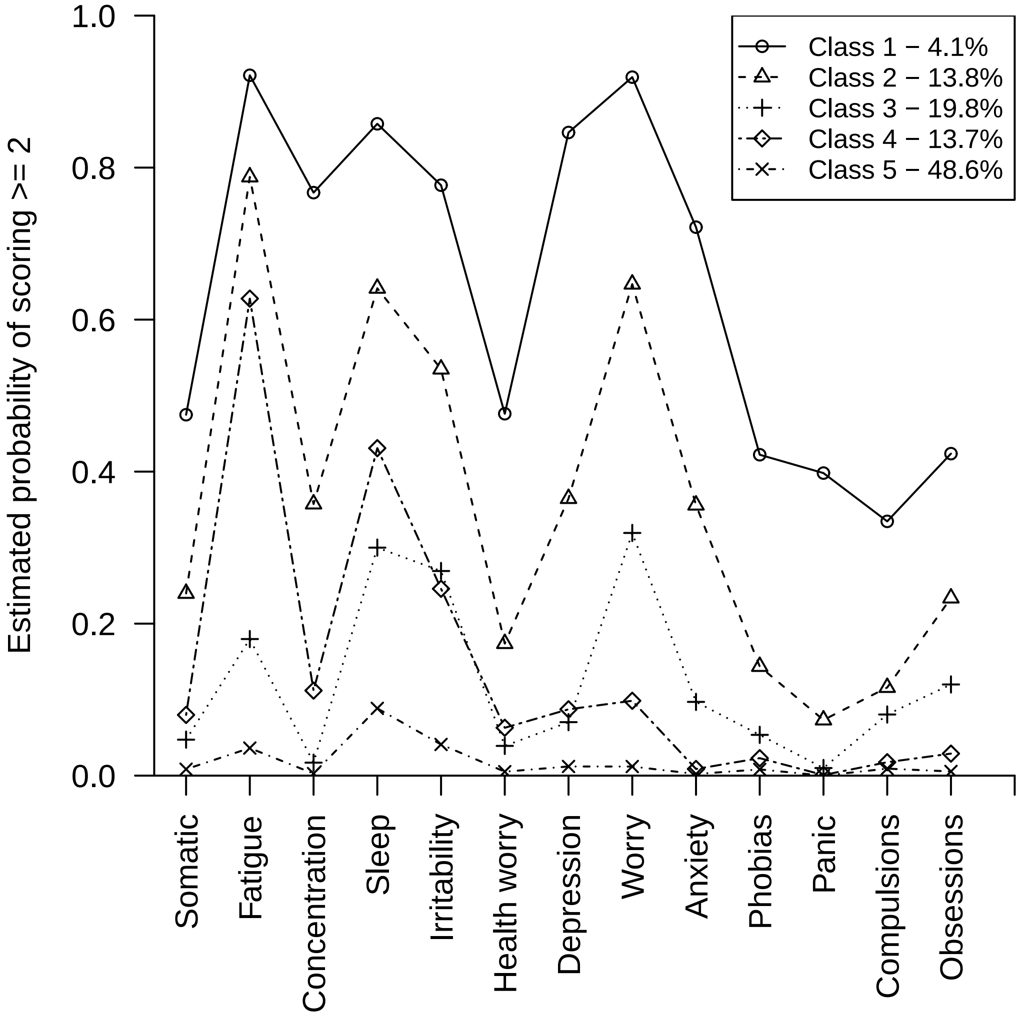

Figure 5.4: Estimated probabilities of scoring 2 or more on each symptom by class for the 5 class latent class model

class comparisons since they can all be seen in a single plot), it is important to note that each model actually estimates class specific probabilities for *all* of the 5 possible scores on each symptom. An example of a plot showing the full set of estimated probabilities is shown in Figure 5.6 — this is the same model as in Figure 5.4.

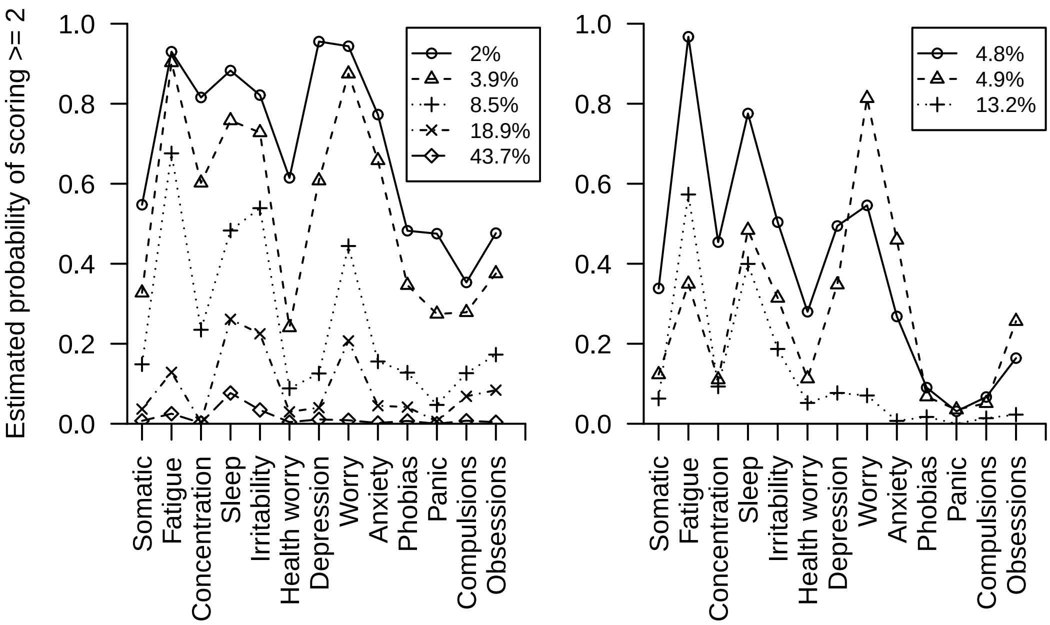

Figure 5.5: Estimated probabilities of scoring 2 or more on each symptom by class for the 8 class latent class model

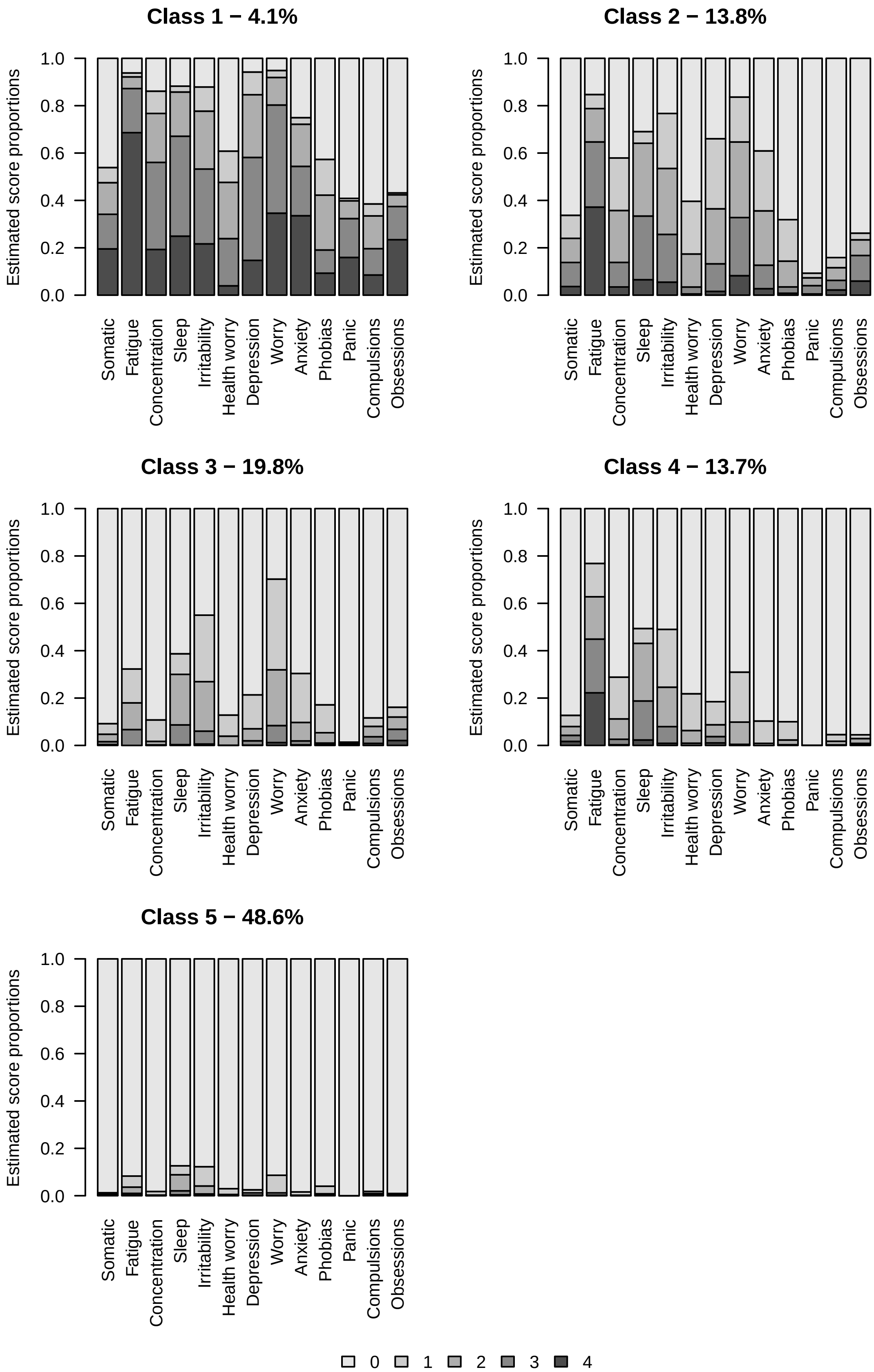

Figure 5.6: Estimated response probabilities from the latent class model with 5 classes (all 5 response categories shown)

5.2.3 Goodness-of-fit and residual analysis

As was the case for the factor models considered in the previous chapter, bivariate goodness-of-fit statistics and residuals can be used to investigate how well a model is able to reproduce the bivariate relationships observed in the data (as discussed on page 61). A selection of these statistics is shown in Table 5.2. This table shows that both the number of item pairs with significant lack of fit and the number of 'large' standardised residuals decrease as the number of latent classes increases. This is not particularly surprising, since models with more classes have more parameters with which to describe the data; the 8 class model contains twice the number of parameters as the 4 class model. The largest bivariate Pearson chi-square value in Table 5.2 (chi-square = 140) occurred for the relationship between worry and anxiety in the 4 class model, indicating that this relationship was recreated particularly poorly by the 4 class model. The 5 class model also struggled with this item pairing (chi-square = 93), although the 8 class model managed somewhat better (chi-square = 37). Some other item pairings that were particularly problematic were the relationships between phobias and panic, and between phobias and compulsions — for these item pairings, the fit remained poor even in the 8 class model (chi-square = 73 and 64 respectively).

Table 5.2: Bivariate goodness-of-fit statistics and residuals for latent class models

Model	Overall bivariate Pearson chi-square	Pairs with significant lack of fit (out of 78)	Largest 5 bivariate Pearson chi-square statistics					Bivariate standardised residuals > 3 (total = 1950)
4c	1,979	8	140	96	91	85	59	19
5c	1,667	4	96	93	80	47	41	11
8c	1,270	2	73	64	38	37	30	6

Abbreviations: c, class.
P value used for significance cut-off = 0.01; critical value for $\chi^2(24) = 43.0$.

Although the 4 class model appears to fit the data relatively poorly compared to the 8 class model, it is useful at this point to compare the fit statistics in Table 2 with some comparable results from the factor models in Table 4.9 on page 81. While the 4 class model has 8 item pairs with significant lack of fit and 19 bivariate residuals greater than 3, the unidimensional factor model had 35 item pairs with significant lack of fit (nearly half of the total number of item pairs), and 95 bivariate residuals greater than 3. Furthermore, given the large total numbers of item pair combinations and bivariate residuals for all models, some of these might be expected to exceed the chosen cut-off by chance even in a well-fitting model.

Therefore, although all three latent class models struggle with some item pairings (particularly those relating to phobias, panic and compulsions), the models in Table 2 do appear to provide a reasonable approximation to the data. The 8 class model does the best job of recreating the observed bivariate relationships, but at the cost of a considerable increase in the number of parameters. Relative to the 4 class model, the 5 class model achieves much of the improvement in fit that the 8 class model manages, while requiring far fewer additional parameters. For this reason, it may be appropriate to ignore the recommendation of the bootstrapped likelihood ratio test, and to select a model with no more than 5 classes to describe the CIS-R data.

5.2.4 Stability in other split half

In terms of the general patterns of symptoms across classes, all of the latent class models with 2 to 8 classes replicated in the second split half of the data. Looking more closely at the 5 class model, the pattern of conditional probabilities across the 5 classes was extremely similar, although there were very minor differences on a couple of symptoms (and also in the estimated class prevalences) for a couple of the classes. A set of bar plots showing the estimated response probabilities for the 5 class model from the second split half of the data is available for comparison in Figure E.1 on page 290 of Appendix E; this can be compared with Figure 5.6 on page 95.

The pattern of classes in the 8 class model was also replicated in the second split half of the data. As in the first half of the data, there were 5 clearly ordered classes, as well as 2 classes with high levels of fatigue and sleep problems with relatively low levels of other symptoms, and 1 class with relatively low levels of fatigue and sleep problems but high levels of worry. Some class prevalence estimates differed by 1 or 2%, and there were a few small differences on a couple of symptoms, but in general, the two 8 class solutions from the different split halves of the data were remarkably similar. The profile plot from the 8 class model estimated in the second split half of the data is available for comparison in Figure E.2 on page 291 of Appendix E; this can be compared with Figure 5.5 on page 94.

Interestingly, the 9 class model (which was estimated successfully in the first split half of the data, although not supported as making a substantial improvement to model fit by any of the fit statistics in Table 5.1) did not replicate in the second split half of the data: the 9 class model was not identified. This suggests that the 9 class model in Table 5.1 may have been overfitting the data, and vindicates the rejection of this model by the bootstrapped likelihood ratio test.

5.3 The meaning of the latent classes

As discussed earlier, the clearly ordered class profiles in Figures 5.3 and 5.4 suggest that at least some of the latent classes may be performing an 'indirect' function: rather than representing distinct subgroups of individuals, the classes may instead be approximating the distribution of one or more underlying latent dimensions. The 4 class model appears to be describing a single latent dimension, while the 5 class model may be indicating the presence of two latent dimensions (although it is also possible that the distinctive 'high fatigue' class may be playing a direct role and correspond to a real group in the population). In order to see how closely the solutions of the 4 and 5 class models correspond to the dimensional results from the factor models of the previous chapter, Figures 5.7, 5.8 and 5.9 plot the factor scores from the one and two dimensional factor models indicating the most likely class assignments for each individual.

Figures 5.7 and 5.8 suggest that the ordered classes from the 4 class model are providing a discrete representation of the same latent dimension measured by the unidimensional factor model. Figure 5.9 suggests that the 5 class model provides a similar discrete representation of this main dimension. However, two of the classes appear to be contrasting individuals with low to moderate scores according to their scores on the highly correlated 'fatigue' and 'worry' dimensions identified in the two factor E/CFA: one class contains those with moderate fatigue but low worry, while the other contains those with moderate worry but low fatigue. The 6 and 8 class models (not shown) add similar distinctions among individuals with higher scores on each factor. These observations seem to imply that the latent classes in all of these models are describing one or two continuous dimensions with a series of discrete points.

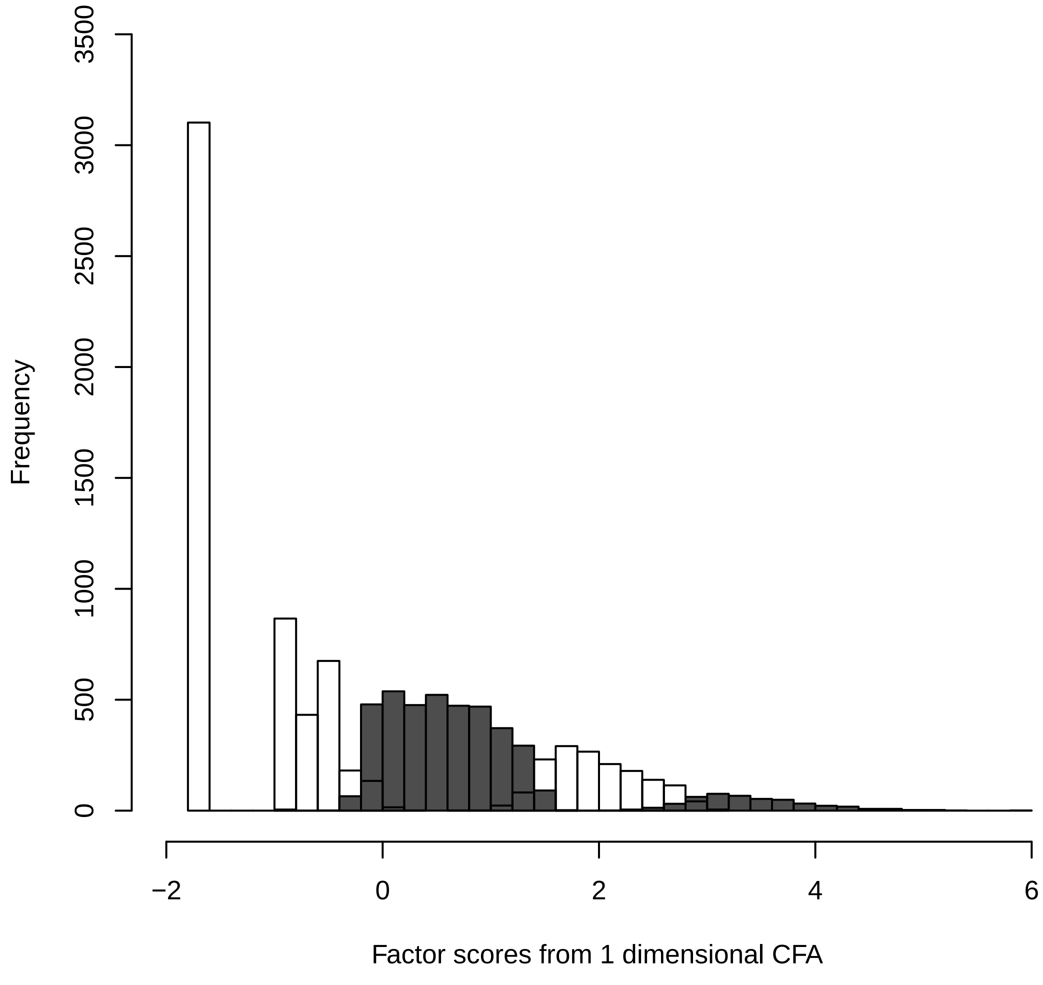

Figure 5.7: Comparison of most likely class membership allocations from the 4 class latent class model with estimated factor scores from the 1 dimensional confirmatory factor analysis. Individuals allocated to adjacent classes are shown in contrasting colours.

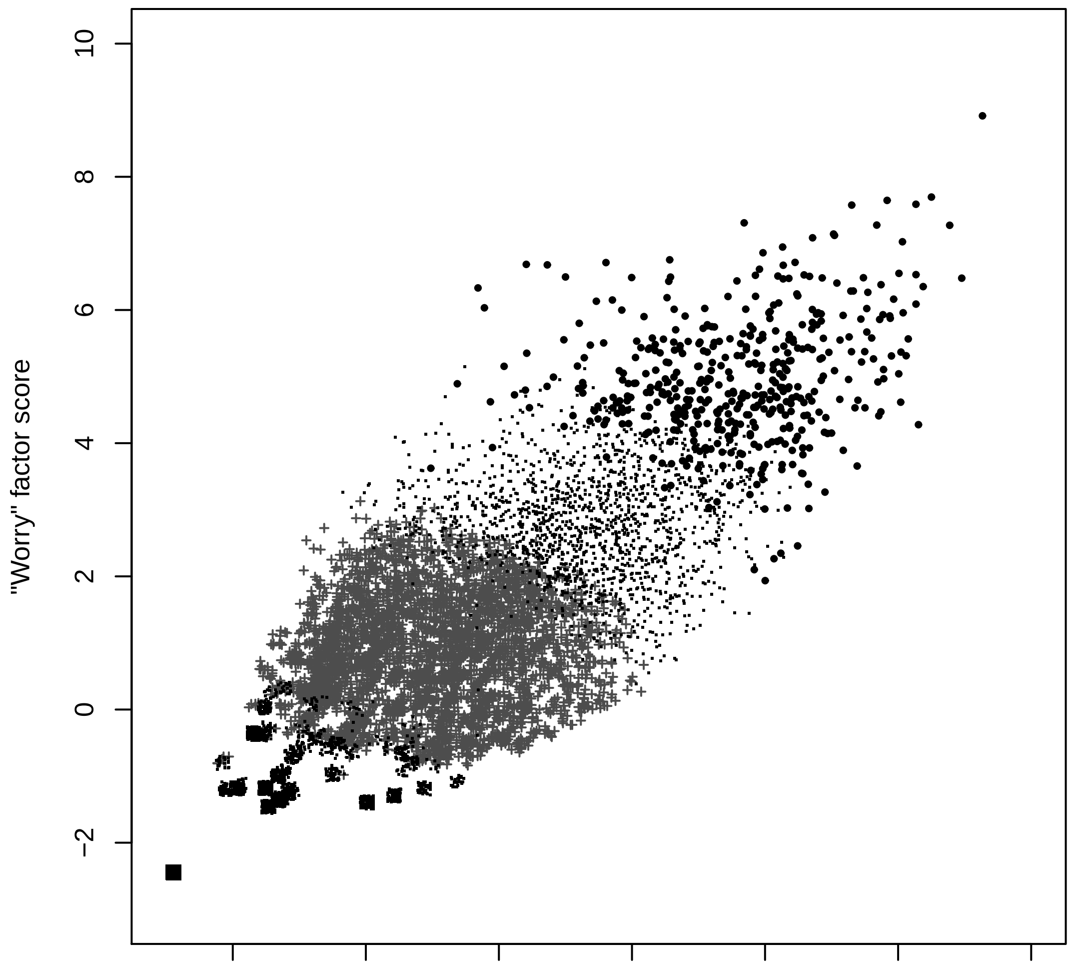

Figure 5.8: Comparison of most likely class membership allocations from the 4 class latent class model with estimated factor scores from the 2 dimensional E/CFA. Factor scores have been jittered by a small random amount to indicate where many individuals share the same factor score.

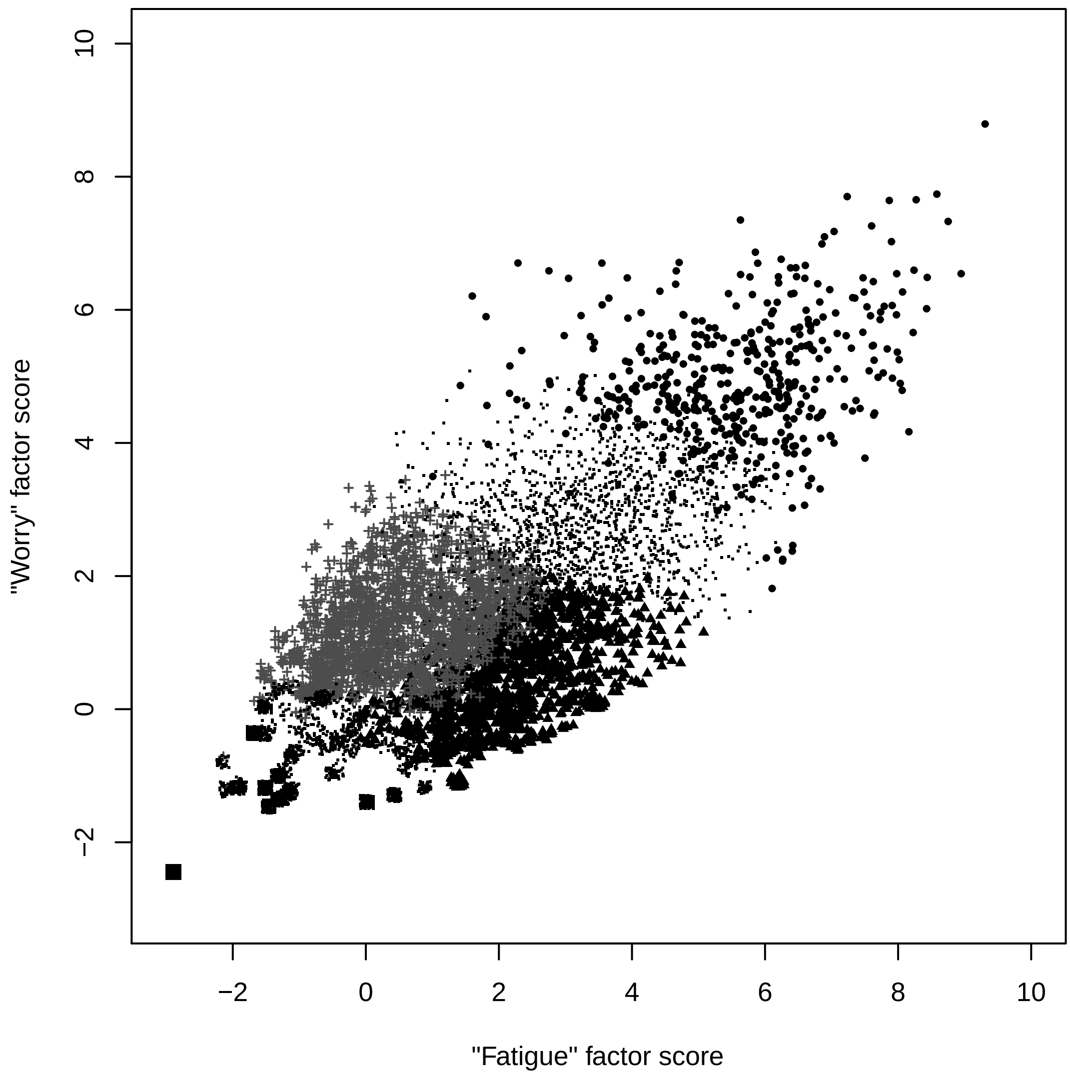

Figure 5.9: Comparison of most likely class membership allocations from the 5 class latent class model with estimated factor scores from the 2 dimensional E/CFA. Factor scores have been jittered by a small random amount to indicate where many individuals share the same factor score

5.4 Conclusion

The choice between latent class models appears to come down to a similar set of alternatives to those discussed at the end of the previous chapter relating to factor models. The classes in all of these latent class models appear to be performing an indirect function, providing a discrete representation of underlying latent dimensions (which may be continuous or discrete). The CIS-R data again appear to be largely unidimensional; however, when a 4 class 'unidimensional' latent class model is fitted, there are some large residuals and poor bivariate goodness-of-fit statistics, suggesting the presence of residual correlations that are not accounted for by the model. The 5 class model addresses some of these residual correlations by extending the unidimensional approximation to accommodate some aspects of a two-dimensional model, while the 8 class model appears to take this further to provide a more complete and well-fitting approximation of the two highly correlated latent dimensions.

Therefore, while the 4 class model summarises the unidimensional core of the latent structure in the data, models with more classes provide a better fit to the data by accounting for the stronger patterns of correlations between some groups of symptoms. Whether these improvements in fit are considered useful or important will be a matter of choice for individual readers. The author favours the simpler 4 class model with its 'minimal assumption' summary of the *key feature* of the latent structure (as identified by the factor and latent models): a single underlying latent dimension.

Finally, it should be remembered that none of the latent class models in this chapter could satisfactorily account for the relationships between phobias, panic and compulsions (which were addressed, at least to some extent, by the third dimension in the three factor E/CFA, even though the residuals of this model suggested that its overall fit was poor). It will be interesting, therefore, to see if any of the more complex hybrid models to be discussed in the next chapter can provide a satisfactory description of the relationship between these variables, as well as whether the hybrid models can improve on the fit of the factor models in a more parsimonious way than the latent class models described in this chapter. There is also the possibility that hybrid models may be able to detect latent subgroups in the data that have been overlooked by the factor and latent class models applied so far, given the need for any model to account for the powerful dimensionality of the data.

Chapter 6

Latent variable modelling III: Factor mixture analysis — unidimensional models

6.1 An overview of factor mixture modelling

The previous two chapters have discussed two families of latent variable models that appear to make very different assumptions about the underlying latent structure. In the factor analysis family of models, the relationships between variables in the model are assumed to be explained by the presence of one or more continuous latent variables. Conversely, in the latent class analysis family of models, the relationships between variables in the model are assumed to be explained by the presence of a categorical latent variable.

Typically, a researcher's choice about which of these two approaches to use will be guided by their prior beliefs about whether the latent construct in question is continuous or categorical in nature. Therefore, researchers who have been looking to identify homogeneous groups of individuals corresponding to different types of mental disorders have naturally employed latent class analysis. However, the conditional independence assumption of latent class analysis means that there can be no severity differences within a class (as discussed in Section 5.1.1). This may be undesirable, since it is reasonable to expect that many mental disorders will vary in severity; the result may be the extraction of 'spurious' latent classes to account for variations in severity (Lubke & Neale, 2006). Factor mixture analysis, which allows for severity differences within a latent class, is increasingly being seen as an attractive alternative for modelling such data (Muthen, 2006).

Factor mixture models are hybrid latent variable models that combine both continuous and categorical latent variables into a single model. While keeping the general goal in latent class analysis of dividing individuals into groups sharing common characteristics, they relax the assumption of conditional independence within each class (which specifies that all items must be uncorrelated within each class) — this is achieved by allowing individuals in the same class to differ along one or more continuous latent dimensions. This could allow individuals to vary in the severity of their symptoms within a single 'disorder' class. For example, one latent class may describe individuals with depressive symptoms, but their depressive symptoms may range from mild to severe along a latent severity dimension. As has been noted by Wittchen *et al.* (2009) and Leoutsakos *et al.* (2010), this makes factor mixture models consistent with the proposed DSM-5 conception of categorical mental disorders that can vary along a dimension of severity (American Psychiatric Association, 2010a).

Even though such an introduction implies that factor mixture models are primarily used with the expectation that they are performing a direct function of identifying latent subgroups in a population (see section 5.1.2 on page 84 for the difference between direct and indirect roles), there are in fact many different specifications for possible models that combine continuous factors and latent classes, and some of these may be chosen precisely because of the indirect functions that these models can perform. One of the key differences between different types of factor mixture models is in the extent to which measurement invariance is imposed between classes on the parameters of the within-class factor model. The next sections will give a brief overview of the importance of measurement invariance within both multi-group factor analysis and factor mixture models, along with a description of how the imposition of different degrees of measurement invariance can allow the specification of a flexible set of models which have some very different implications for the latent structure in the data.

6.2 Measurement invariance in multi-group factor analysis models

While the goal of this section is to clarify the concept of measurement invariance in relation to the role it plays in factor mixture models, the principles of measurement invariance are the same in factor mixture models as for factor analysis with two or more groups where the grouping variable is already known (such as gender) — this is known as multi-group factor analysis. In multi-group factor analysis separate factor models are specified simultaneously for each group — the model parameters for each group can then either be constrained to be equal across the two groups, or they can be freed to vary and hence estimated separately in each group. Multi-group factor analysis may be carried out either with the primary goal of exploring and testing for potential differences in the factor structure between the groups, or as an important prelude to building a more complex model that requires assumptions of measurement invariance between the groups (for example, a model for longitudinal data where the different groups represent repeated measurements of the factor indicators on two or more occasions, and where the goal is to investigate changes in factor scores over time).

In order to see whether different parameter estimates for certain types of parameters are needed in each group, or whether the model fits well even when these parameters are constrained to be equal across groups, a likelihood ratio test (for twice the difference in model log-likelihoods) may be used to compare the fit of the models with and without parameter equality constraints on the chosen parameters. Equality constraints are generally imposed or relaxed for particular types of model parameters in sequence (Brown, 2006) — the degree of measurement invariance that is said to be present relates to which types of parameters may be constrained to be equal across the groups without significantly impairing the fit of the model.

Given that the principles of measurement invariance are the same in the multi-group setting where group membership is known as in the factor mixture setting, an overview of the different levels of measurement invariance will first be given in the context of multi-group factor analysis. In later sections, this will be extended to the more complex situation where class membership is not known and class prevalences must be estimated alongside the parameters of the within-class factor models. The discussion in this chapter will mainly focus on factor models with only one factor. However, all of what follows applies equally to multi-dimensional factor analysis models, although these models must additionally take into account the appropriateness of equality constraints on covariances/correlations between factors.

It can be difficult to visualise what is happening in multi-group factor analysis when equality constraints are relaxed for particular types of parameters. The next subsection will therefore start by reviewing the parameters of the standard factor analysis model and how they can be visualised. One of the key differences between models with different degrees of measurement invariance lies in the number of parameter constraints that are required in order to ensure that the factor model is identified, and the implications this has for how the results can be interpreted. Therefore, the next subsection will also review the subject of model identification in the context of single group factor models. Subsequent subsections will then move on to a discussion of multi-group models with differing degrees of measurement invariance, in each case considering the identification constraints that are required at this level and the consequences these have for model interpretation.

6.2.1 Review of the single group factor model and model identification

In order to interpret and visualise the differences between multi-group factor models with differing degrees of measurement invariance, it will be helpful to review the standard factor model for ordinal categorical data and how this is identified. In this model, the probability of an individual i responding to item j in category k or higher is given by

$$P(Y_{ij} \geq k) = \frac{e^{\alpha_j \theta_i + \gamma_{jk}}}{1 + e^{\alpha_j \theta_i + \gamma_{jk}}}, \tag{6.1}$$

or equivalently

$$logit(P(Y_{ij} \geq k)) = \alpha_j \theta_i + \gamma_{jk}. \tag{6.2}$$

Equations 6.1 and 6.2 describe how this probability is determined by the individual's latent trait/factor score θ_i, the factor loading (or slope) for the item α_j, and the intercept for the category γ_{jk}. This parameterisation is sometimes known as the slope/intercept parameterisation; since it takes the form of an ordinal logistic regression model, this parameterisation emphasises the role of logistic regression in the factor model for categorical data. (The probit link function may also be used instead of the logit link). Equation 6.1 can be rewritten in what is known as the item response theory (IRT) parameterisation as

$$P(Y_{ij} \geq k) = \frac{e^{\alpha_j (\theta_i - \delta_{jk})}}{1 + e^{\alpha_j (\theta_i - \delta_{jk})}}. \tag{6.3}$$

This parameterisation gives the item response theory model known as the Graded Response Model (Samejima, 1969) — the intercept parameter γ_{jk} has now been replaced by a category difficulty parameter δ_{jk}. The two models are equivalent, and both will fit the data equally well. Furthermore, if the slope/intercept parameterisation is used, the item response theory difficulty parameters δ_{jk} from the second parameterisation can be recovered from the intercepts γ_{jk} by the transformation

$$\delta_{jk} = -\frac{\gamma_{jk}}{\alpha_j}. \tag{6.4}$$

The reverse transformation can also be applied. In other chapters of this project the slope/intercept parameterisation is preferred when describing models and their parameters, primarily since this parameterisation extends more naturally to multidimensional factor models, and since it may be understood more intuitively by researchers who are familiar with the generalised linear modelling framework but not item response theory. However, the use of difficulty parameters may make it easier to visualise the requirements for model identification to be discussed later for different configurations of measurement invariance. Therefore, in this chapter intercepts and difficulty parameters will be used almost interchangeably, according to whichever makes the interpretation more intuitive.

In order to be able to switch comfortably between these two parameter types, it will be helpful to focus on how these parameters can be interpreted and to see how they relate to each other graphically. The single group factor model described above in Equations 6.1 and 6.3 is pictured in Figures 6.1 and 6.2 for an item with 5 ordered categories (k) corresponding to symptom scores of 0, 1, 2, 3 and 4; such a model requires the estimation of four ($k - 1$) intercept or difficulty parameters corresponding to the probability of obtaining a score of 1 or more, 2 or more, 3 or more, and 4. Figure 6.1 shows how the probability of gaining a particular score (or higher) changes in relation to an individual's latent trait score. Figure 6.2 shows the same model but this time on the original logit scale used for parameter estimation.

The factor loadings/slopes/discrimination parameters α_{jk} can be visualised in both figures. In Figure 6.1 the factor loading represents the slope of the curves at the point where the probability of obtaining the symptom score k or higher is 0.5 (indicated by the horizontal dashed line). Note that the standard factor model assumes that the factor loading is the same for all categories of an item — therefore, the four curves are parallel. In Figure 6.2, the factor loading α_{jk} is simply the slope or gradient of the lines. The higher the value of the factor loading, the steeper the line, and hence the better the item is at discriminating between those with low latent trait scores θ_i and those with high scores.

The intercept parameters γ_{jk} relate to the vertical axis measured on the logit scale, therefore they do not appear on Figure 6.1. (If the inverse logit transformation was applied to the intercepts, each would give the probability of obtaining score k or higher for an individual with a latent trait score (θ_i) of 0, as would be indicated by the intersections of the curves with the vertical dashed line in Figure 6.1.) However, the intercepts do appear in Figure 6.2 — here, they reflect the values on the vertical axis at which the plotted lines cross 0 on the horizontal axis. The values of the intercept parameters are indicated by the 4 symbols marked on the vertical axis.

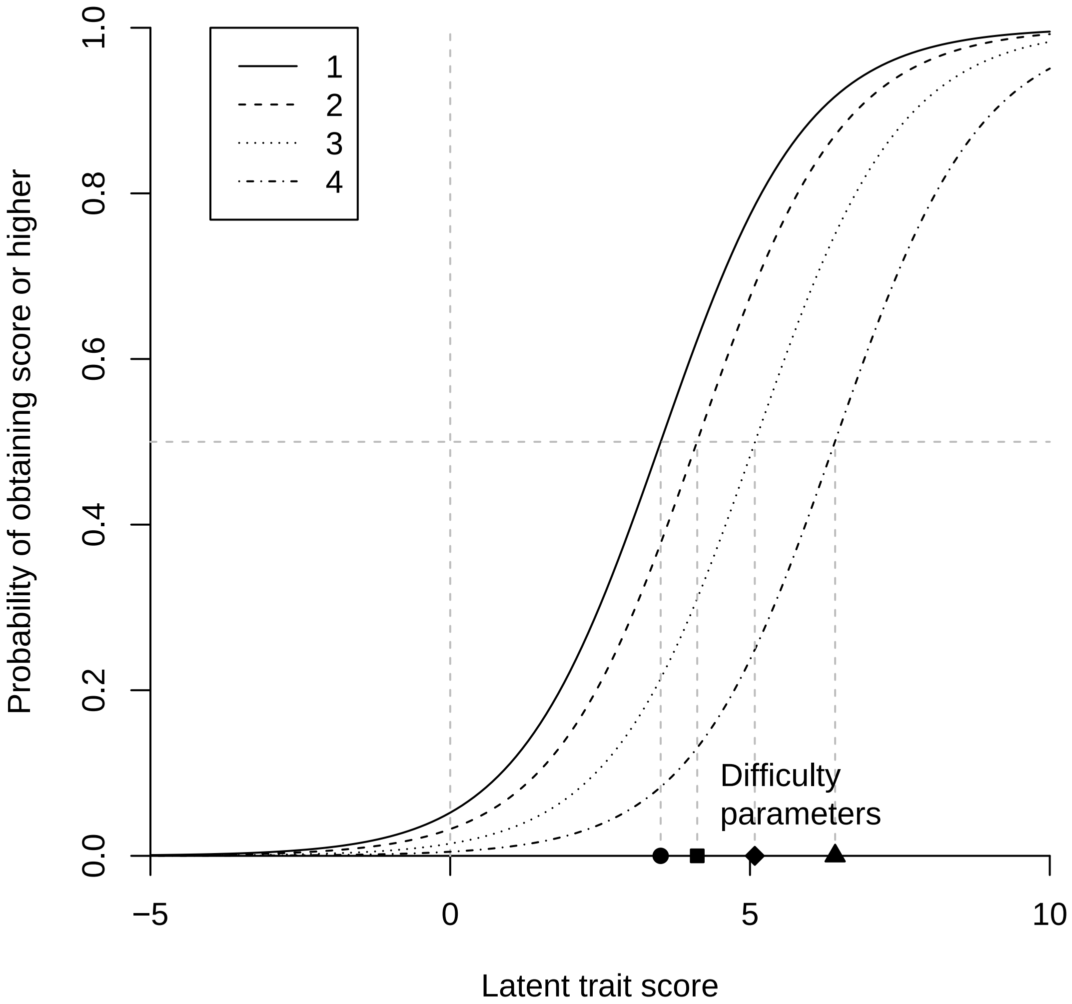

Figure 6.1: Response functions for an item with 5 ordered categories (scores of 0,1,2,3 and 4) on the probability scale. The values of the difficulty parameters for scores 1,2,3 and 4 are indicated by the symbols on the horizontal axis.

The higher the intercept for a category, the greater the expected prevalence of that score or higher at all levels of the latent trait. Since the factor model for ordinal data models the *cumulative* probabilities of scoring in category k *or higher*, the $k-1$ intercept parameters will always be strictly ordered.

In contrast to the intercept parameters measured on the vertical axis, the difficulty parameters δ_{jk} are actually measured on the same scale as the latent trait scores (θ_i) and are plotted along the horizontal axes of Figure 6.1 and 6.2. It is the fact that these parameters can be plotted on the latent trait scale that will make them particularly useful during the discussion of model identification. The difficulty parameters can be interpreted as the score on the latent trait at which an individual

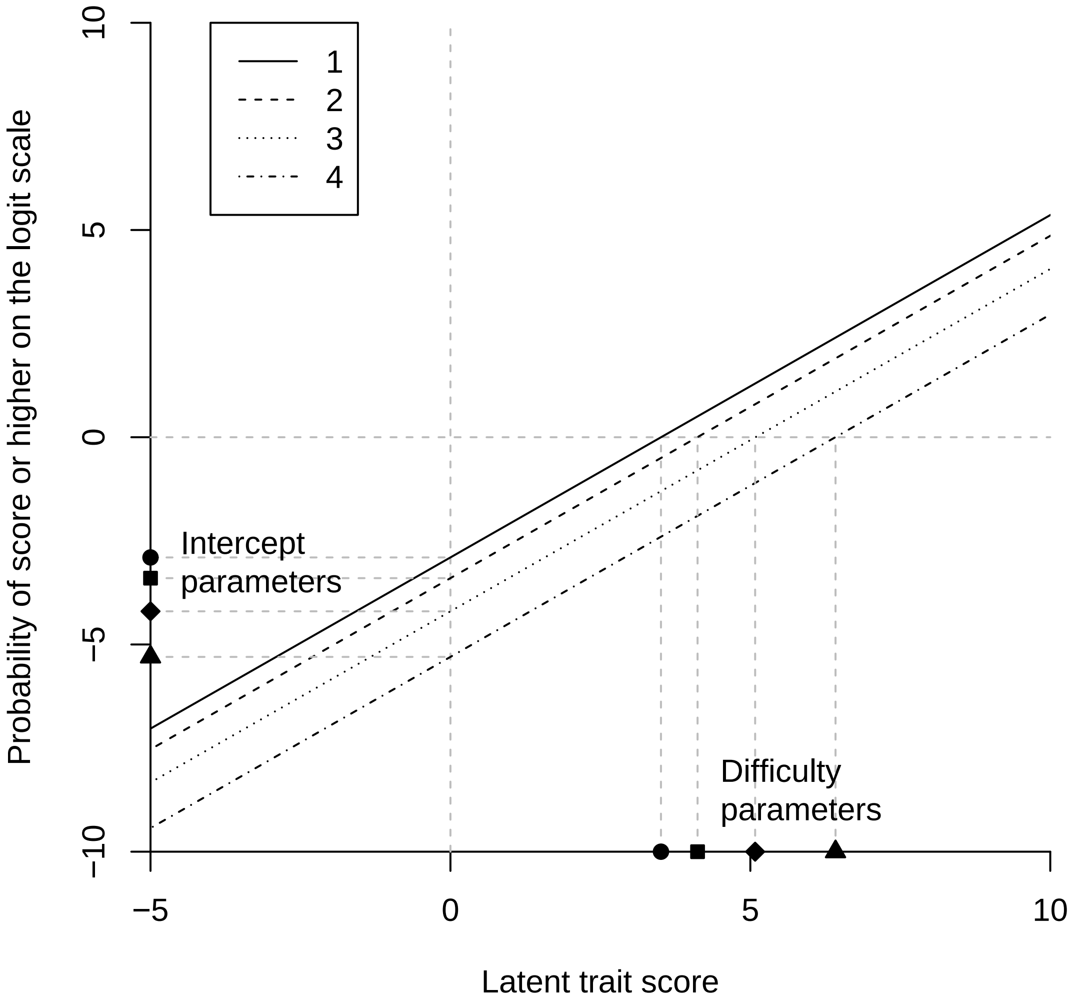

Figure 6.2: Response functions for an item with 5 ordered categories (scores of 0,1,2,3 and 4) on the original logit scale. The values of the difficulty parameters for scores 1,2,3 and 4 are indicated by the symbols on the horizontal axis, while the intercept parameters for those scores are indicated by the corresponding symbols on the vertical axis.

has a 50% chance of obtaining a score of k or higher for that item. The name of the 'difficulty' parameter relates to the origins of item response theory in the field of academic ability testing — the higher the 'difficulty' of an item, the further along the latent ability trait a student would need to be to have a good chance of answering the item correctly.

A low difficulty parameter for a category indicates that even individuals with relatively low latent trait scores are likely to obtain symptom scores of k or higher on this item, whereas a high difficulty parameter reflects the fact that only individuals with high latent trait scores are likely to obtain this symptom score. The higher the difficulty parameter for a particular category, the lower the expected prevalence of that symptom score at all values of the latent trait. The values of the difficulty parameters are indicated in Figure 6.2 by the 4 symbols marked on the horizontal axes of both figures — this figure illustrates how the intercepts and difficulty parameters are inversely related (as can also be seen from Equation 6.4). However, Figure 6.2 also highlights the fact that both types of parameter are in fact measuring the same thing, albeit in different ways: the vertical positions of the four regression lines.

A key issue with factor and item response theory models is that the models as described above have no unique solution; different sets of values for the parameters can yield exactly the same item response probabilities. For example, looking at the item response theory parameterisation in Equation 6.3, the values $\alpha_j = 1$, $\theta_i = 4$ and $\delta_{jk} = 2$ yield a probability of responding in that particular category or higher of 0.88. However, the values $\alpha_j = 2$, $\theta_i = 2$ and $\delta_{jk} = 1$ yield exactly the same result. In general, any valid combination of parameter values can be transformed into an equally valid combination simply by *multiplying* the α_js by a factor of z and *dividing* the θ_is and δ_{jk}s by the same factor z. Since all of these combinations are equally likely to have generated the observed data, attempting to estimate the model as it stands will result in infinite standard errors for the parameters and maximum likelihood estimation will fail; the model is said to be "not identified". This occurs because the latent trait has no inherent metric. In order to estimate the model successfully and ensure that there is a unique solution, arbitrary constraints must be imposed to fix two aspects of the latent trait: its scale, and its location.

Figure 6.3 illustrates the importance of fixing the scale of the latent trait. The scale may be conceptualised as the distance between the units or 'tick marks' on the horizontal axes of the two graphs in this figure. The panels show just two out of the many possible scales for the latent trait — the units or 'tick marks' are closer together in the left hand panel than in the right hand panel. Given that the factor loading/slope is simply the gradient of the lines (i.e., the vertical rise in units of the y axis for each 1 unit horizontal change along the x axis), Figure 6.3 illustrates how the value of the slope parameter must change to reflect a change in the scale of the latent trait. This change in the factor loading occurs even though the relationship modelled in the graph stays exactly the same.

In order to fix or 'identify' the scale of the latent trait (and therefore to ensure that there is a unique solution for the factor loadings), an arbitrary scale must be selected. Common choices are to fix the slope of the first item in the model to be 1 or to fix the variance of the factor to be equal to 1. The first of these constraints tends to be more commonly used in the factor analysis and structural equation modelling frameworks, whereas the latter is preferred in item response theory settings. However, in most situations it makes no difference which constraint is used — both are equally valid. Factor loadings estimated under one scale constraint may be transformed to correspond to another choice of constraint simply through multiplication by the appropriate constant (Kamata & Bauer, 2008).

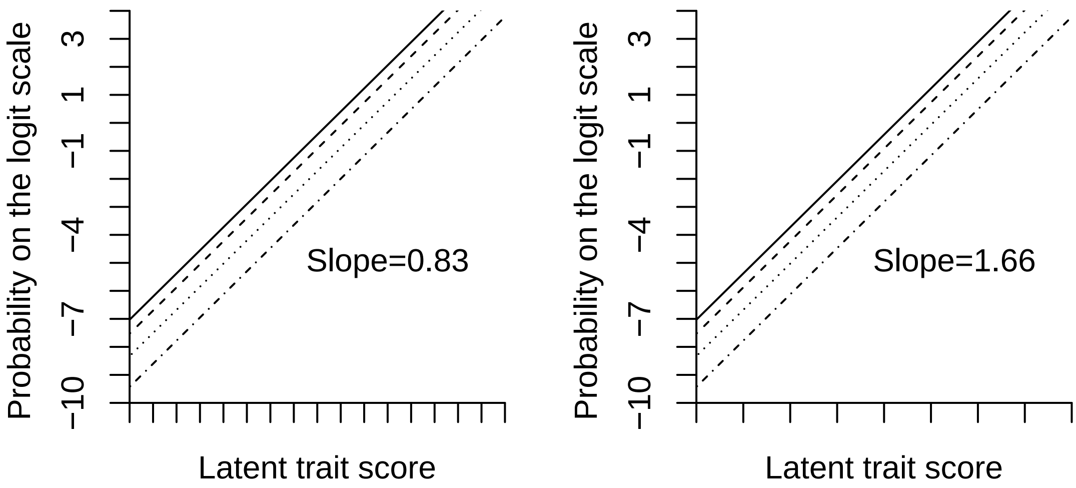

Figure 6.3: Illustrating how the value of the slope parameter depends on the choice of scale constraint for the latent trait (which affects the distance between the unit markings on the horizontal axis). The panels represent two different but equally valid choices of scale constraint applied to the same data. The units are more closely spaced in the left-hand panel than in the right hand panel; hence, the slope parameter is lower for the left-hand panel than for the right-hand panel.

In Figure 6.3 the focus was on factor loadings and the *scale* of the latent trait. In this setting it was not necessary to have any values marked along the latent trait axis, since what was important was the distance between the units, not the actual values corresponding to those units. However, in order to estimate values for the intercept or difficulty parameters, the numerical values corresponding to each of the units or 'tick marks' must be known — these values refer to the *location* of the latent trait. Figure 6.4 shows two out of the many possible choices that can be made for the location of the latent trait: the zero value for the latent trait score appears in different positions along the horizontal axes for the two graphs.

Since the values of the difficulty parameters δ_{jk} are read straight off the horizontal latent trait axis, it is immediately clear that the values of these parameters will depend on the choice of location for the latent trait — the difficulty parameters are higher for the choice of location in the left hand panel than for the choice in the right hand panel, even though the relationship being modelled is exactly the same. In addition, since the intercepts γ_{jk} of the lines are defined at the latent trait score of 0, these parameters will also change according to where on the horizontal axis the value of 0 is placed. Therefore, in order to ensure that there is a single unique solution for the intercepts and/or difficulty parameters, it is necessary to fix the *location* of the latent trait, in addition to its scale. The most common form of constraint is to fix the mean of the factor/latent trait scores in the sample to be 0. However, it is also possible to fix the value of a single difficulty parameter or the mean of the difficulty parameters. As with the choice of scale constraint, whichever location constraint is selected, the parameter estimates may be transformed to correspond to an alternative choice of constraint by the application of a straightforward transformation (Kamata & Bauer, 2008).

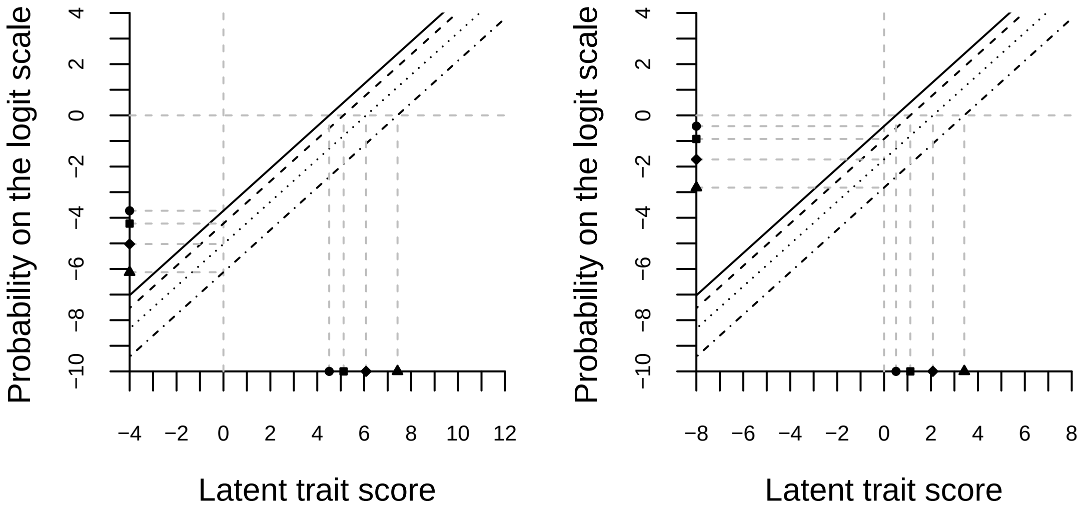

Figure 6.4: Illustrating how the values of the difficulty parameters (symbols on the horizontal axis) and intercept parameters (symbols on the vertical axis) depend on the choice of location constraint for the latent trait — the choice of location constraint affects the position of the 0 score on the horizontal latent trait axis. The panels represent two different but equally valid choices of location constraint applied to the same data — the same scale constraint has been used in each panel.

One of the difficulties associated with the arbitrary nature of scale and location constraints is that it is only meaningful to compare parameters (such as slope parameters or factor scores) that are measured on the same metric. If a single factor model is applied to all individuals in a population, this is not a problem — the slope parameters for different items can be easily compared, and we can make statements about whether one individual has a higher latent trait score than another individual. However, it becomes more difficult when the model is based on only a subgroup of the population. If the location and scale constraints of the latent trait are fixed based on the sample characteristics of a particular group (such as its factor mean and variance), the values of all parameters will depend on the nature of the group that was selected.

For example, if a group is relatively homogeneous (such as a clinical sample where everyone tends to have high scores on the latent trait), fixing the scale constraint so that the factor score variance is 1 will mean that the 'tick marks' from Figure 6.3 for this group will be much closer together than they would be if the same model was fitted to the entire clinical and non-clinical population — this means that the factor loadings will be lower for the clinical sample (and so the relationships may appear to be weaker), even though the actual relationships between the symptoms and the latent trait are the same for both situations. This highlights the fact that parameter values for factor models have no inherent meaning when they are separated from the context of the sample in which they are measured. Furthermore, this implies that it will only be meaningful to say that the factor score of an individual from one group is higher than the factor score of an individual from another group if we can be sure that the individuals are measured on exactly the same scale — whether this is possible for factor models with two or more groups is an important consideration.

The discussion above has concentrated on model identification for standard single group factor models. However, in multi-group settings, the situation becomes a little more complex. The absolute minimum requirement for a unidimensional multi-group factor analysis model to be identified is that the location (factor mean) and scale (a factor loading or the factor variance) is fixed in one of the groups, as in the standard single group model. For some model configurations, these constraints are sufficient to identify the metric of the latent trait in all groups. Nonetheless, models with some levels of measurement invariance will require identification constraints to be specified for *each* group. The various commonly used configurations of measurement invariance will be described in the following subsections, along with the implications of each configuration for comparing parameter values or factor scores between groups.

6.2.2 Strong factorial invariance

In general, for a factor analysis model with ordinal categorical data there are four different types of parameter that may be constrained or estimated within each group:

1. the item intercepts or corresponding difficulty parameters (there are $k-1$ intercepts associated with an ordinal item with k categories);

2. the factor loadings (one for each item);

3. the factor mean;

4. the factor variance (factor variances *and* covariances in a multidimensional factor model).

The simplest multi-group situation occurs when factor loadings and intercepts are constrained to be equal in each group. The terminology for describing different levels of measurement invariance varies in the literature (see Brown, 2006), and is further complicated by the fact that the situation is slightly different for factor analysis with continuous variables. However, the specification of equal loadings and equal intercepts can be described as ‘strong factorial invariance’. Since the intercepts and loadings are the same in all groups, the factor is measuring the same latent concept in each group. Furthermore, the fact that these parameters are known to be the same in each group means that the latent trait metric will be the same in each group — therefore, it makes sense to compare factor scores for individuals from different groups, or to compare the mean factor scores of the groups themselves.

As illustrated in Figure 6.5 on the following page, the scale and location of the latent trait only need to be fixed in one group (which may be termed the reference group), for example by fixing the mean of the latent trait in that group to 0 and the loading of the first item to 1. The identical loadings in each group will ensure that the scale is the same for all groups. In terms of the location of the latent trait, since this is fixed in the reference group, the mean factor score in this group is known to be 0; the mean factor scores for the other groups can now be freely estimated, but this requires the application of the location constraint from the reference group to the latent trait in the other groups.

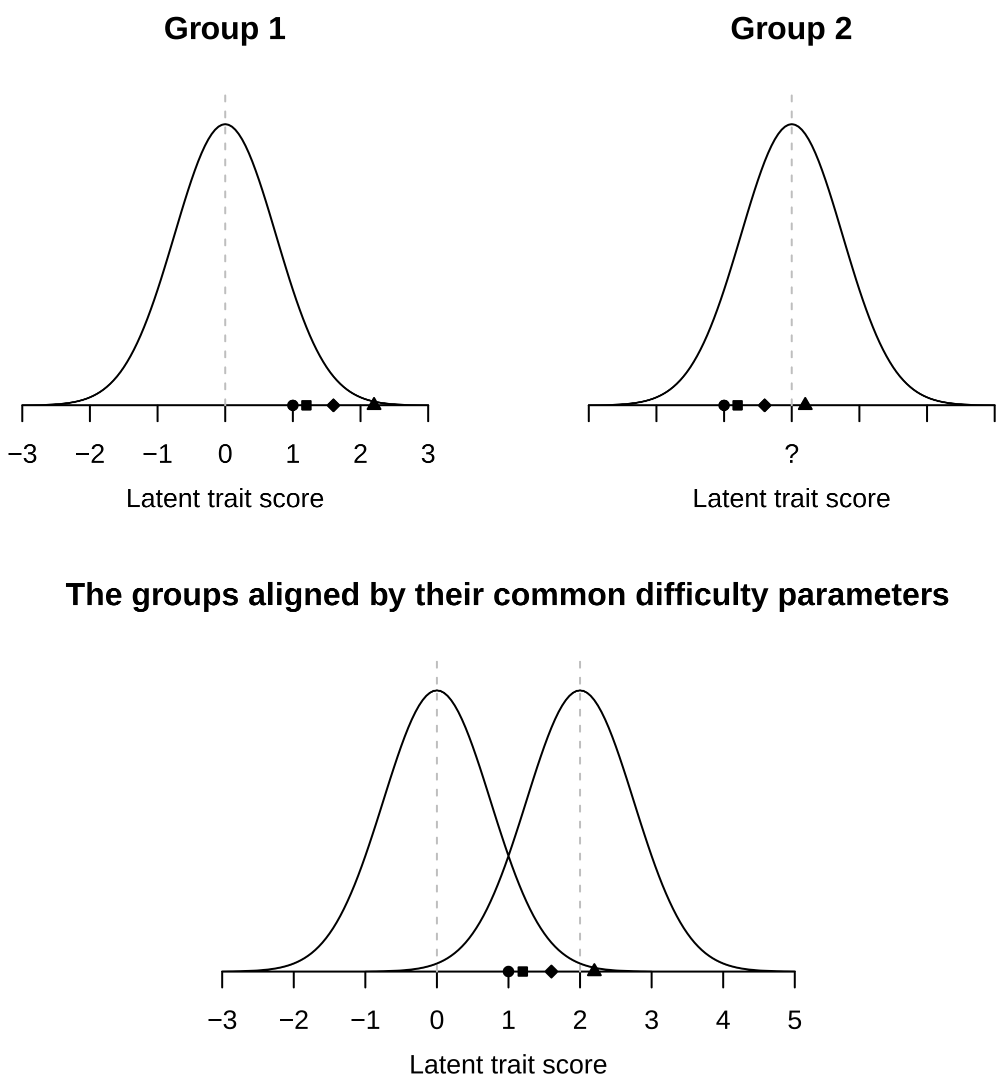

Figure 6.5: Factor score distributions for two hypothetical groups showing strong factorial invariance, i.e. the factor loadings and intercepts/difficulty parameters are the same in each group. The scale of the latent trait in both groups has been fixed by constraining one of the factor loadings to equal 1, while its location has been fixed in Group 1 by fixing the factor mean in this group to 0, but the location has not been fixed in Group 2. Nonetheless, since both groups share the same values for the difficulty parameters (the symbols on the horizontal axes), these parameters can be considered as anchor points that allow the metric from Group 1 to be applied to the latent trait in Group 2. This is illustrated in the lower panel. The estimated factor mean for Group 2 can be interpreted as the difference between the mean factor scores of Group 1 and Group 2.

Since all groups are known to have the same values for their intercept parameters (and hence the same values for their difficulty parameters on the latent trait scale) the shared values of these difficulty parameters can be considered as 'anchor points' that allow the metric from the reference group to be applied to the other groups. For example, if a difficulty parameter for the highest category is known to take the value of 2.2 in the reference group, this difficulty parameter is constrained by the equality constraint to take the same value of 2.2 in all of the other groups — the values at all other positions along the latent trait can then be deduced on the basis of their positions relative to this point (as illustrated in Figure 6.5). The estimated factor means in groups other than the reference group may be interpreted as the difference in mean factor scores between that group and the reference group (in which the factor mean was fixed to 0). This means that the groups may differ substantially in their overall levels of the latent trait, even though the intercepts and loadings are constrained to be the same in both groups.

It is possible to constrain the factor variances to be the same in each group. However, this may be inappropriate and impair the fit of the model, particularly if there are large differences in the overall level of the latent trait between groups. For example, a group with clinically diagnosed disorders may be much more homogeneous in terms of the range of its latent trait scores than the remainder of the population without a clinical diagnosis (which could encompass an enormous range of symptom levels from perfect health to distressing but sub-threshold conditions).

6.2.3 Weak factorial invariance

A more complex situation occurs where the factor loadings are constrained to be equal across groups but the intercepts are allowed to vary — this has been described as 'weak factorial invariance'. The presence of differing intercepts does not merely indicate differences in overall latent trait severity between the groups (which could be accommodated easily by different factor means in the presence of strong factorial invariance as described in the previous subsection). The fact that intercepts are required to differ between groups implies that the *relative* probabilities of the items or categories vary in some way. For example, symptoms of anxiety may be much more prevalent than symptoms of depression in one group, while the opposite pattern may hold in another group. Such a situation is illustrated in Figure 6.6 — in this hypothetical example with binary symptom measures, the relative prevalences of symptoms differ between the groups. In group 1 (in the lower panel), symptoms such as fatigue and sleep problems are relatively rare, whereas symptoms such as anxiety and panic are much more prevalent. However, in group 2 (upper panel) it is the symptoms such as fatigue and sleep problems that are relatively prevalent, whereas anxiety and panic are much rarer.

(It is important to note here that the intercept parameters correspond to the expected symptom prevalences only for individuals with the mean latent trait score in each group. Since the symptoms are allowed to have different slope/loading parameters, the regression lines for some symptoms may have different gradients and may cross; therefore, the expected probabilities of experiencing particular symptoms need not be ordered the same across all levels of the latent trait. Correspondingly, even in the presence of strong measurement invariance, groups that differ substantially in their average levels of symptom severity may show different orderings of the intercepts for certain symptoms. This means that care is needed in interpreting observed differences in the pattern of intercepts — not all differences necessarily imply the absence of measurement invariance between groups.)

Since weak factorial invariance requires that the factor loadings/slopes are the same in the two groups, the individual symptoms can all be said to be equally good measures of the latent trait in the two groups. Theoretically, since the factor loadings are constrained to be the same in each group, the scale of the latent trait can also be said to be the same in each group; that is, the 'tick marks' on the horizontal axis will be equally spaced for each group, and a one unit change in latent trait severity should have the same meaning in each group. As a consequence, the scale of the latent trait only needs to be fixed in one group (although if a factor loading is fixed to 1 for one group, the presence of equal loadings across classes will mean that this constraint applies in the other group also).

However, in the situation illustrated in Figure 6.6 it seems clear that scores on the latent trait are not measuring the same form of symptom severity in each group — the latent trait appears dominated by anxiety-related symptoms in group 1, but by symptoms such as fatigue and sleep problems in group 2. Furthermore, since there are no common threshold/intercept parameters to act as 'anchor points' between the two groups, the location of the latent trait must be identified separately in the two groups by fixing the mean of the latent trait to 0 in each group — therefore, it is not possible to say whether a latent trait score from one group is higher or lower than a latent trait score from another group. The same trait score may be associated with very different probabilities of endorsing items in each group.

For analysis with ordinal items, there is a second reason why different intercepts may be required in each group, even if the relative ordering of expected probabilities is the same for the items in each group: this is that the *distances between* the intercept parameters may be different for the two groups. This is illustrated for two

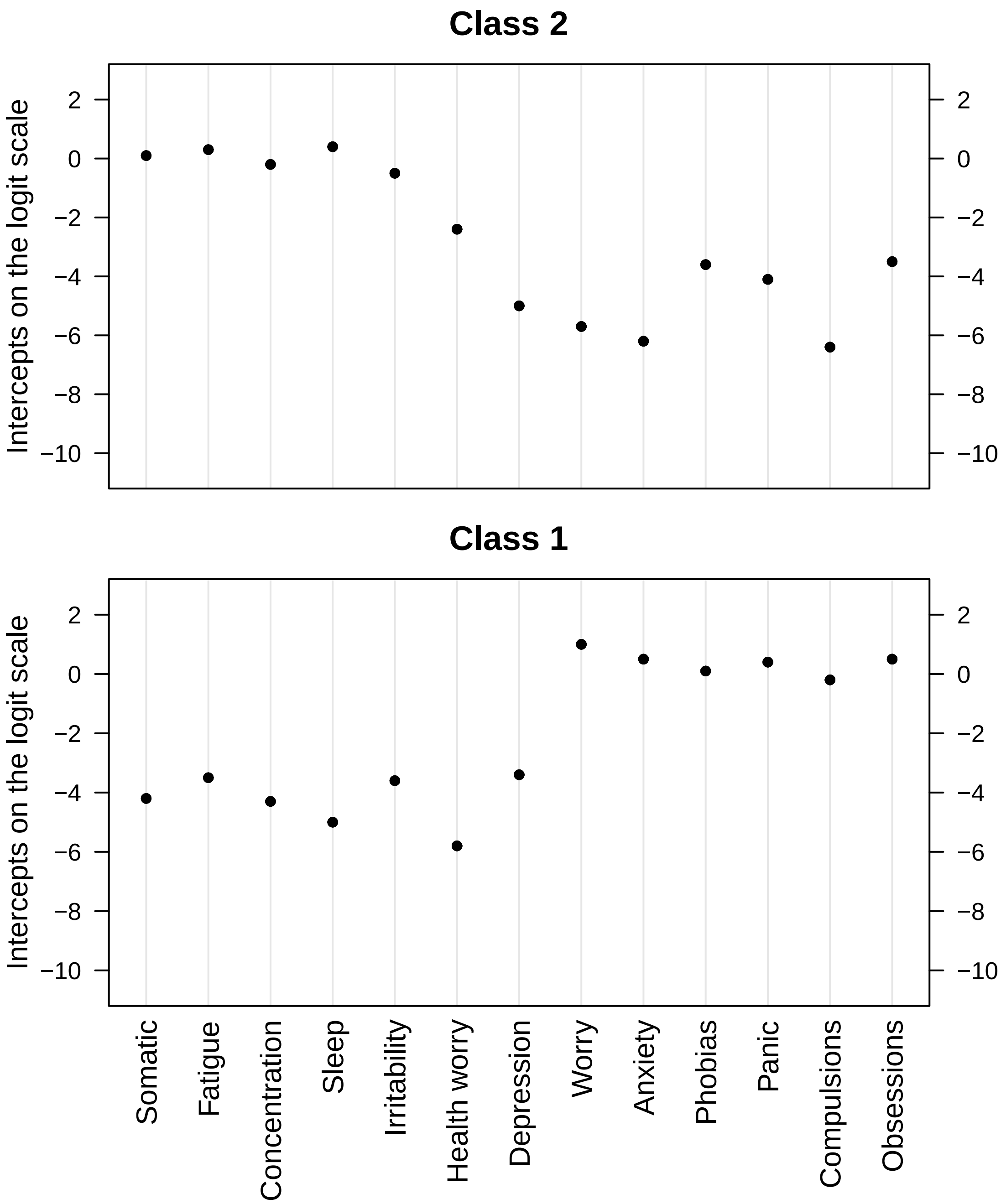

Figure 6.6: Hypothetical intercept parameters for a set of binary items in two groups showing weak factorial invariance. While the factor loadings are the same in each group, the intercepts parameters are allowed to vary between groups.

hypothetical groups with ordinal symptom items in Figure 6.7 — in group 1 (in the lower panel), the four intercept parameters are approximately equally spaced, whereas in group 2 (upper panel) the highest two intercepts are spaced much closer together than the lower intercepts.

(It is important to note here that the spacing between the regression lines for the four ordered scores on any item is constant at all levels of the latent trait; this is because all categories of an item share a single loading/slope parameter, as was illustrated in Figure 6.9. This means that, unlike differences in the relative orderings of intercepts, differences in the spacings between intercepts *cannot* be related to differences in average trait severity between the two groups. Substantial differences in the spacings between intercepts will *always* reflect an absence of measurement invariance between the two groups.)

As with the case of strong factorial invariance in the previous subsection, the factor variances for the two groups may be constrained to be equal or allowed to differ. However, the fit of the model may be impaired if the variances of the two groups are inappropriately forced to be equal.

6.2.4 No measurement invariance

The final situation occurs when both factor loadings and intercepts are allowed to vary across groups. For multidimensional factor models, this may or may not entail 'configural invariance' or 'equal form', in which there are the same number of factors in each group and the same items have loadings on each factor. However, this project is focussing on *latent* groupings in which it is not feasible to explore the factor structure separately in each group without some clear *a priori* theory. (We can't just run an exploratory factor analysis in each of the latent groups if we don't know who belongs to each group.) Therefore, where models are described as having 'no measurement invariance', this will simply refer to the specification of differing factor loadings and intercepts (but with the same pattern of factors and loadings). The scale and location of the latent trait must be set by fixing the factor mean and either the variance or a factor loading within each group. As in the case of weak measurement invariance, the lack of a shared metric means that it is again meaningless to compare the factor scores of individuals in different groups or to compare the mean factor scores of different groups. In fact, if the factor loadings differ substantially between the groups, factor scores may be measuring very different latent constructs in each group.

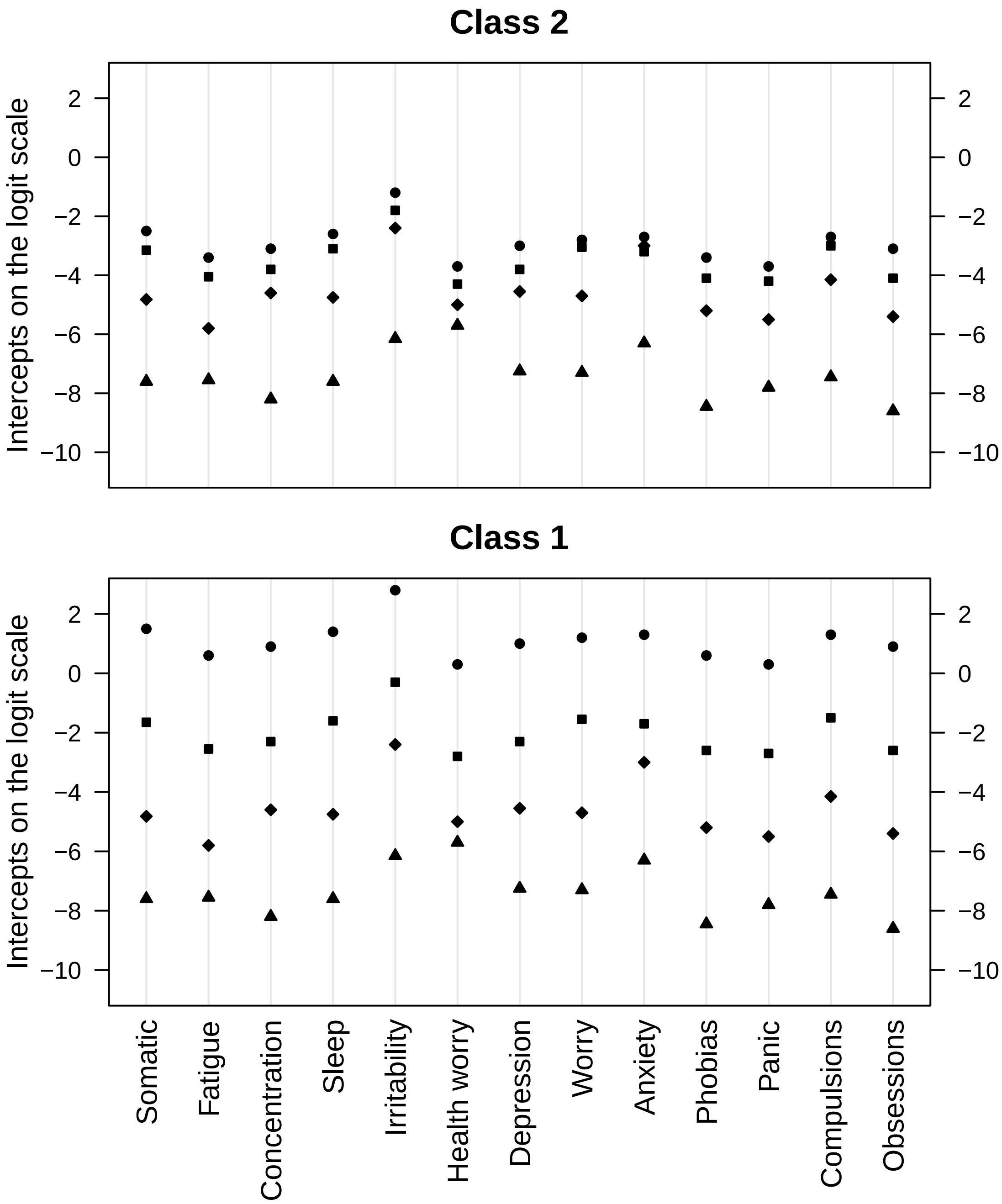

Figure 6.7: Hypothetical intercept parameters for a set of ordinal items in two groups showing weak factorial invariance.

6.2.5 Partial measurement invariance

The three scenarios above described situations in which all items were treated the same — if intercepts or loadings were allowed to vary, they were allowed to vary for every item. In the absence of *a priori* hypotheses about which items might be expected to lack measurement invariance, this seems a sensible way to proceed, since there are many possible combinations of parameters where measurement invariance could be relaxed, and selecting which constraints to relax on the basis of model fit indices would be likely to lead to overfitting. However, it is possible to specify multi-group factor analysis models with *partial* measurement invariance across a subset of items, and this may be particularly desirable in situations where a researcher suspects that a few specified items may behave differently for members of a particular group (such as items about forgetfulness for an elderly group of respondents). If the other items are known to have invariant intercepts/thresholds and loadings, then those items can be used to equate the location and scale of the latent trait across groups.

This may be useful for settings where the main aim of the analysis is to score individuals (such as educational testing): here, a multi-group model with partial measurement invariance may be able to 'adjust' scores for differential item functioning on a few items. However, the researcher should ideally know in advance which items may show differential item functioning and which items can be specified as non-invariant — where this is not the case, multiple testing may lead to some items being identified incorrectly as showing differential item functioning. This chapter will consider only models in which measurement non-invariance is allowed for all or no items. Since this project will not consider any models with this form of partial measurement invariance, it will not be discussed further.

6.3 Measurement invariance in factor mixture models

The principles of measurement invariance for multi-group factor analysis also apply to factor mixture models where the values of the grouping variable (class membership) are unknown and the class prevalences must be estimated along with the parameters of the within-class factor models. In factor analysis settings, the goal of comparing models with different degrees of measurement invariance will be to find which type of model fits the data best, and whether the data are suitable for certain types of analysis which require an assumption of measurement invariance. For this reason, researchers are recommended to follow a particular series of steps from fitting the least restrictive model to fitting the most restrictive model, at each stage checking that the addition of new constraints does not result in a serious deterioration in model fit before moving on to a more restrictive model (Brown, 2006).

In factor mixture modelling where the grouping variable is unknown, the situation is not so clearly defined — some configurations of measurement invariance can be seen as performing particular analytical roles, and so may be of interest to the researcher even where there is evidence that a model with fewer between-class equality constraints provides a substantially better fit to the data. The following paragraphs will describe how the different degrees of measurement invariance discussed in the previous section can be applied to factor mixture models, and will aim to illustrate the specific roles that each type of model can play.

6.3.1 Semi-parametric factor models

In the first type of factor mixture model, factor loadings and intercepts are constrained to be equal across latent classes (strong factorial invariance), and identification constraints are only required in one class. As a result of the loadings and intercepts being the same, the factor metric is identical in each class and equal factor scores represent the same levels of symptoms — therefore, it makes sense to compare the factor scores of individuals in different latent classes. Since continuous latent variables are assumed by most software to be normally distributed (or to follow a multivariate normal distribution in the case of two or more latent dimensions), the estimation procedure here is essentially trying to model the latent trait with a mixture of two or more normal distributions. This is very similar to the finite mixture modelling situation described in Section 5.1.2 on page 84 and illustrated in Figure 5.1, apart from the fact that the distributions being modelled are now of latent variables rather than observed variables (for example, severity of mental distress, as opposed to blood glucose levels).

As described in Section 5.1.2, the mixture components may be performing a direct or indirect role; that is, they could represent discrete sub-populations with their own symptom distribution (perhaps relating to a particular disease mechanism), or they could simply be a way of approximating the distribution of the latent variable with a mixture of normal distributions. Unfortunately, it is never possible to tell from the data alone which role the latent classes are playing; furthermore, the model may be hindered in identifying the 'ideal' mixture distribution by the fact that some regions of the latent trait may be poorly measured by items on the questionnaire

(Pickles & Angold, 2003). Since this type of factor mixture model is able to relax the assumption of the standard factor model that the latent variable must be normally distributed, it may be described as 'semi-parametric factor analysis' (Masyn *et al.*, 2010) — this term will be used for the rest of the present study to describe factor mixture models in which factor loadings and intercepts are constrained to be equal, and where the factors in each class differ only in terms of their means and variances/covariances. Sample Mplus code for estimating a semi-parametric factor model is provided in Section D.4 on page 280 of Appendix D.

6.3.2 Latent class factor models

A slightly different factor mixture model can be obtained by again constraining factor loadings and intercepts to be equal across classes (so this model also demonstrates strong factorial invariance). In this case, the factor variance is fixed to be 0 within each class — this means that all individuals belonging to that class share the same factor score, which is identical to the mean factor score for that class. Such models may perform a direct function by representing an underlying latent dimension that is composed of a fixed number of discrete categories rather than being continuous in nature. While the classes may represent discrete categories, the locations of these classes along the latent dimension (as reflected by their factor means) can be measured on a continuous scale, unlike the classes in a standard latent class model that can sometimes be ordered but cannot be quantitatively compared.

In addition to their direct applications, these models can also be very useful in their indirect role — this is because of their ability to approximate the distribution of an underlying latent dimension by a set of discrete classes without requiring any assumptions that the trait is normally distributed (either overall, or within a latent class as for semi-parametric factor analysis described in the previous section). This is similar to the approach described as 'non-parametric maximum likelihood estimation' by Rabe-Hesketh *et al.* (2003) in the context of using a latent variable to adjust for measurement error in a skewed covariate of a regression model.

This ability to approximate any underlying continuous distribution with a set of discrete classes is related to the situation illustrated for 'standard' latent class analysis in the left-hand panel of Figure 5.2 on page 86, hence the description as 'latent class factor analysis' (e.g., in Muthen, 2006). However, given the fact that these models require no distributional assumptions about the nature of the underlying latent dimension, the models are sometimes described as 'non-parametric factor analysis' (Masyn *et al.*, 2010). Nonetheless, the name 'latent class factor model' will be used for the rest of this thesis to describe factor mixture models in which loadings and intercepts are constrained to be equal while factor variances are fixed to 0 within each class — this is to avoid the possibility of confusion with some non-parametric approaches used in the following chapter. Sample Mplus code for estimating a latent class factor model is included in Section D.5 on page 281 of Appendix D.

Although there are clear similarities between latent class analysis and latent class factor analysis in their ability to approximate continuous underlying dimensions, it is important to note that these latent class factor analysis models are quite different from latent class models. In latent class analysis, the probabilities of obtaining each item score are directly estimated within each class — this means that, even where the classes are clearly ordered, no type of mathematical response function need be specified to describe how the probabilities of obtaining each score change across classes of increasing severity. However, estimating the probabilities for every item category in every class requires a large number of parameters, and where some classes are small the estimated probabilities may be sensitive to random fluctuations in the data.

In the case of latent class factor analysis (as with standard categorical factor or IRT models), a mathematical function such as the logit function must be used to describe how the probability of scoring in each category changes for classes of increasing severity. While latent class factor models rely on the assumption that this function is appropriate for the data, they require far fewer parameters than latent class models to provide a reasonable approximation of a continuous underlying distribution with a particular number of classes: in addition to the single set of loadings and intercepts needed for the overall factor model, each extra class requires only the estimation of the mean factor score for that class and the class prevalence. In contrast, each additional class in latent class analysis requires an estimate of the class prevalence along with a full set of item category intercepts (a total of 53 parameters per class for these data).

While it may seem that there is a considerable loss of information in latent class factor analysis compared to standard factor analysis as a result of allowing the factor to take on only a handful of discrete values, this is not a serious limitation. Since allocation of individuals to classes is probabilistic, the factor score for each individual is calculated as a weighted average of the mean factor scores for each class, with the weights being the individual's estimated probabilities of belonging to each class. For even a small number of classes, these 'remixed' factor scores can approximate a continuous distribution of factor scores as well as the 'correct' smooth distribution (Pickles & Angold, 2003) — three or four classes may be sufficient for a factor measured by a small number of indicators, although more may be needed where there are many items. In fact, this latent class approach is very similar to the method that Mplus and many other software packages use for maximum likelihood estimation of the standard factor model — this is known as Gaussian quadrature, and is described in the following paragraph.

Since the likelihood equations that are required for estimating the factor analysis model cannot be written in a way that allows them to be solved algebraically, a procedure known as Gaussian quadrature is used to approximate the likelihood. With this method, the continuous latent variable is replaced with a pre-specified number of discrete quadrature 'nodes' (perhaps as few as 5, although more is better; Schilling & Bock, 2005); these quadrature nodes correspond closely to the representation of the latent variable by a set of classes in latent class factor analysis. The likelihood for a particular response pattern can then be approximated by a weighted sum of the likelihoods at each of the quadrature nodes.

The key difference between a standard continuous latent variable model estimated with Gaussian quadrature and a latent class factor model is that during model estimation with quadrature the weights of the nodes are pre-determined from the known properties of the normal distribution, whereas the weights of the latent class factor model (the class prevalences) are estimated from the observed distribution of the data (Rabe-Hesketh *et al.*, 2003). This means that, while standard factor analysis relies on the assumption that the factor is normally distributed, latent class factor analysis requires no assumptions about the distribution of the latent variable. Thus, the latent class factor model may be useful in cases where the distribution of the latent trait is believed to be non-normal, at least as a sensitivity analysis — if the estimated factor loadings are very different under the two approaches, this implies that the results are sensitive to the normality assumption of the standard factor model.

6.3.3 Factor mixture models without strong factorial invariance

The two preceding sections have covered types of factor mixture analysis in which the imposition of strong measurement invariance allows for the direct comparison and 'remixing' of factor scores between classes; this is because the metric of the factor is exactly the same in each class. However, when the item intercepts (and possibly also the factor loadings) are allowed to vary, the metric of the factor must be fixed separately within each class, and therefore no direct comparisons of factor scores or 'remixing' is possible — within each class, the factor mean must be fixed at 0, and either the factor variance or one of the loadings must be fixed to define the scale of factor within that class. While we can compare the factor scores of individuals in the same class, all we can say about individuals in different classes is that they are in different classes.

In line with Masyn *et al.* (2010), the term 'factor mixture model', will from now on be reserved for such models without strong factorial invariance, even though the semi-parametric and latent class factor models described earlier are also special cases of the factor mixture model. Sample Mplus code for a factor mixture model in which factor loadings are constrained to be equal across classes but intercepts are allowed to differ is provided in Section D.6 on page 283 of Appendix D. Sample code for a factor mixture model in which both loadings and intercepts are allowed to vary is provided in Section D.7 on page 285.

It is possible to specify different numbers of factors or different patterns of factor loadings in the different classes. However, such models will not be considered here since the author has no *a priori* theoretical grounds for choosing one of the many possible configurations over another. Therefore, all factor mixture models estimated in this study will have the same number of factors and the same pattern of factor loadings in each class. Nonetheless, these types of factor mixture model still offer an enormous degree of flexibility, and they may play many different roles. Where the factor loadings are approximately the same across classes and the pattern of intercepts is mostly similar, the factors can be considered to be measuring the same construct in each class — in this case, the latent classes *may* be describing latent subgroups that exhibit differential item functioning (DIF); for example, some symptoms may be more strongly related to the latent trait in one group than another.

However, where there are large differences between classes in the factor loadings or the ordering of item intercepts, the factor may be measuring completely different constructs in each latent class. For example, in one class the factor may measure predominantly the severity of anxiety symptoms, while in another class the factor may measure predominantly symptoms related to depression. This type of factor mixture model may be seen as an extension of the 'direct' function of latent class analysis (see Section 5.1.2) in which the conditional independence requirement is relaxed within class, i.e., the classes really do represent distinct but unobserved subgroups of the population, although the groups are not completely homogeneous since individuals can differ in trait severity within each class.

While both of the situations above describe 'direct' functions of the classes in a factor mixture model (they identify subgroups exhibiting differential item functioning or subgroups where the latent trait measures a different concept in each class, respectively), factor mixture model classes may also play an 'indirect' role. (This will be discussed in detail in the next chapter, which will focus on how the models to be reported later in this chapter can be interpreted.) Therefore, care must be taken when interpreting the results of any factor mixture analysis — just as with the correspondence between factor and latent class models, it may be difficult to determine from the data whether the classes of a favoured factor mixture model do indeed represent qualitatively distinct categories of individuals, or whether they are playing some form of indirect role.

6.3.4 Alternative representations of the same structure

It is important to note that there will be no single 'correct' choice for the number of factor mixture model classes in a particular set of data — the make-up and number of the classes that are extracted will be closely bound up with the way that the within-class model is specified and the role that the classes are performing. For example, more classes will be needed to adequately describe the relationships in the data if two correlated factors are constrained to be uncorrelated within each class (as illustrated in Figure 6.8 on the next page), or if factor variances and covariances are fixed to 0 within each class (as in the latent class factor models described above). However, this does not make the model with more classes (as in the right-hand panel of Figure 6.8) incorrect — it is simply an alternative way to describe the same latent structure. In some cases, models with conditional independence or 0 factor variance within each class may be considerably easier to interpret than those that allow for severity differences, since individuals are homogeneous within each class; this means that the class properties can be easily summarised by their factor means or estimated item prevalences.

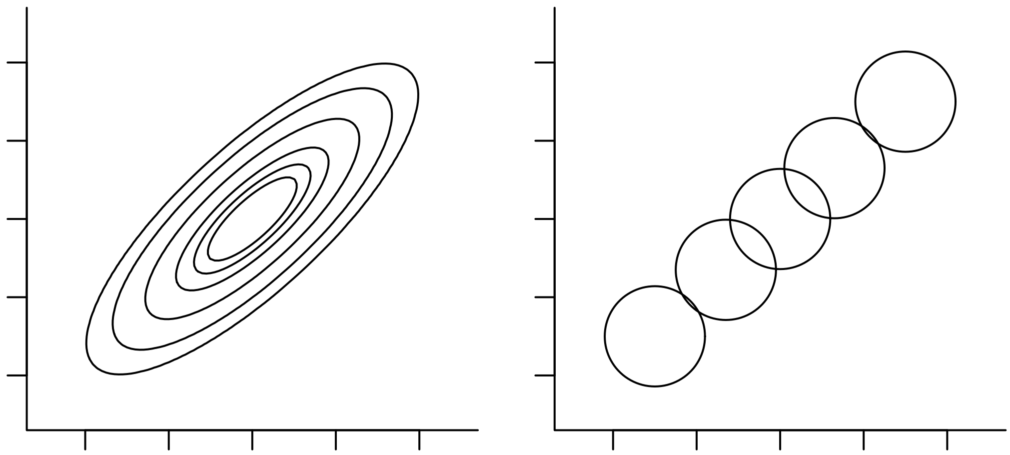

Figure 6.8: Alternative representations of the latent structure of a hypothetical data set containing two correlated latent dimensions. In the left-hand panel the factors are allowed to be correlated and so their relationship can be described well by a single class factor mixture model (equivalent to a standard factor model), whereas in the right-hand panel the factors are constrained to be uncorrelated within each factor mixture model class.

6.4 Factor mixture model results

The next sections will examine the results of fitting semi-parametric and latent class factor models to the CIS-R data. Given that the CIS-R data can be described as essentially unidimensional (as was discussed in Section 4.2.3), this chapter will focus solely on unidimensional factor mixture models; this means that models may include any number of latent classes, but can only allow individuals to vary along a single latent dimension within each class. (The following chapter will consider multi-dimensional extensions to some of these models, in order to check whether possible miss-specification of the factor structure is resulting in the extraction of additional latent classes.)

Since there are so many different factor mixture models to consider in this chapter, the key results relating to likelihood-ratio tests, model fit indices, replication in the second split half and bivariate residuals will be presented in a single table. Table 6.1 shows these statistics for the semi-parametric and latent class factor models fitted to the first split half of the CIS-R data, along with the fit statistics from the standard single factor model for comparison.

Table 6.1: Model comparison table for unidimensional factor mixture models with measurement invariance

Model	# par	LL	BIC	AIC	Relative entropy	LMR aLRT p value	Smallest class size (proportion)	Validates in split half	Overall bivariate Pearson chi-square	Pairs with significant* lack of fit (out of 78)	Bivariate standardised residuals > 3 (outof 1950)
Standard factor model as baseline for comparison											
1f 1c	65	-90,962	**182,531**	182,055	-	-	-	Y	3,974	35	95
Semi-parametric factor models (intercepts and factor loadings constrained to be the same across classes)											
1f 2c v.s.	67	Not identified									
1f 2c	68	Not identified									
1f 3c v.s.	69	-90,954	182,552	182,047	0.48	-	1,526 (0.14)	Y	4,100	39	98
1f 3c	70	No repeatable and/or replicated solution									
1f 4c v.s.	71	Not identified									
Latent class factor models (intercepts and factor loadings the same across classes — variances fixed to 0)											
1f 2c	66	-93,381	187,378	186,894	0.87	0.1683	3,003 (0.27)	Y	13,525	74	393
1f 3c	68	-91,452	183,538	183,040	0.80	0.0000	1,058 (0.09)	Y	5,848	62	166
1f 4c	70	-91,045	182,742	182,229	0.76	0.0000	396 (0.04)	Y	4,438	44	111
1f 5c	72	-90,971	182,613	182,086	0.70	0.0003	205 (0.02)	Y	4,207	40	100
1f 6c	74	-90,958	182,607	182,064	0.68	0.2112	63 (0.01)	N	4,136	39	98
1f 7c	76	-90,954	182,616	182,060	0.63	0.0421	36 (0.00)	N	4,136	39	98
1f 8c	78	Not identified									

Abbreviations: # par, number of parameters estimated in the model; LL, log-likelihood; BIC, Bayesian information criterion; AIC, Akaike information criterion; LMR aLRT, Lo-Mendell-Rubin adjusted likelihood ratio test; f, factor; c, class; v.s., variances constrained to be equal. All models fitted to the same random half of the data. n = 11,230. *P value used for signficance cut-off = 0.01.

6.4.1 Semi-parametric factor models

As can be seen from Table 6.1, considerable difficulties were encountered when trying to fit semi-parametric single factor models to the CIS-R data. Models with 2 classes (each class including a normally distributed factor) were not identified, even after trying hundreds of different sets of random starting values. While model non-identification may be an indication that the researcher is trying to estimate a model with too many latent classes (Uebersax, no date), this difficulty may also be related to the large number of individuals with 'all 0' response patterns (who didn't report any of the symptoms at all, and whose scores are undefined without making some kind of assumption about the shape of the overall distribution of factor scores — see Section 4.2.2 on page 66). In these 2 class models, either the mean factor score or the class prevalence was not identified in the class containing the 'all 0' response pattern. This suggests that the added flexibility in the shape of the factor distribution may have a downside: it may leave factor scores for individuals with 'all 0' response patterns undefined.

The three class semi-parametric model with factor variances constrained to be equal was identified. However, Table 6.1 indicates that the improvement in log-likelihood for this model over the standard factor model is small. The increased Bayesian information criterion (introduced on page 58), suggests that the improvement in likelihood is too small to justify inclusion of the 4 additional parameters required for this model. The overall mixture distribution implied by the estimated factor means and variances of the three classes in the semi-parametric model are shown in Figure 6.9 — the overall distribution appears to be roughly normal, although with rather fat tails. Given the very limited information in the data about the lower end of the factor score distribution, little can be concluded from the slight asymmetry of the distribution. There is certainly no evidence of the kind or bi-modality or multi-modality that might be suggestive of the classes performing a direct rather than indirect function. This model could also be estimated in the second split half of the data, again resulting in an approximately normal overall mixture distribution (although with a different shape to the lower tail) — a graph is available for comparison in Figure E.3 on page 292 of Appendix E.

The relative entropy of 0.48 for the 3 class semi-parametric model seems low (relative entropy was introduced on page 60). If the classes were performing a direct function by representing true subgroups of the population, low relative entropy would be a concern — this would mean that the model leaves us with a lot of uncertainty about which classes the individuals in our sample should be allocated to. However, low relative entropy is not really a concern when the latent classes are believed

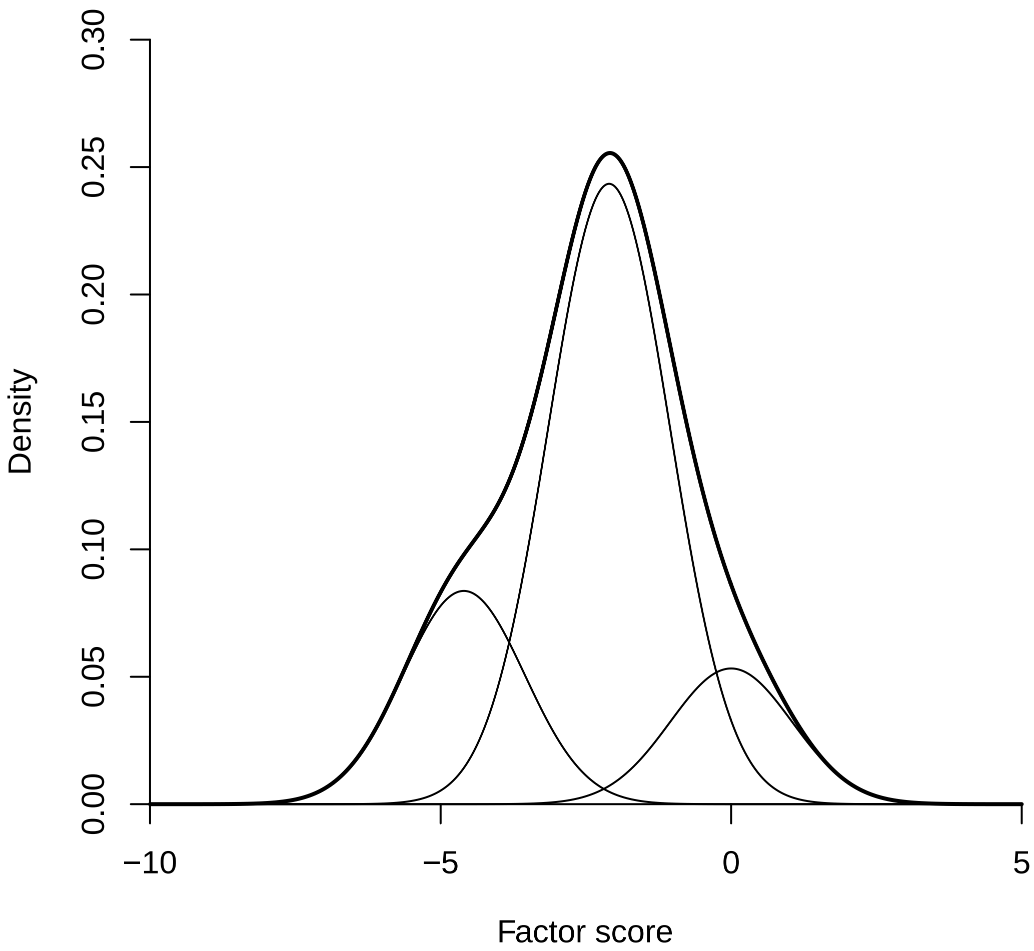

Figure 6.9: The mixture of normal distributions implied by the three class semi-parametric factor model in Table 6.1 — the variances of the three classes were constrained to be equal.

to be performing an indirect function: when the classes all represent parts of one single group, it doesn't matter if we are uncertain about which individuals should be allocated to each class.

As Table 6.1 indicates, the 3 class semi-parametric model in which the variances were allowed to vary in each class produced no satisfactory solution — the best log-likelihood could not be replicated, even using the parameter estimates associated with the best log-likelihood as starting values. Estimation of the factor variance for the class including the 'all 0' response pattern appeared to be problematic for this model. Furthermore, the 4 class semi-parametric model was not identified, even when the factor variances were constrained to be equal. Given the difficulties estimating these models, no further semi-parametric models were considered.

Factor scores for individuals can be estimated from the semi-parametric factor model. The factor score can be estimated by 'remixing' the factor scores: the remixed factor score is a weighted sum of the score that the individual would receive if they belonged to each class, weighted by their estimated probability of belonging to that class. (Although an individual's estimated factor scores might be expected to be the same in each class given that the factor model parameters are the same in each class, they may vary a little since factor scores in each class are shrunk towards the mean factor score in each class.) The estimated factor scores from the standard factor model and 3 class semi-parametric models are compared in Figure 6.10. The upper halves of the distributions appear to be very similar, although the distribution of factor scores for those with mild but not 'all 0' symptoms appears slightly different. A scatterplot comparing the estimated factor scores for each individual from the standard factor model and 3 class semi-parametric model is shown in Figure 6.11. The points fall extremely close to a straight line, apart from the very highest scores. This suggests that relaxing the normality assumption of the latent trait distribution through a semi-parametric factor model makes little difference for these data to the factor scores that are estimated for each individual.

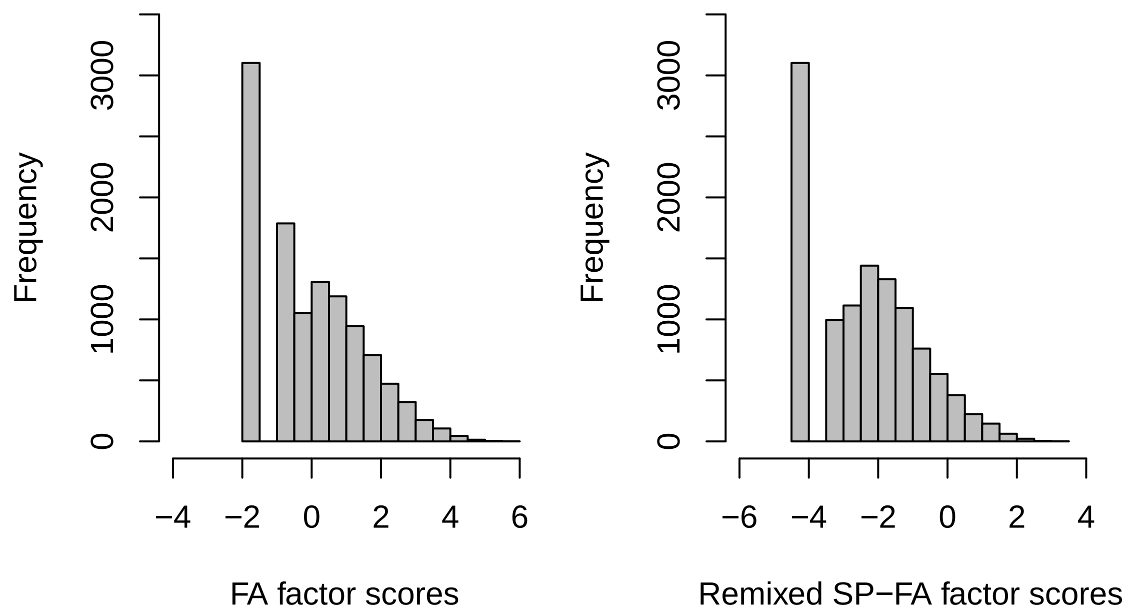

Figure 6.10: Histograms comparing the distributions of factor scores from the standard single factor model (FA) and the remixed factor scores from the 3 class semi-parametric factor model (SP-FA).

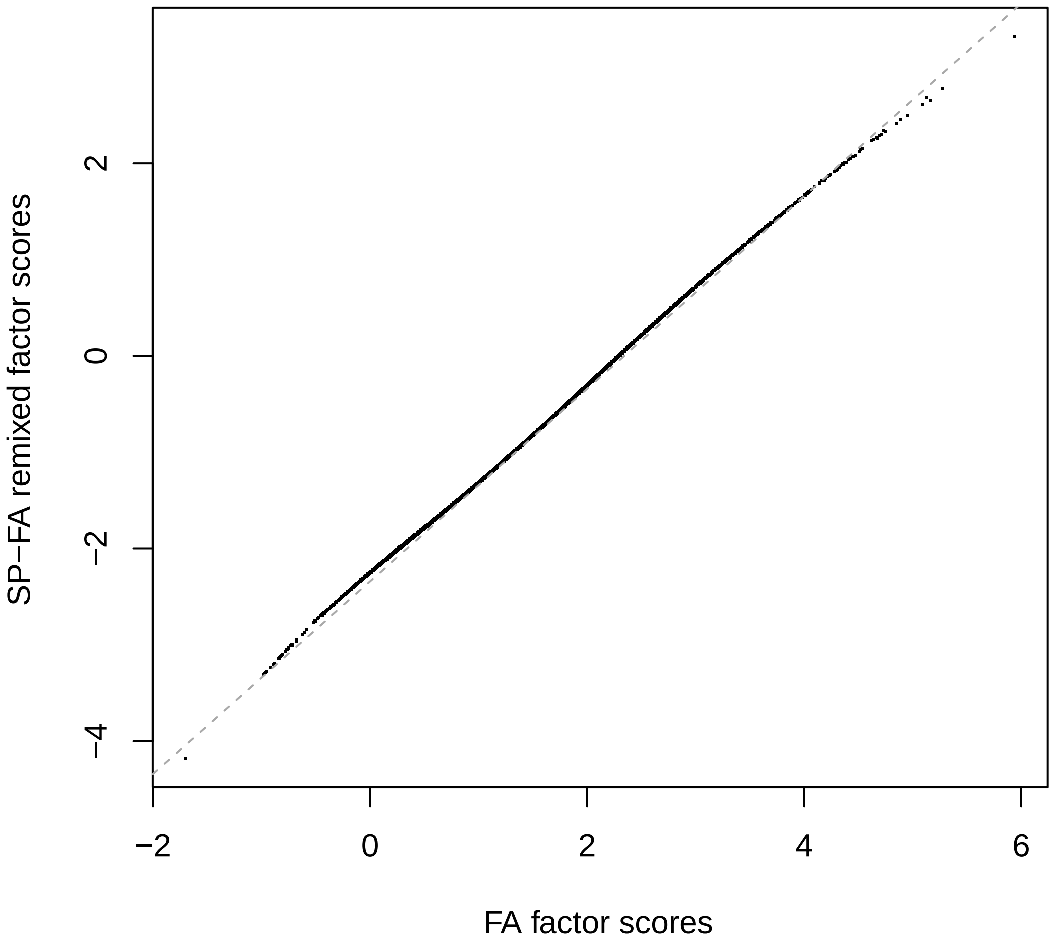

Figure 6.11: Scatterplot comparing the distributions of factor scores from the standard single factor model (FA) and the remixed factor scores from the 3 class semi-parametric factor model (SP-FA).

6.4.2 Latent class factor models

Table 6.1 also shows the fit statistics from a number of latent class factor models that have been applied to the CIS-R data. The most classes that could be successfully estimated in the first half of the data was 7 — the left-hand panel of Figure 6.12 shows the estimated factor means and prevalences of these 7 classes, while the right-hand panel shows the 'remixed' factor scores of individuals based on this model. The distribution of factor scores appears very similar to that from the semi-parametric factor model in Figure 6.10, which itself was similar to the estimated scores from the standard factor model. As in the case of the remixed scores from the semi-parametric model, the correspondence between individuals' estimated factor scores from the standard factor model and their remixed scores from the 7 class latent class factor model is very close (as shown in Figure 6.13), apart from those for a handful of individuals at the very highest levels of the latent trait.

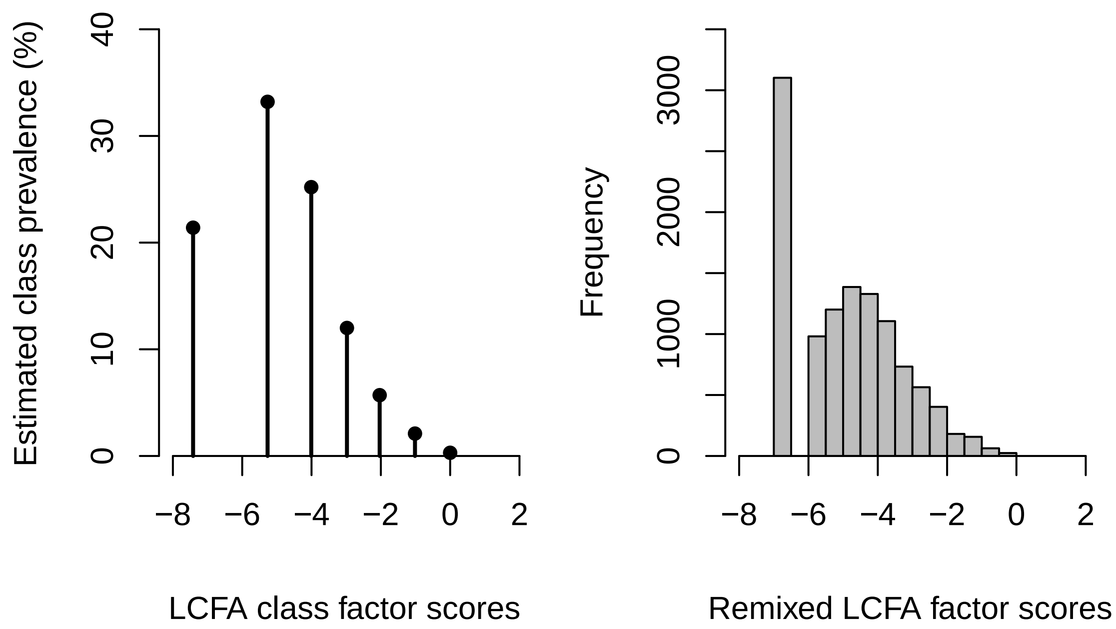

Figure 6.12: Factor scores from the 1 dimensional 7 class latent class factor model (LCFA). The left-hand panel shows the estimated mean factor scores in each of the classes. The right-hand panel shows the remixed factor scores for all individuals in the sample (which are a sum of the 7 class means weighted by the individual's probability of belonging to that class).

Since the factor variance is specified to be 0 within each class of the latent class factor model, the probabilities of endorsing symptoms within each class can be summarised

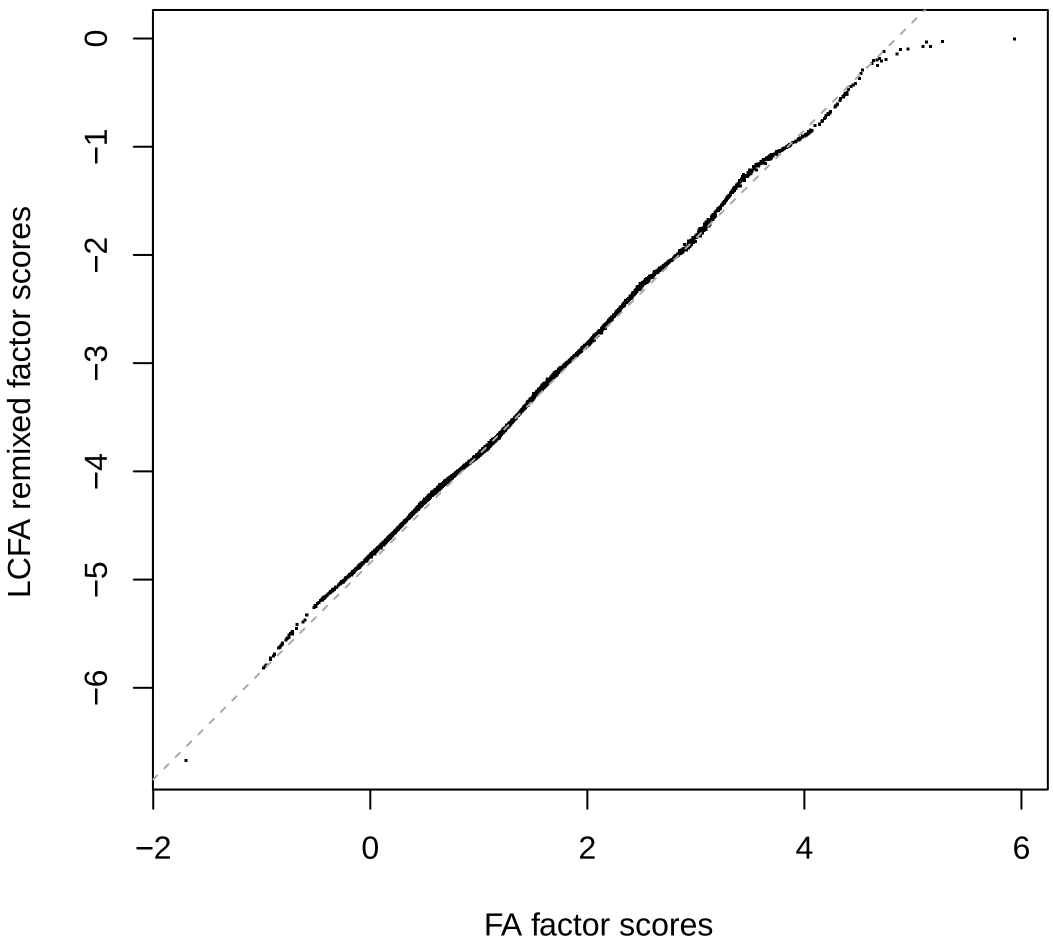

Figure 6.13: Scatterplot comparing the distributions of factor scores from the standard single factor model (FA) and the remixed factor scores from the 7 class latent class factor model (LCFA).

efficiently in a profile plot in the same way as for latent class models. Figure 6.14 shows the estimated probabilities of scoring 2 or more on each symptom in the 7 classes; what is striking about this chart is its similarity to the ordered profiles of some latent class models in Chapter 5. The 4 classes of the latent class model in Figure 5.3 on page 92 match very closely with classes 2, 3, 4 and 6 (counting from the bottom) in Figure 6.14. (The clearest exception is the absence of the small 'peak' for panic symptoms in the highest class of the latent class model). This highlights the fact that the observed class profiles in latent class models do not need to track each other's rises and falls exactly (sometimes described as being 'parallel') in order to indicate that a dimensional model may fit the data very well.

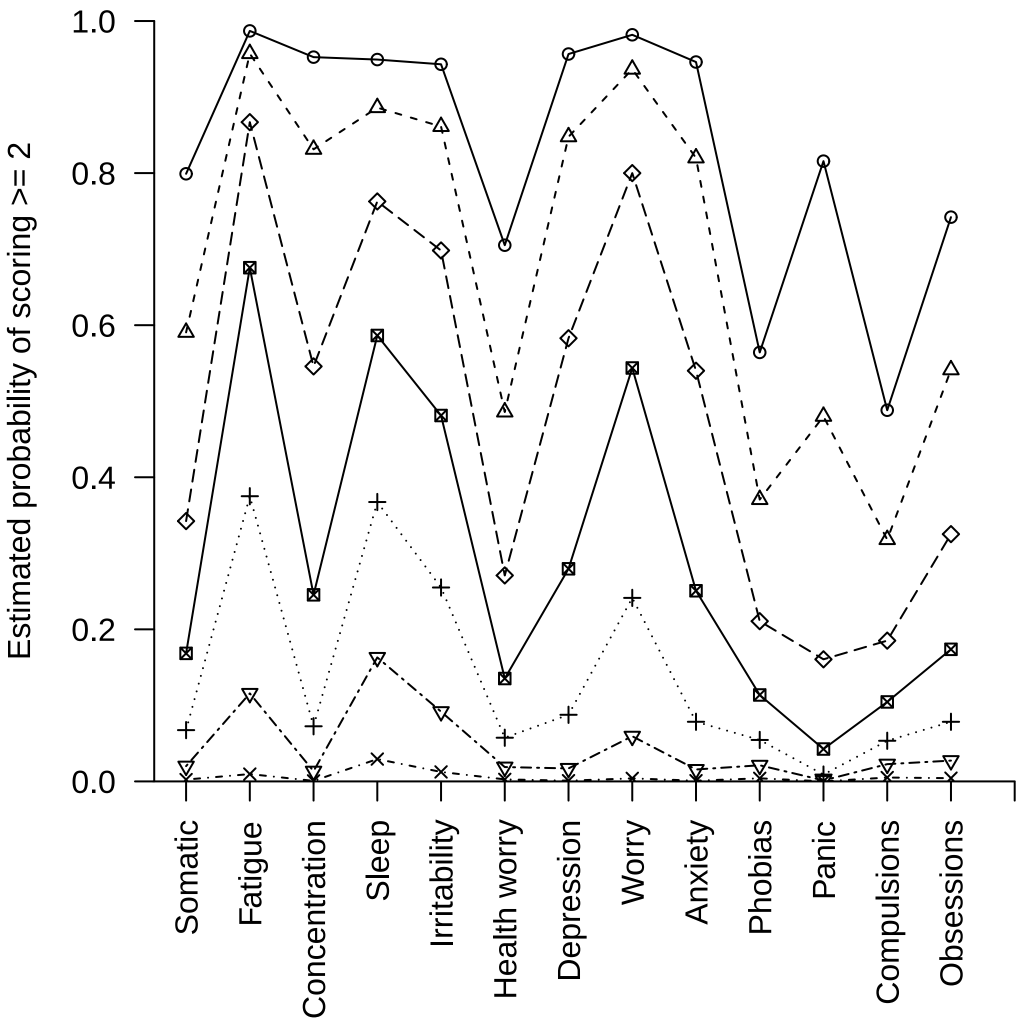

Figure 6.14: Probabilities of scoring 2 or more on each symptom for each class in the 7 class latent class factor model.

Although the previous paragraphs have focussed on the results of the 7 class latent class factor model, Table 6.1 indicates that the maximum number of classes that could be estimated in the second split half of the data was 5. Since the classes are being assumed here to be playing the indirect function of approximating a continuous dimension, the fact that the 7 class model does not replicate is not a major cause for concern. However, it does indicate that the maximum possible number of classes that can be used to approximate an underlying distribution will be determined by the characteristics of the data for each sample, not just by the underlying latent structure.

The results of the 5 class model for the first half of the data are shown in Figure 6.15 — it appears that this model does not do such a good a job of approximating the continuous underlying distribution as the 7 class model in Figure 6.12 (although the 'bins' used for plotting this histogram may be exaggerating the irregularities in the distribution of remixed scores). The equivalent 5 class latent class factor model from the second split half of the data can be seen for comparison in Figure E.4 on page 293 — in the second split half, the remixed distribution is even more irregular. These results seem to suggest that more classes are better when it comes to approximating a continuous distribution with latent class factor analysis, even though there is little or no improvement in model fit or residuals (see Table 6.1).

Even though the semi-parametric and latent class factor models produce remixed factor scores that are extremely similar to those from the standard factor model, it is still possible that the estimated factor loadings may vary between the different types of model. This might indicate that inappropriately assuming a normal factor distribution results in biased estimates of the factor loadings. For comparison, the estimated unstandardised loadings from the three different types of factor model are shown in Table 6.2.

The loadings are very similar for each model type, suggesting that the assumption of a normal factor distribution is not biasing these parameter estimates. The standard errors are also very similar for each of the models, suggesting that there is no loss of precision from using the semi-parametric or latent class form of the model. The one noticeable difference is in the variance of the estimated factor scores — these are not the model estimates for the variance of the latent trait calculated during marginal maximum likelihood estimation, but the variances of the 'expected a posteriori' factor scores (which include remixed factor scores as well as standard factor scores). The variance of these scores is a little smaller for the standard model than for the semi-parametric and latent class factor models; this may reflect greater 'shrinkage' towards the overall factor mean in the standard factor model.

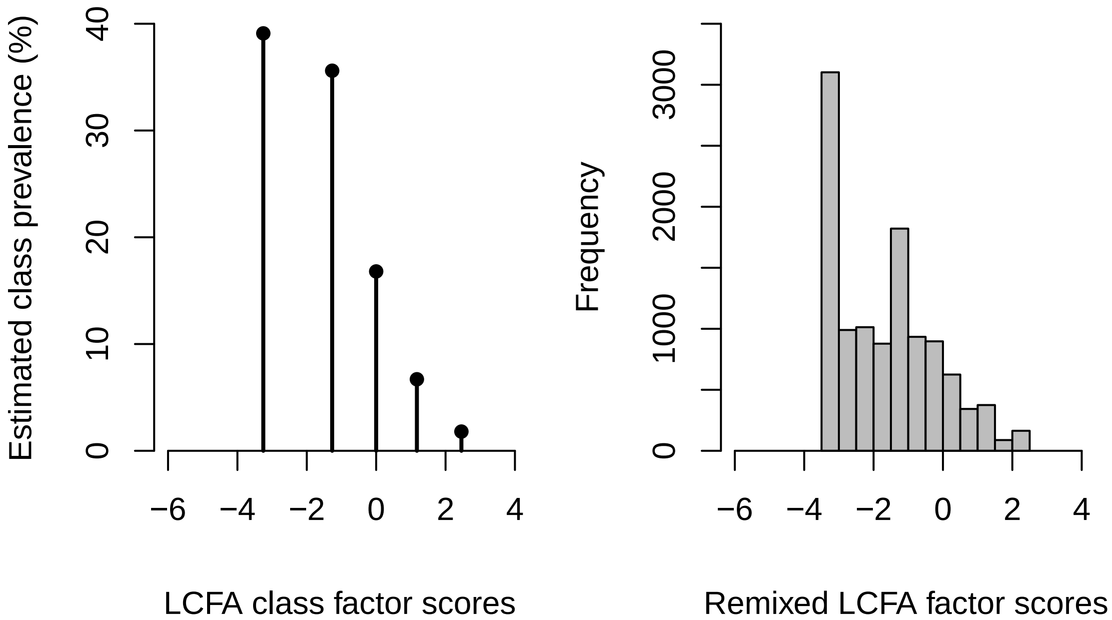

Figure 6.15: Factor scores from the 1 dimensional 5 class latent class factor model (LCFA). The left-hand panel shows the estimated mean factor scores in each of the classes. The right-hand panel shows the remixed factor scores for all individuals in the sample (which are a sum of the 5 class means weighted by the individual's probability of belonging to each class).

Table 6.2: Comparison of **unstandardised** loadings from the standard factor model (FA), the semi-parametric factor model (SP-FA) with 3 classes and the latent class factor model (LCFA) with 7 classes

	FA	SP-FA (3c)	LCFA (7c)
Symptom	Estimate (SE)	Estimate (SE)	Estimate (SE)
Somatic symptoms	1*	1*	1*
Fatigue	1.21 (0.04)	1.21 (0.04)	1.21 (0.04)
Concentration/forgetfulness	1.37 (0.05)	1.39 (0.05)	1.38 (0.06)
Sleep	0.87 (0.03)	0.87 (0.03)	0.87 (0.03)
Irritability	0.97 (0.04)	0.97 (0.04)	0.97 (0.04)
Worry about physical health	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)
Depression	1.34 (0.05)	1.36 (0.05)	1.36 (0.05)
Worry	1.28 (0.05)	1.28 (0.05)	1.28 (0.05)
Anxiety	1.32 (0.05)	1.33 (0.05)	1.33 (0.05)
Phobias	0.78 (0.04)	0.78 (0.04)	0.78 (0.04)
Panic	1.53 (0.08)	1.54 (0.08)	1.55 (0.08)
Compulsions	0.71 (0.04)	0.71 (0.04)	0.71 (0.04)
Obsessions	0.88 (0.04)	0.88 (0.04)	0.88 (0.04)
Estimated factor variance	2.77	-	-
Variance of estimated scores	2.12	2.57	2.62

* These loadings were fixed to 1 to identify the scale of the latent variable.

6.4.3 Factor mixture models: intercepts allowed to vary

The model comparison statistics for factor mixture models in which the intercepts are allowed to vary across classes are shown in Table 6.3 (as well as for the models in which both intercepts and loadings are allowed to vary that will be discussed in the following section). This table also includes the fit statistics for the standard (1 class) factor model for comparison. According to the Bayesian information criterion, the 3 class model provides the best description of the data. The Lo-Mendell-Rubin adjusted likelihood ratio test (aLRT, introduced on page 59) also supports the choice of a 3 class model, while rejecting the need for a 4 class model. However, it may be noted that the Lo-Mendell-Rubin test is a little ambiguous about whether or not the 2 class model actually describes the data better than a standard (1 class) factor model — the p value is 0.08, which would not be rejected by standard cut-off criteria despite the large likelihood ratio test statistic of 1,017. If indeed the 2 class mixture model cannot describe the data better than the standard 1 class model, significant p values for the addition of further classes are somewhat irrelevant.

The measures of bivariate goodness-of-fit (described on page 61) all suggest that the 2, 3 and 4 class factor mixture models provide a much better fit to the data than the standard factor model. The overall bivariate Pearson chi-square statistic falls dramatically from 3,974 to 2,135 with the inclusion of the second class, and then falls further to 1,667 with the inclusion of the third class. The reduction to 1,431 with the inclusion of the fourth class is less impressive. The number of variable pairs showing poor fit and the number of standardised bivariate residuals also fall dramatically with the inclusion of the second class, along with more modest changes at the inclusion of the third and fourth classes.

However, none of the factor mixture models achieve the levels of classification accuracy that would be desirable in order to include class allocation as a variable in further analyses (as discussed on page 60). The 2 class model has a relative entropy of 0.66, which is well below the desired level of 0.8, while the 3 class model has an even worse relative entropy of 0.58. This does not reflect the quality of the models or their goodness-of-fit, but it does mean that many individuals in the sample cannot be confidently classified into a particular latent class. High levels of certainty would be desirable for classes that corresponded to diagnostic groupings, but this is not provided by any of the models here. The 2 and 3 class models both resulted in similar classes when applied to the second split half of the data, but the additional class of the 4 class model appeared to have a very different interpretation in the second half, and so will not be considered further as a candidate model to describe the latent structure of these data.

Table 6.3: Model comparison table for unidimensional factor mixture models without strong measurement invariance

Model	# par	LL	BIC	Relative entropy	LMR aLRT statistic	LMR aLRT p value	Smallest class size (proportion)	Validates in split half	Overall bivariate Pearson chi-square	Pairs with significant* lack of fit (out of 78)	Bivariate standardised residuals > 3 (out of 1950)
Standard factor model as baseline for comparison											
1f 1c	65	-90,962	182,531	-	-	-	-	Y	3,974	35	95
Factor mixture models — equal loadings, intercepts allowed to differ											
1f 2c	119	-90,453	182,015	0.66	1,017	0.0778	2,134 (0.19)	Y	2,135	8	17
1f 3c	173	-90,141	**181,895**	0.58	622	0.0251	1,385 (0.12)	Y	1,667	4	9
1f 4c	227	-90,010	182,137	0.65	261	0.2106	467 (0.04)	N	1,431	0	6
Factor mixture models — loadings and intercepts allowed to differ											
1f 2c	131	-90,420	182,062	0.70	1,080	0.2315	2,186 (0.19)	N	2,071	14	16
1f 3c	197	-90,080	181,997	0.56	680	0.1530	1,325 (0.12)	N	1,573	4	7
1f 4c	Not identified										

Abbreviations: # par, number of parameters estimated in the model; LL, log-likelihood; BIC, Bayesian information criterion; LMR aLRT, Lo-Mendell-Rubin adjusted likelihood ratio test; f, factor; c, class.
All models fitted to the same random half of the data. n = 11,230. *P value used for significance cut-off = 0.01.

The proportions of individuals with each symptom score for the classes of the 2 class factor mixture model are shown in Figure 6.16. The first large class has relatively low levels of all symptoms, although with around 30% of the class reporting primarily low or moderate levels of fatigue, sleep problems, irritability or worry. The second small class is more symptomatic: as well as showing much higher levels and more severe symptoms of fatigue, sleep problems, irritability and worry, it also shows substantial levels of depression and anxiety, as well as more modest levels of the other symptoms. The proportions presented in Figure 6.16 represent the overall proportions of each symptom score that were endorsed by all individuals in each class (although the proportions for any one class are actually obtained by weighting the responses from the entire sample by their model estimated probabilities of belonging to that class).

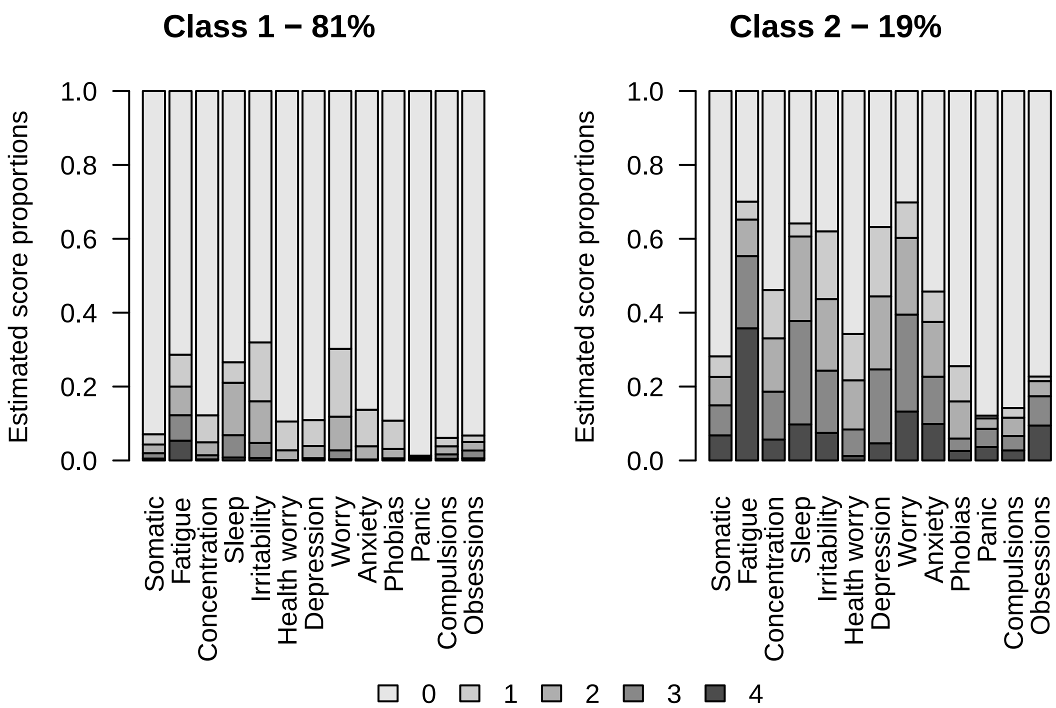

Figure 6.16: Estimated score proportions in each class of the 2 class factor mixture model with intercepts allowed to vary but loadings constrained to be equal across classes.

Since the factor mixture model allows individuals to vary in severity within classes, it is important to remember that individuals may vary considerably in their probabilities of endorsing symptoms within each class. This is in contrast to the latent class models in the previous chapter, where the estimated probabilities of responding in each category were identical for all individuals in a particular class, and corresponded directly to the category intercepts for that class. For the factor mixture models shown here, applying the inverse logit transformation to the category intercepts would yield the probabilities of responding in each category for an individual with a 0 factor score (the mean score) in each class. The corresponding factor mixture model estimated on the second split half of the data shows a similar pattern of symptoms for the two classes, and these are available for comparison in Figure E.5 on page 294 of Appendix E.

The estimated score proportions for the factor mixture model with 3 classes are shown in Figure 6.17. The largest class includes modest levels of fatigue, sleep problems and irritability, but low levels of all other symptoms. The medium sized class is distinguished by high levels of mild or moderate worry, in addition to the symptoms present in the largest class. The smallest class presents a similar profile to the smallest class in the 2 class model, although the estimated class prevalence is lower and the levels of depression, worry and anxiety are slightly higher. The corresponding model estimated on the second split half of the data shows a very similar pattern of symptoms for the three classes, and these are available for comparison in Figure E.7 on page 296.

As noted above, the additional class in the factor mixture model with 4 classes differed in the two split halves of the data. This suggests that the model may be overfitting the data. In the first split half, the model includes a class in which every member reports some degree of phobia symptoms. However, in the second split half the four classes roughly correspond to: low levels of all symptoms; high levels of fatigue but low levels of worry; low levels of fatigue but high levels of worry; and lastly, high levels of both fatigue and worry (along with high levels of the other symptoms). This pattern is actually suggestive of the presence of two correlated dimensions dominated by fatigue and worry that are being ignored by the choice of a single factor inside the factor models — this possibility will be returned to again in the next chapter. For the interested reader who wishes to check this '2-dimensional' pattern, the estimated score proportions from the 4 class model in the second split half are presented in Figure E.9 on page 298.

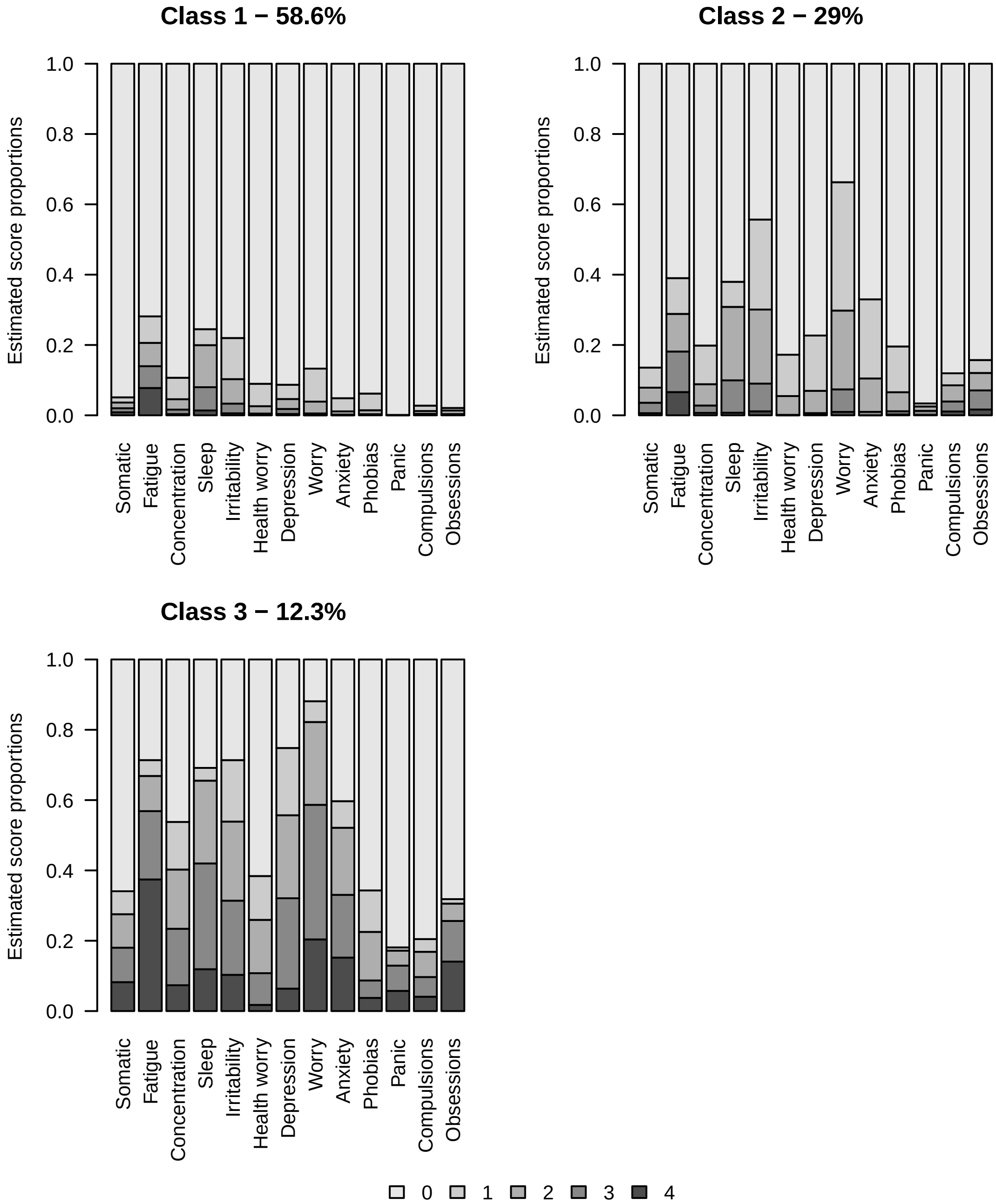

Figure 6.17: Estimated score proportions in each class of the 3 class factor mixture model with intercepts allowed to vary but loadings constrained to be equal across classes.

6.4.4 Factor mixture models: loadings and intercepts allowed to vary

The fit statistics for factor mixture models in which loadings are also allowed to vary were presented in Table 6.3 on page 142. The class prevalences and estimated score proportions from the 2 class and 3 class models are actually very similar to those from the models in the previous subsection in which intercepts alone were allowed to vary, so to save space they are not shown here. Relaxing the equality constraints on the factor loadings theoretically allows the factor to measure disorder severity for different constellations of symptoms in each class — the symptom clusters being measured in such a situation would be revealed by the patterns of high factor loadings for each class. Given this, it seems reasonable to require that patterns of high factor loadings are replicated in the second split half of the data — otherwise, any attempts to identify the meanings of the different classes will be misleading. However, this replication does not occur for either the 2 class or 3 class models. The four class model was not identified.

Table 6.4: Estimated loadings within each class for the 2 class factor mixture model in which loadings and intercepts are allowed to differ between classes

	Class 1 (80.5%)	Class 2 (19.5%)
Symptom	Loading (SE)	Loading (SE)
Somatic	1.5 (0.2)	0.8 (0.1)
Fatigue	1.5 (0.5)	0.6 (0.1)
Concentration/forgetfulness	1.7 (0.7)	1.2 (0.1)
Sleep	1.0 (0.2)	0.7 (0.1)
Irritability	1.5 (0.2)	1.0 (0.1)
Health worry	1.1 (0.3)	0.8 (0.1)
Depression	1.6 (0.1)	1.4 (0.3)
Worry	**2.0** (0.4)	**1.7** (0.4)
Anxiety	**2.1** (0.3)	**1.7** (0.3)
Phobias	1.1 (0.2)	1.3 (0.3)
Panic	**2.4** (0.4)	**2.1** (0.3)
Compulsions	1.2 (0.2)	1.3 (0.4)
Obsessions	1.6 (0.4)	1.1 (0.2)

Note: Factor variances were fixed to 1 in each class to identify the scale of the latent variable. The largest three loadings in each class are highlighted in bold.

Table 6.4 shows the estimated loadings for the 2 class model — the highest loadings in both classes are for the symptoms of worry, anxiety and panic. However, for the same model estimated on the second split half, the highest loadings shown in

Table E.2 on page 299 are for fatigue, concentration and panic in the first class, and for fatigue, concentration and depression in the second class. The pattern of highest loadings for the 3 class model also fails to replicate, as can be seen by comparing the loadings in Table 6.5 with those from Table E.3 on page 300. This instability suggests that the great flexibility of the model formulation is allowing the factor loadings to change in response to chance fluctuations in the data, rather than in response to the underlying latent structure. For this reason, factor mixture models in which the loadings are allowed to vary between classes will not be considered further in this project.

Table 6.5: Estimated loadings within each class for the 3 class factor mixture model in which loadings and intercepts are allowed to differ between classes

	Class 1 (54.7%)	Class 2 (33.5%)	Class 3 (11.8%)
Symptom	Loading (SE)	Loading (SE)	Loading (SE)
Somatic	1.7 (0.1)	1.1 (0.1)	0.9 (0.1)
Fatigue	**3.1** (0.3)	**1.8** (0.2)	**1.5** (0.2)
Concentration/forgetfulness	2.2 (0.1)	**2.1** (0.2)	**1.7** (0.1)
Sleep	1.5 (0.1)	0.9 (0.1)	0.9 (0.1)
Irritability	1.5 (0.1)	1.0 (0.1)	1.0 (0.1)
Health worry	1.6 (0.1)	0.9 (0.1)	1.0 (0.1)
Depression	1.8 (0.1)	1.3 (0.2)	0.9 (0.1)
Worry	2.0 (0.5)	1.0 (0.1)	0.7 (0.1)
Anxiety	**3.1** (0.9)	1.0 (0.2)	0.7 (0.1)
Phobias	0.8 (0.1)	1.0 (0.2)	1.2 (0.1)
Panic	**2.9** (1.4)	**1.8** (0.3)	**1.4** (0.2)
Compulsions	0.7 (0.1)	0.9 (0.1)	1.1 (0.1)
Obsessions	1.5 (0.3)	0.8 (0.1)	0.5 (0.1)

Note: Factor variances were fixed to 1 in each class to identify the scale of the latent variable. The largest three loadings in each class are highlighted in bold.

6.5 Conclusion

This chapter started with an overview of factor mixture models and a discussion of the implications that different levels of measurement invariance have for the roles that the factor mixture model can perform. The first models to be considered were those with strong measurement invariance, where intercepts and factor loadings were constrained to be equal across groups: semi-parametric and latent class factor models. It was noted that these models may be particularly useful for relaxing the normality assumption in standard factor models. However, relaxing the assumption of a normally distributed latent trait appeared to make very little difference to the fit of the models, and the estimated factor scores and factor loadings were also extremely similar to those from the standard factor model. Therefore, these semi-parametric and latent class factor models will not be considered further in this project.

In contrast, factor mixture models with weak factorial invariance (where intercepts were allowed to vary but not loadings) appeared to show a substantial improvement in fit over the standard factor model — the mixture model with 3 classes appeared to fit the best, but the 2 class model also showed a clear improvement over the standard model. However, none of the models in which loadings were additionally allowed to vary across classes could be successfully replicated in the second split half of the data — this was also the case for the 4 class model in which loadings but not intercepts were constrained to be equal. These discrepancies highlight the risks of overfitting the data that are inherent in these models, as well as the usefulness of splitting the data in this way.

Given that mixture models with 2 or 3 classes appear to describe the data better than the standard factor model (while also requiring fewer parameters than the well-fitting latent class models described in the previous chapter), the next important task is to try to identify what role the mixture components are playing — do they correspond to true groups in the population (perhaps individuals with and without a common mental disorder), or are they simply accommodating some form of miss-specification in the standard model? This question is crucial to the interpretation of the results of the factor mixture models reported here, and will be the focus of the next chapter.

Chapter 7

Latent variable modelling IV: Factor mixture analysis — interpretation

7.1 How can we interpret the classes of the factor mixture models?

The previous chapter demonstrated that factor mixture models with 2 or 3 classes provide a better fit to the data than a standard factor model in which all individuals are assumed to belong to a single population. A natural response to this may be to assume that the underlying population is heterogeneous, and is composed of at least two distinct groups (perhaps corresponding to those with and without a common mental disorder). In this case, a careful examination of the parameters within each latent class and of the allocations of individuals to classes should allow the characteristics of the classes to be identified. However, there is an alternative possibility that must also be considered before this kind of conclusion can be drawn: the classes in the factor mixture analysis may be playing an 'indirect' rather than a 'direct' role. (See Section 5.1.2 on page 84 for a fuller discussion of what it is meant by the distinction between 'direct' and 'indirect' roles of latent classes.) In this case, the latent classes may just be relaxing some of the assumptions of the standard factor model, rather than dividing individuals in the sample into distinct and meaningful groups.

The aim of this chapter is to investigate as fully as possible whether the latent classes from the 2 and 3 class factor mixture models described in the previous chapter are performing a direct or indirect role. This question has not been investigated explicitly in any of the previous papers that have used factor mixture models to investigate the latent structure of particular psychiatric symptoms, so there is no standard approach that can be adopted; therefore, the conclusions from this exercise may be tentative. However, the question seems too important to leave unaddressed; indeed, attempting to resolve this ambiguity seems vital if the results are to be used to inform theories about the nature of mental disorders. Furthermore, if an answer cannot be suggested by the researcher who has total access to the model results and original data, it will be virtually impossible for other researchers to provide an answer armed with only briefly reported findings.

While it may not be possible to answer this question definitively, a careful consideration of the alternatives should at the very least suggest that some interpretations are more plausible than others. The chapter will therefore go through the possible roles that the classes may be performing in turn, in each case looking for evidence that the classes are indeed performing that role. The chapter will start by considering whether the factor mixture model classes are consistent with a 'direct' interpretation — this will involve investigating the extracted classes to see whether they appear to correspond to clinically meaningful groups. The chapter will then move on to consider three particular indirect roles that classes in a factor mixture model may play: accommodating non-normality of the latent trait; accommodating miss-specification of the model structure; and finally, accommodating some form of misfit from the logistic function that is used in the standard factor model.

7.2 A direct role? Examining the classes from the 2 and 3 class factor mixture models

7.2.1 The 2 class model

The 2 class factor mixture model consisted of one large class with an estimated prevalence of 81% and one small class with an estimated prevalence of 19%. The proportions endorsing each symptom score among those allocated to each class were illustrated in Figure 6.16 on page 143 — this indicated that the small class had much higher levels of all symptoms (with particularly high levels of fatigue, sleep problems, irritability, depression and worry), while the large class showed much lower levels of symptom endorsement (although around a third of the class reported mostly mild levels of fatigue, sleep problems, irritability and worry). From such a summary of the two classes, it may appear that theses two classes differ simply in their overall levels of symptom severity. However, if this were indeed the case a much more parsimonious model would be able to account for such a difference in severity, as described in Section 6.2.3 on page 116: in the semi-parametric factor model described in the previous chapter, a single parameter could measure the difference in mean factor scores between the two classes, as opposed to the set of 52 intercepts that must be estimated for each additional class in the factor mixture model. In fact, the semi-parametric and non-parametric factor models showed very little improvement in fit over the standard factor model (as reported in Table 6.1 on page 130), suggesting that the situation is not this simple. As discussed in Section 6.2.3, the need for differing intercepts implies that the classes must differ either in the relative *orderings* of the symptoms, or in the *spacings* between the intercept parameters. These differences in intercept parameters may correspond to patterns of symptoms associated with particular mental disorders.

Since many of the symptoms reported by those in the larger 'mild' class are low levels of non-specific symptoms such as fatigue, sleep problems, irritability and worry, it is possible that this class represents those without any disorder (although members may still report a number of general symptoms), while the smaller 'severe' symptoms class represents those who do have a true disorder. However, since Figure 6.8 gives only aggregate proportions endorsing each symptom, this chart may conceal the presence of individuals with high levels of important symptoms within the 'mild' class.

For this reason, it may be worth examining the symptom profiles of the individuals allocated to Class 1 who had the most severe symptoms. Table 7.1 shows the symptom profiles for the 14 individuals in this class who had the highest total symptom scores (obtained from summing the 13 separate symptom scores), along with their probabilities of belonging to Class 1. It is clear from this table that despite Class 1 appearing to be the 'mild' class, some of the individuals allocated to this class have high overall levels of symptoms and report key symptoms such as depression, anxiety and panic alongside non-specific symptoms. The highest total score among those allocated to Class 1 is 34. The table also shows the ICD-10 common mental disorders diagnosed for each individual by the CIS-R algorithm: all of the individuals in this table meet the criteria for at least one ICD-10 disorder, and half meet the criteria for two or more, ruling out any possibility that the symptoms of these individuals could be dismissed as 'sub-clinical' or transient. These findings do not support the idea that the classes represent a simple division into individuals with and without a common mental disorder.

Table 7.1: The response profiles, probabilities of class membership and ICD-10 diagnoses for the 14 individuals in class 1 with the highest total symptom scores

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total score	Class probability	ICD-10 diagnoses
4	4	4	4	4	1	0	3	2	1	3	4	0	34	0.549	Panic disorder
1	4	3	2	2	2	1	2	1	2	3	4	3	30	0.971	Agoraphobia, OCD
3	3	2	3	2	2	4	2	1	1	3	1	3	30	0.961	Depressive episode (severe), Panic
2	4	3	3	3	1	0	2	2	3	3	3	0	29	0.921	GAD, Specific phobia
4	3	3	4	1	1	2	3	2	1	0	2	3	29	0.617	GAD
4	3	3	4	1	1	2	3	2	1	0	2	3	29	0.617	GAD
3	4	4	3	3	0	0	4	2	1	0	1	3	28	0.590	GAD
0	4	3	2	3	1	1	1	2	2	4	2	3	28	0.904	GAD, Agoraphobia, Social phobia
1	2	2	2	3	0	1	2	2	3	2	4	4	28	0.880	OCD, Social phobia
2	2	4	2	3	0	2	4	0	3	3	1	1	27	0.529	Specific phobia
0	4	3	4	2	1	0	3	1	2	3	4	0	27	0.522	OCD, Agoraphobia
1	4	3	3	3	0	2	3	2	3	0	2	0	26	0.506	Specific phobia
0	3	4	3	2	1	2	3	2	0	2	1	3	26	0.824	Depressive episode (moderate)
1	4	4	4	1	1	2	3	2	1	0	0	3	26	0.504	GAD
3	4	4	3	3	0	2	2	2	1	2	0	0	26	0.761	Depressive episode (moderate), GAD

Abbreviations: OCD, obsessive compulsive disorder; GAD, generalised anxiety disorder.

Note: The total score is the sum of the 13 symptom scores. Symptom scores of 3 or 4 have been highlighted for easier identification.

Some of the individuals listed in Table 7.1 have fairly low probabilities of belonging to Class 1: those with probabilities close to 0.5 have roughly similar probabilities of belonging to either of the two classes, so the classification certainty for these individuals is very poor. However, other individuals have very high probabilities of belonging to this class, and those with low and high class probabilities cannot be easily distinguished in terms of their total symptom scores or the presence of particular symptoms or diagnoses. Given that some of the individuals allocated to Class 1 with high probabilities have such high levels of symptoms, it is clear that the classes are not just dividing individuals by the likely presence or absence of any common mental disorder, or even into groups with mild/absent versus severe symptoms. At this point, it is not clear why some individuals are assigned to one class rather than the other. This suggests the need for further analysis of the characteristics of individuals allocated to each of the two classes.

Bar charts summarising overall levels of symptom endorsement in each class clearly cannot give us enough detail about the types of individual that are allocated to each class to characterise the classes. A useful next step would therefore be to examine the ranges of symptom levels that are present in the two classes — these are shown in Figure 7.1. This figure considers total symptom scores obtained by summing the 13 separate symptom scores; this is similar to the 'CIS-R score' obtained by summing all 14 sections scores from the CIS-R interview. However, the reader should note that the scores used here exclude any score obtained on the 'depressive ideas' section of the CIS-R interview, since this item was excluded from all of the latent variable modelling. Nonetheless, for most individuals the total scores presented here will be the same or very similar to their CIS-R score (particularly for those with 0 scores for the depression section), and so the scores will have roughly similar interpretations. The convention for CIS-R scores is that a score of 12 or more can be used to identify a probable psychiatric 'case' (Lewis *et al.*, 1992), and in fact all individuals with a score of 12 or more who do not meet the criteria for any other ICD-10 common mental disorder are given the diagnosis of 'mixed anxiety and depressive disorder' by the CIS-R algorithm.

Figure 7.1 reveals that Class 1 is dominated by a large number of individuals with very low or zero total scores. However, this class also includes individuals with higher levels of symptoms, and there is a considerable overlap in the ranges of total scores between the two groups. In fact, for total scores roughly in the range of 12-16 there appear to be similar numbers of individuals allocated to each of the two classes. It may therefore be helpful to examine individuals within this particular overlapping section of the symptom score distribution to see what distinguishes individuals with high probabilities of belonging to one class or the other: since both

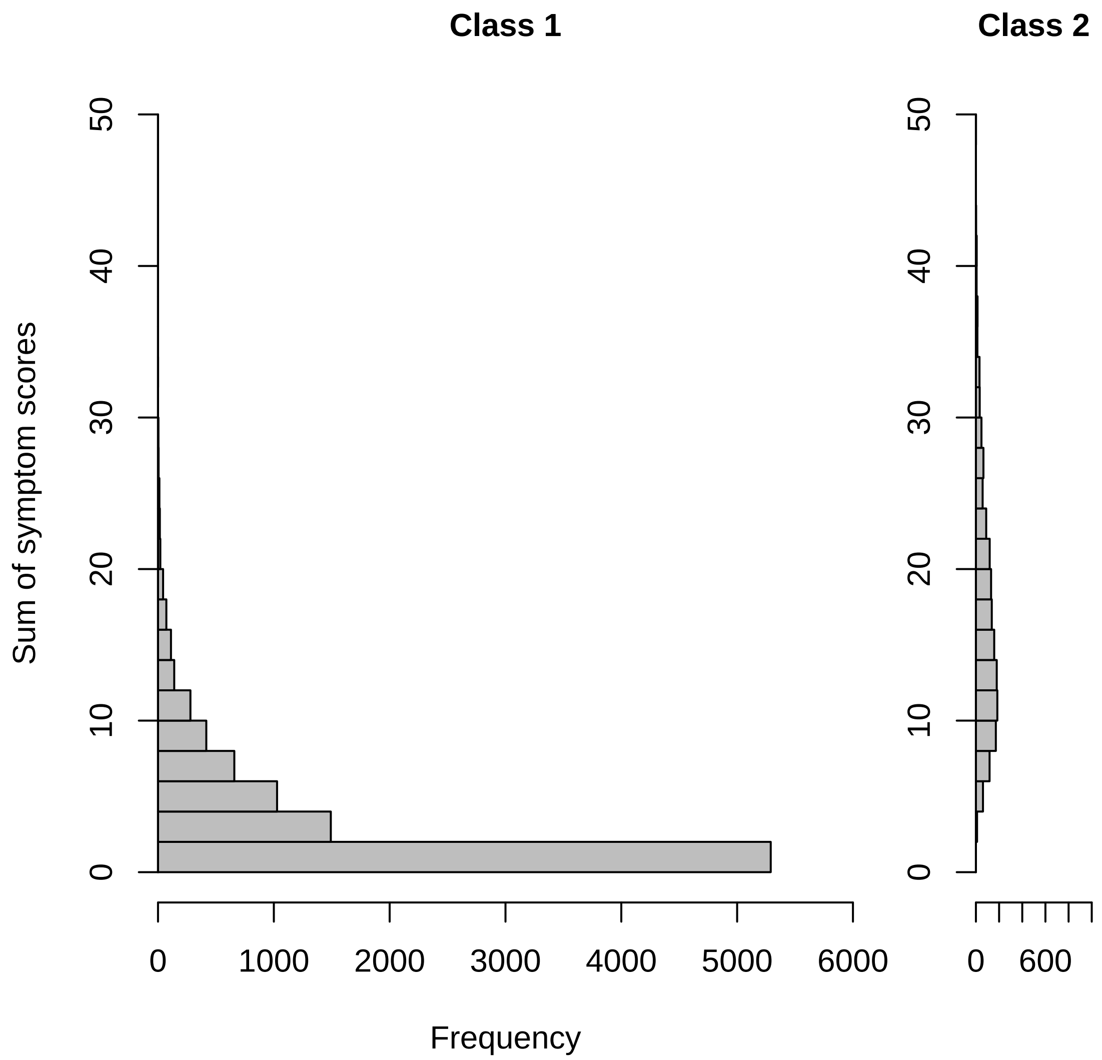

Figure 7.1: Sideways histograms showing the frequencies of total symptom scores in each of the classes of the 2 class factor mixture model — the total symptom scores are a sum of the 13 individual section scores.

groups will have the same overall level of symptoms, examining the response profiles may reveal particular patterns of symptoms that influence an individual's probability of belonging to one class or the other. Table 7.2 presents the response profiles of the 30 individuals obtaining total scores of 12-16 with the highest probabilities of belonging to Class 1, while Table 7.3 does the same for the 30 individuals with the highest probabilities of belonging to Class 2.

Table 7.2: The 30 response profiles with the highest probabilities of belonging to class 1 among individuals with total symptom scores between 12 and 16

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total	Probability
1	1	0	0	3	1	0	2	1	1	0	2	2	14	0.998
0	4	2	2	2	0	0	1	0	1	0	1	1	14	0.994
2	2	2	0	3	0	0	1	1	1	0	0	2	14	0.994
0	0	1	1	2	2	1	1	1	2	0	0	1	12	0.994
0	3	1	1	1	1	1	0	1	2	0	3	1	15	0.992
0	2	0	2	0	1	1	2	1	1	0	1	2	13	0.992
1	3	2	0	1	1	0	1	1	0	0	2	0	12	0.990
2	4	1	1	2	0	2	1	1	0	0	0	1	15	0.989
0	2	1	2	3	0	1	1	1	1	0	0	0	12	0.988
0	2	1	2	2	0	0	2	1	0	0	0	2	12	0.985
0	1	1	2	3	0	1	2	1	1	0	0	3	15	0.984
3	3	1	1	1	1	1	1	1	0	0	0	0	13	0.982
2	4	0	1	4	0	0	1	1	1	0	2	0	16	0.981
0	0	0	1	2	2	0	2	1	1	0	3	0	12	0.981
0	2	2	2	1	1	0	2	0	1	1	0	0	12	0.979
0	1	1	2	2	0	2	1	0	1	0	2	0	12	0.979
2	2	2	2	0	1	0	0	1	0	1	2	0	13	0.978
0	2	2	2	1	1	1	1	2	3	0	0	1	16	0.976
2	3	0	1	2	1	0	2	1	0	0	0	3	15	0.976
1	4	0	0	2	0	0	1	1	2	0	3	0	14	0.974
2	3	2	1	1	2	1	2	1	1	0	0	0	16	0.972
2	2	1	0	2	2	1	1	1	0	0	0	0	12	0.972
0	3	1	2	1	1	0	1	2	0	0	0	2	13	0.971
4	2	1	2	2	0	0	1	1	0	0	0	0	13	0.970
0	3	0	2	1	0	0	2	1	1	0	3	0	13	0.970
0	3	0	2	2	0	1	2	1	1	0	2	0	14	0.969
1	3	0	2	2	1	0	2	1	0	0	0	0	12	0.969
1	2	1	0	1	1	0	2	0	0	0	2	3	13	0.968
0	4	0	3	1	0	0	1	1	0	1	2	0	13	0.968
0	4	1	3	2	0	1	1	1	0	0	1	0	14	0.967

Note: Symptom scores of 3 and 4 have been highlighted in order to facilitate comparisons with Table 7.3.

Table 7.3: The 30 response profiles with the highest probabilities of belonging to class 2 among individuals with total symptom scores between 12 and 16

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total	Probability
0	4	1	0	2	0	2	3	4	0	0	0	0	16	1.000
0	4	0	4	0	3	3	2	0	0	0	0	0	16	1.000
0	0	2	3	2	0	0	4	4	1	0	0	0	16	1.000
0	0	0	3	1	0	0	4	4	0	0	0	4	16	1.000
0	0	0	2	3	0	1	3	4	2	0	0	1	16	1.000
0	1	2	0	2	0	3	3	4	0	0	0	0	15	1.000
2	2	0	0	0	3	3	2	3	0	0	0	0	15	1.000
4	0	0	0	0	2	0	3	4	2	0	0	0	15	1.000
0	4	0	2	2	0	2	0	4	0	0	1	0	15	1.000
0	4	0	2	2	0	0	3	4	0	0	0	0	15	1.000
0	2	0	3	2	0	1	2	4	0	0	0	0	14	1.000
0	0	0	0	4	0	2	4	4	0	0	0	0	14	1.000
0	3	0	3	1	0	0	3	4	0	0	0	0	14	1.000
0	0	1	2	3	1	0	3	4	0	0	0	0	14	1.000
0	2	1	3	2	0	1	1	4	0	0	0	0	14	1.000
0	0	0	2	0	2	3	0	4	0	0	0	3	14	1.000
0	2	1	0	1	0	2	0	4	0	3	0	0	13	1.000
1	1	0	2	2	0	0	3	4	0	0	0	0	13	1.000
0	0	2	3	0	0	1	3	4	0	0	0	0	13	1.000
0	2	1	0	0	0	0	2	4	2	2	0	0	13	1.000
0	0	0	3	0	0	0	4	4	1	0	0	0	12	1.000
0	1	0	3	0	0	2	2	4	0	0	0	0	12	1.000
0	0	0	3	0	0	0	3	4	0	2	0	0	12	1.000
0	1	0	3	1	0	0	0	4	0	0	3	0	12	1.000
0	4	3	0	0	0	3	4	0	2	0	0	0	16	0.999
0	0	2	0	0	0	3	2	3	0	0	0	4	14	0.999
3	4	0	0	0	3	2	3	0	0	0	0	0	15	0.998
3	0	0	2	0	2	3	0	3	0	0	0	0	13	0.998
0	4	0	2	0	3	4	0	2	0	0	0	0	15	0.997
0	0	0	3	0	0	3	3	0	0	0	0	4	13	0.997

Note: Symptom scores of 3 and 4 have been highlighted in order to facilitate comparisons with Table 7.2.

The clearest thing that is revealed by a comparison of Tables 7.2 and 7.3 is that many of those with a high probability of belonging to Class 2 have high scores for anxiety (scores of 3 or 4), while most of those with high probabilities of belonging to Class 1 have mild (but not absent) symptoms of anxiety. Those in Class 2 who *do not* have severe symptoms for anxiety tend to have severe symptoms for at least one other symptom out of worry over physical health, depression or worry. Most of those without high scores on these particular symptoms have 0 scores on that item; there are few individuals in Table 7.3 with mild scores of 1 on the items in question. However, among the individuals from Class 1 in Table 7.2 there are many mild symptoms of worry over physical health, depression, worry and anxiety, but no severe scores.

The preceding examination of response profiles in the region of total symptom scores from 12-16 seems to suggest that the two classes may be distinguished by the presence of relatively mild versus severe or absent symptoms on a few key items. However, the difference between the two classes is still not clear cut. For this reason, it may be helpful to compare the actual parameters of the model in each of the classes to see how they differ. In this particular factor mixture model, the factor loadings were constrained to be equal in both classes, but the intercepts were allowed to differ. Therefore, differences in the estimated intercepts may clarify the differences between the two classes. Since the location of the latent trait is identified in each class by fixing the mean of the latent trait to 0, the intercepts represent the cumulative probabilities (on the logit scale) of each symptom score for individuals at the mean latent trait score for that class; applying the inverse logit transformation to the intercepts will yield the actual probabilities. Figure 7.2 compares the intercepts for the two classes; there are 4 ordered intercepts for each item, corresponding to scores of 1 or higher, 2 or higher, 3 or higher and the score of 4 (in descending order).

It should be noted that one of the intercepts for Class 1 is not shown because it is off the bottom of the plot (this is the intercept corresponding to a score of 4 for anxiety). This is because none of the individuals allocated to Class 1 received a score of 4 for anxiety, resulting in an intercept on the logit scale of minus infinity (or some very large negative number at which the estimation routine stops trying to further maximise the likelihood). On the logit scale, values below approximately -5 correspond to probabilities of less than 0.01, while values below -7 correspond to probabilities of less than 0.001. At intercept values below -10, large differences in the actual value will make virtually no practical difference to the predicted probabilities.

In terms of the *orderings* of the intercepts, Figure 7.2 suggests fairly similar overall patterns of symptoms in the two classes. Nevertheless, the prevalences of depression

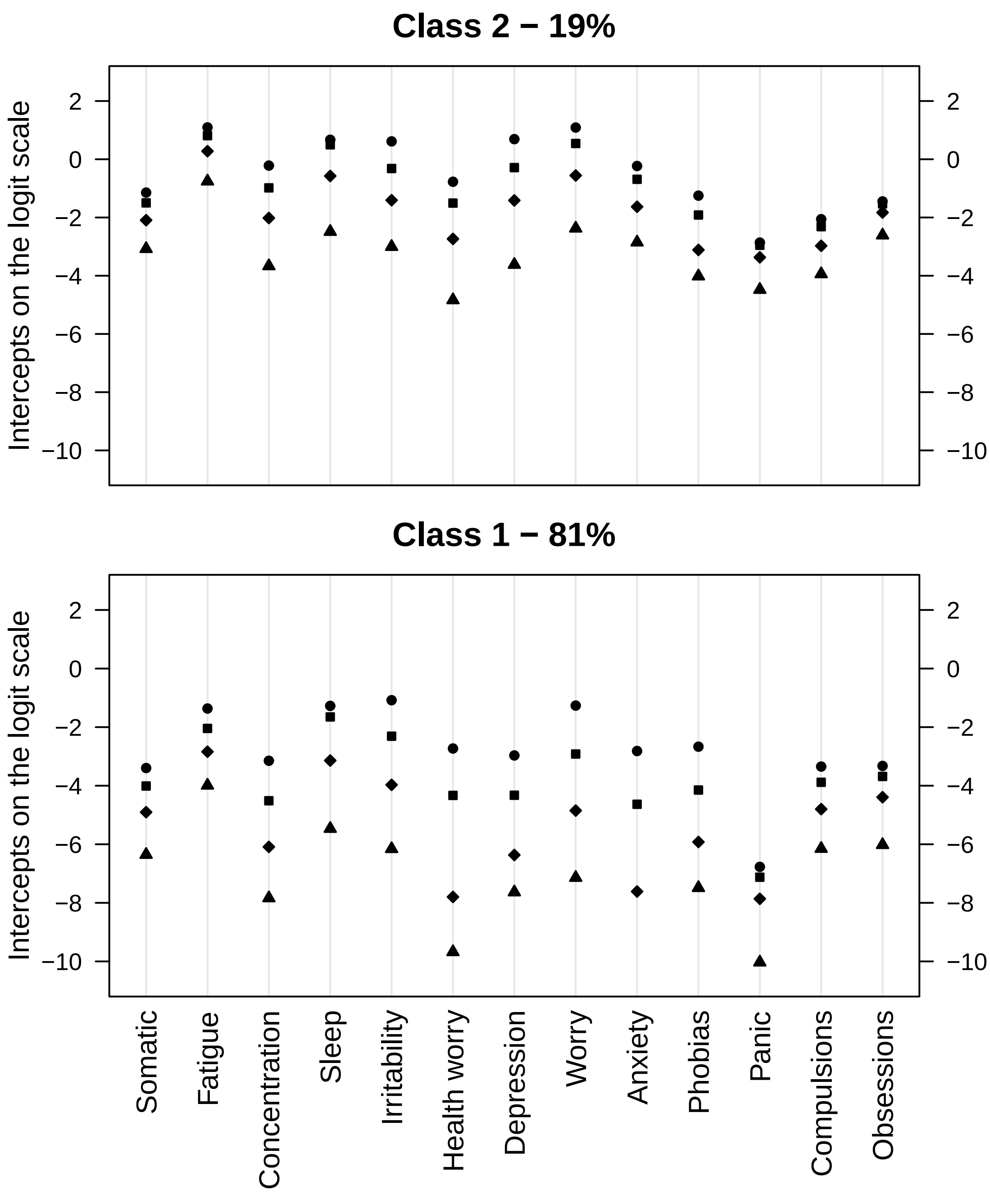

Figure 7.2: Intercept parameters for each of the classes in the 2 class factor mixture model. The mean latent trait score for each class was fixed to 0.

and particularly panic appear somewhat greater relative to other symptoms in Class 2 than in Class 1 (although this may simply reflect differences in overall symptom severity between the mean trait score of Class 2 and the mean trait score of Class 1).

However, when it comes to the *spacings* between the ordered intercepts for each item, the differences between the two classes are more noticeable: the intercepts for many symptoms are spaced much more widely in Class 1 than in Class 2; this occurs most dramatically for the symptoms of health worry, worry, anxiety and phobias, and it is clearest for the spacings between the highest three ordered intercepts. However, there appears to be a similar (if smaller) effect for many of the other symptoms. (These differences between the spacings of intercepts in the two classes are not just quirks resulting from the small number of cases in Class 2 — the same pattern is seen in the second half of the data in Figure E.6 on page 295 of Appendix E.)

While the intercepts shown in Figure 7.2 only correspond to response probabilities for individuals with latent trait values of 0 in each class, the item response functions for all categories of a symptom are constrained to be parallel by the proportional odds formulation of the factor model for ordinal categorical data. (See Figure 6.2 in Section 6.2.1 for an example.) This means that the spacings between the response functions will stay the same within each class for all values of the latent trait. Therefore, the increase in the spacings between the highest three intercepts in Class 1 will correspond to increased prevalences of the mild symptom scores of 1 and 2, relative to the more severe scores of 3 and 4.

It therefore seems possible that Class 1 is distinguished from Class 2 by a relatively high prevalence of mild symptoms, particularly with respect to the symptoms of worry over physical health, depression, worry and anxiety. Conversely, in Class 2, the spacings between the first pair of ordered intercepts for some symptoms are very small; this will correspond to very low prevalences of symptom scores of 1 for those individuals allocated to this class. Members of Class 2 are likely to have moderate/severe symptoms for these items, or no symptoms at all.

The difference between the classes in the 2 class factor mixture model therefore appears to be subtle; it certainly does not correspond to a simple distinction between individuals with a particular mental disorder and those without. However, it is still not clear whether the groups correspond to true subgroups in the population. It is possible that having milder levels of symptoms such as worry and anxiety than might otherwise be expected may have clinical relevance: for example, it could be a feature of individuals who have long-term rather than acute conditions (and who might therefore have adjusted to the presence of certain symptoms over time), or it might reflect symptoms with a different underlying cause (for example, symptoms associated with personality disorders or psychotic disorders). Although such 'post-hoc' explanations are highly speculative, they could be used to generate testable hypotheses (for example, that highly symptomatic individuals in Class 2 will have experienced symptoms for a shorter period of time than highly symptomatic individuals in Class 1, or that they will be more responsive to treatments specifically for anxiety and depression). However, it may be that the classes are in fact playing an indirect rather than a direct role — this possibility will be considered later on in this chapter. Before that, it will be useful to investigate whether any of the latent classes in the 3 class factor mixture model appear to have a direct clinical interpretation.

7.2.2 The 3 class model

While the 2 class model considered in the previous section improved the fit of the model relative to the standard factor model, Table 6.3 indicated that the best fitting of all the replicable models considered in the previous chapter was the factor mixture model with 3 latent classes: this model had the lowest Bayesian information criterion, as well as the smallest bivariate residuals. Therefore, perhaps the classes from the 3 class model will have a more obvious clinical interpretation than those from the 2 class model. Figure 6.17 on page 145 implies that the classes are ordered in terms of symptom severity, although fatigue appears to predominate in Class 1 whereas worry appears to predominate in Class 2. However, as we saw in the previous section, this type of aggregate chart may be obscuring what is actually going on within each of the classes. For this reason, it may be helpful to use the same approach to examine the 3 classes as was used for the 2 class model in the previous section.

The first consideration is whether either Class 1 (the 'mild' class) or Class 2 (the 'intermediate' class) could be considered as a 'no disorder' class. Tables 7.4 and 7.5 show the response profiles and total scores for those with the highest total scores among those allocated to these two classes. (The different numbers of individuals included in tables such as these result from the sorting of these tables by total score: individuals with tied total scores were either *all included* where space allowed, or else *all excluded*. This was done to avoid inadvertently 'cherry picking' individuals for inclusion in the table who appeared to fit with some hypothesised pattern.) Again, some individuals in both classes show high levels of overall symptoms and report key symptoms such as depression, anxiety and panic. Furthermore, more than half of the individuals in the two tables meet the criteria for at least one ICD-10 diagnosis. Again, there is no class that clearly corresponds to a 'no disorder' class.

Table 7.4: The response profiles, probabilities of class membership and ICD-10 diagnoses for the 16 individuals in class 1 (the 'mild' class) with the highest total symptom scores

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total score	Class probability	ICD-10 diagnoses
4	4	4	4	4	1	0	3	2	1	3	4	0	34	0.549	Panic disorder
4	4	4	3	2	1	3	1	3	0	3	0	0	28	0.500	Depressive episode (severe), Panic, GAD
2	4	3	3	3	3	3	1	1	1	0	3	0	27	1.000	Depressive episode (severe)
2	4	3	2	4	1	2	2	1	1	0	3	0	25	1.000	
0	4	2	2	3	1	4	2	1	0	0	4	2	25	0.999	Depressive episode (severe), OCD
0	4	3	3	3	0	4	2	1	1	0	3	0	24	1.000	Depressive episode (severe)
2	4	3	3	2	3	4	2	0	0	0	0	0	23	0.547	Depressive episode (severe)
0	4	3	3	3	2	3	1	0	2	0	0	2	23	1.000	Depressive episode (severe)
3	4	3	4	2	1	1	2	2	0	0	0	0	22	1.000	GAD
0	4	3	4	4	2	2	1	0	2	0	0	0	22	0.735	
4	3	2	3	3	2	2	1	1	0	0	0	1	22	1.000	
3	4	1	4	3	1	0	1	0	3	1	0	0	21	0.999	Specific phobia
3	4	3	4	3	0	0	2	2	0	0	0	0	21	0.498	GAD
4	3	4	4	1	3	1	1	0	0	0	0	0	21	0.717	
4	4	1	3	2	1	2	3	0	0	0	0	1	21	0.475	
3	4	4	2	3	0	2	1	2	0	0	0	0	21	0.429	GAD

Abbreviations: OCD, obsessive compulsive disorder; GAD, generalised anxiety disorder.

Note: The total score is the sum of the 13 symptom scores. Symptom scores of 3 or 4 have been highlighted for easier identification.

Table 7.5: The response profiles, probabilities of class membership and ICD-10 diagnoses for the 11 individuals in class 2 (the 'intermediate' class) with the highest total symptom scores

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total score	Class probability	ICD-10 diagnoses
1	4	3	2	2	2	1	2	1	2	3	4	3	30	0.964	Agoraphobia, OCD
3	3	2	3	2	2	4	2	1	1	3	1	3	30	0.682	Depressive episode (severe), Panic
2	4	3	3	3	1	0	2	2	3	3	3	0	29	0.801	Specific phobia, GAD
0	4	3	2	3	1	1	1	2	2	4	2	3	28	0.803	Agoraphobia, Social phobia, GAD
1	2	2	2	3	0	1	2	2	3	2	4	4	28	0.573	OCD, Social phobia
0	3	4	3	2	1	2	3	2	0	2	1	3	26	0.680	Depressive episode (moderate)
3	4	4	3	3	0	2	2	2	1	2	0	0	26	0.780	Depressive episode (moderate), GAD
2	4	3	3	3	2	1	3	2	0	0	2	0	25	0.564	GAD
3	4	3	2	2	1	2	3	2	0	0	0	3	25	0.561	
2	2	4	2	2	2	2	2	2	1	0	0	3	24	0.874	Depressive episode (moderate)
4	3	1	4	3	2	0	2	2	2	0	0	1	24	0.655	

Abbreviations: OCD, obsessive compulsive disorder; GAD, generalised anxiety disorder.
Note: The total score is the sum of the 13 symptom scores. Symptom scores of 3 or 4 have been highlighted for easier identification.

Table 7.4 suggests strikingly high levels of severe depressive episodes among the high scorers of Class 1: perhaps the highly symptomatic individuals in this class are those with relatively 'pure' fatigue and depression (so not accompanied by high levels or worry, anxiety or panic)? However, closer examination of Table 7.4 suggests that the picture is not nearly so clear: all of the individuals in this table report at least mild levels of worry and often anxiety. Furthermore, several individuals in this group meet the full criteria for an ICD-10 anxiety disorder while having 0 scores for the symptom of depression. Therefore, it doesn't seem realistic to consider this as a 'pure' depression class. Although symptoms such as fatigue, concentration/forgetfulness, sleep problems and irritability appear to dominate Table 7.4, it is not immediately clear what distinguishes these individuals from the high scoring individuals in Class 2 shown in Table 7.5.

As in the previous section, it would be useful to be able to compare the ranges of overall symptom levels in the three classes; therefore, Figure 7.3 shows the distributions of total symptom scores for the three groups. While the mean total score in Class 3 appears to be higher than that of the other two classes, there is again a substantial degree of overlap in the distributions of total scores across the classes. The 'mild' and 'intermediate' severity classes actually cover similar ranges, although Class 1 is dominated by individuals with 0 or very low scores. Again, the numbers of individuals with total scores in the range of 12-16 are fairly similar in each of the three classes, so it would be useful to examine the response profiles of individuals at this level to see what distinguishes the individuals allocated to each class.

Tables 7.6, 7.7 and 7.8 present the response profiles for individuals with total scores in the 12-16 range who have the highest probabilities of belonging to each of the three classes.

- Class 1 in Table 7.6 shows particularly high levels of fatigue (only one of the individuals in the table does not have the highest score for this symptom), as well as high levels of somatic symptoms and sleep problems. There are no scores greater than 1 for the symptoms of worry, anxiety, phobia or panic.

- Class 2 in Table 7.7 shows few of the highest symptom scores, although there are a fair number of 3 scores for fatigue, irritability, compulsions and obsessions. Virtually all individuals in this table show mild or moderate symptoms of worry and anxiety.

- Class 3 in Table 7.8 appears to be dominated by severe symptoms of anxiety. Those individuals who don't have scores of 3 or 4 for anxiety have scores of 0 — there are no mild scores for anxiety in this table. In fact,

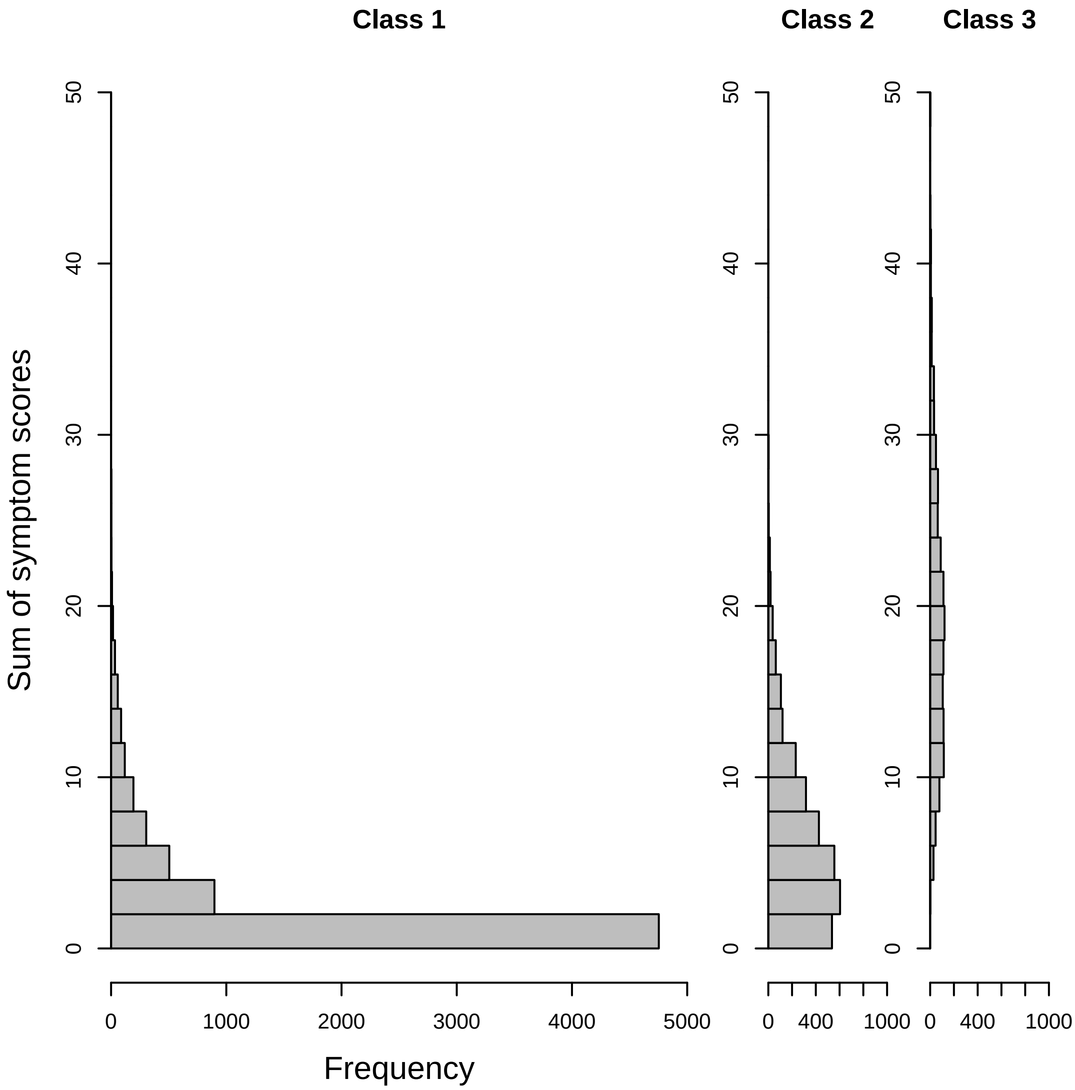

Figure 7.3: Sideways histograms showing the frequencies of total symptom scores in each of the classes of the 3 class factor mixture model — the total symptom scores are a sum of the 13 individual section scores.

Table 7.6: The 30 response profiles with the highest probabilities of belonging to Class 1 among individuals with total symptom scores between 12 and 16

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total	Probability
0	4	2	3	4	1	0	0	1	0	0	0	1	16	0.999
3	4	2	3	0	1	1	0	1	1	0	0	0	16	0.999
1	4	2	2	3	1	0	1	0	0	0	0	0	14	0.998
0	4	2	2	2	0	0	1	0	1	0	1	1	14	0.997
0	4	1	3	0	1	2	1	1	1	0	0	0	14	0.997
2	4	2	2	2	0	0	0	1	0	0	1	0	14	0.997
0	4	2	3	1	1	0	0	1	0	0	0	0	12	0.997
0	4	1	4	0	1	0	0	1	1	0	0	2	14	0.996
0	4	1	4	2	1	0	1	0	0	0	0	0	13	0.996
3	4	0	3	1	1	1	1	0	1	0	1	0	16	0.993
0	4	3	2	0	0	1	1	1	1	0	1	0	14	0.993
0	4	1	2	2	1	1	0	0	0	0	1	0	12	0.993
0	4	2	3	3	2	0	0	1	0	0	0	0	15	0.992
3	4	1	2	1	1	0	0	0	0	0	2	0	14	0.991
3	4	1	2	1	1	1	0	0	0	0	0	0	13	0.991
4	4	1	3	1	0	0	0	0	0	0	0	0	13	0.990
2	4	2	3	0	0	0	1	0	0	0	1	0	13	0.990
0	4	0	4	2	1	1	1	1	1	0	0	0	15	0.989
0	4	2	2	0	1	2	1	0	1	0	0	0	13	0.989
0	4	2	2	1	1	2	1	0	0	0	0	0	13	0.988
0	4	2	3	2	0	0	1	1	0	0	0	0	13	0.988
4	4	0	3	0	2	0	1	1	1	0	0	0	16	0.986
3	4	1	3	2	0	1	0	1	0	0	0	0	15	0.980
4	4	2	0	0	2	0	0	0	0	0	0	0	12	0.978
3	3	2	3	2	1	1	0	1	0	0	0	0	16	0.977
0	4	2	2	0	1	2	1	0	0	0	0	0	12	0.977
3	4	1	2	3	0	0	1	0	0	0	0	0	14	0.974
0	4	3	0	1	1	4	1	0	0	0	0	0	14	0.967
0	4	3	3	0	1	1	0	0	0	0	0	0	12	0.965
4	4	0	3	2	0	0	0	1	0	0	0	0	14	0.964

Note: Symptom scores of 3 and 4 have been highlighted in order to facilitate comparisons with Table 7.7 and 7.8.

Table 7.7: The 30 response profiles with the highest probabilities of belonging to Class 2 among individuals with total symptom scores between 12 and 16

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total	Probability
1	1	0	0	3	1	0	2	1	1	0	2	2	14	0.997
0	0	1	1	2	2	1	1	1	2	0	0	1	12	0.991
0	3	3	0	1	2	0	1	1	2	2	1	0	16	0.982
2	3	0	1	2	1	0	2	1	0	0	0	3	15	0.982
0	1	1	2	3	0	1	2	1	1	0	0	3	15	0.978
0	2	0	2	0	1	1	2	1	1	0	1	2	13	0.978
0	0	0	1	2	2	0	2	1	1	0	3	0	12	0.978
0	3	0	2	3	0	0	2	1	0	1	0	3	15	0.977
0	0	0	2	2	1	0	1	0	1	1	0	4	12	0.977
0	3	1	1	1	1	1	0	1	2	0	3	1	15	0.974
2	3	0	0	2	0	1	1	1	0	0	3	0	13	0.974
0	2	2	2	1	1	1	1	2	3	0	0	1	16	0.973
0	3	0	2	2	0	1	2	1	1	0	2	0	14	0.970
0	3	0	2	1	0	0	2	1	1	0	3	0	13	0.970
1	2	1	0	1	1	0	2	0	0	0	2	3	13	0.966
0	1	1	3	1	0	2	1	1	0	2	0	0	12	0.966
0	1	1	2	3	2	0	2	2	0	1	0	0	14	0.965
0	3	1	2	2	2	1	2	1	0	0	0	2	16	0.964
2	1	0	2	2	0	1	2	1	0	0	0	3	14	0.962
1	4	0	0	2	0	0	1	1	2	0	3	0	14	0.962
0	2	1	2	2	0	0	2	1	0	0	0	2	12	0.960
2	2	2	0	3	0	0	1	1	1	0	0	2	14	0.959
0	3	1	2	1	0	0	2	1	0	0	0	3	13	0.957
0	2	2	0	1	0	2	2	1	1	2	0	2	15	0.956
1	0	0	2	2	0	1	1	2	0	0	1	3	13	0.955
0	3	1	0	1	0	1	1	2	0	0	2	3	14	0.954
2	3	2	1	1	2	1	2	1	1	0	0	0	16	0.952
0	3	0	2	0	2	0	1	1	2	0	3	0	14	0.952
1	2	4	1	2	0	0	2	2	1	0	0	0	15	0.948
0	3	2	2	2	0	1	2	2	0	1	0	0	15	0.947

Note: Symptom scores of 3 and 4 have been highlighted in order to facilitate comparisons with Table 7.6 and 7.8.

Table 7.8: The 30 response profiles with the highest probabilities of belonging to Class 3 among individuals with total symptom scores between 12 and 16

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total	Probability
0	0	0	3	1	0	0	4	4	0	0	0	4	16	1.000
0	0	2	3	2	0	0	4	4	1	0	0	0	16	1.000
0	4	1	0	2	0	2	3	4	0	0	0	0	16	1.000
0	0	0	2	3	0	1	3	4	2	0	0	1	16	1.000
0	4	0	2	2	0	0	3	4	0	0	0	0	15	1.000
0	4	0	2	2	0	2	0	4	0	0	1	0	15	1.000
0	1	2	0	2	0	3	3	4	0	0	0	0	15	1.000
4	0	0	0	0	2	0	3	4	2	0	0	0	15	1.000
0	0	0	2	0	2	3	0	4	0	0	0	3	14	1.000
0	2	0	3	2	0	1	2	4	0	0	0	0	14	1.000
0	0	0	0	4	0	2	4	4	0	0	0	0	14	1.000
0	0	0	2	0	4	1	2	3	2	0	0	0	14	1.000
0	3	0	3	1	0	0	3	4	0	0	0	0	14	1.000
0	2	1	3	2	0	1	1	4	0	0	0	0	14	1.000
0	0	1	2	3	1	0	3	4	0	0	0	0	14	1.000
0	2	1	0	1	0	2	0	4	0	3	0	0	13	1.000
0	0	2	3	0	0	1	3	4	0	0	0	0	13	1.000
1	1	0	2	2	0	0	3	4	0	0	0	0	13	1.000
0	2	1	0	0	0	0	2	4	2	2	0	0	13	1.000
0	1	0	3	0	0	2	2	4	0	0	0	0	12	1.000
0	1	0	3	1	0	0	0	4	0	0	3	0	12	1.000
0	0	0	3	0	0	0	4	4	1	0	0	0	12	1.000
0	0	0	3	0	0	0	3	4	0	2	0	0	12	1.000
0	4	3	0	0	0	3	4	0	2	0	0	0	16	0.999
0	0	2	0	0	0	3	2	3	0	0	0	4	14	0.999
0	0	0	3	0	0	3	3	0	0	0	0	4	13	0.998
0	0	2	3	0	0	3	4	0	0	0	0	0	12	0.998
2	2	0	0	0	3	3	2	3	0	0	0	0	15	0.997
0	0	1	4	0	0	3	3	0	0	0	0	3	14	0.997
0	0	0	4	0	0	4	2	3	0	0	0	0	13	0.995

Note: Symptom scores of 3 and 4 have been highlighted in order to facilitate comparisons with Table 7.6 and 7.7.

there are few mild scores on most of the symptoms apart from fatigue, concentration/forgetfulness, irritability and depression.

In general, Class 3 appears to be fairly similar to the small class from the 2 class factor mixture model described in the previous section: in both cases, individuals with total symptom scores of 12-16 and high probabilities of belonging to these classes have few mild symptoms, particularly for health worry, worry and anxiety. Classes 1 and 2 appear to divide up those who do show such mild symptoms; in particular, Class 1 appears to accommodate those whose somatic symptoms, fatigue and sleep symptoms exceed their other symptoms. In Class 2, all but one of the individuals included in the table have a mild score of 1 or 2 for worry, and all but two have mild scores for anxiety.

To complete the examination of the three class model, Figure 7.4 shows the intercept parameters for each class; again, the intercepts correspond to response probabilities for individuals with the mean latent trait score in each class, and the intercepts for the most severe scores of some symptoms are missed off the bottom of the bottom two panels where their corresponding probabilities are extremely close to 0. (The corresponding plot for the second split half of the data is shown in Figure E.8 on page 297 of Appendix E.)

- In terms of the relative *orderings* of the intercepts, the orderings in Class 1 appear to be roughly the same as those for Class 1 in the 2 class model shown in Figure 7.2 on page 159, while the orderings for Class 3 appear to correspond to Class 2 in the 2 class model. The orderings of the intercepts in the middle class could perhaps be seen as intermediate between the positions in the two outer classes. For Class 1, the symptoms of fatigue, sleep and irritability have the highest expected prevalence at the mean trait score, whilst in the other two classes the prominence of these symptoms decreases relative to worry and anxiety.

- In terms of the *spacings* between the intercepts, Class 2 shows the widest spacings, although the spacings between the higher pairs of intercept parameters in Class 1 are also noticeably wider for many symptoms than the intercepts in Class 3. Therefore, it seems that there is likely to be the same relative excess of mild symptoms in Classes 1 and 2 compared to Class 3 (particularly of health worry, worry and anxiety) as seen in the previous section.

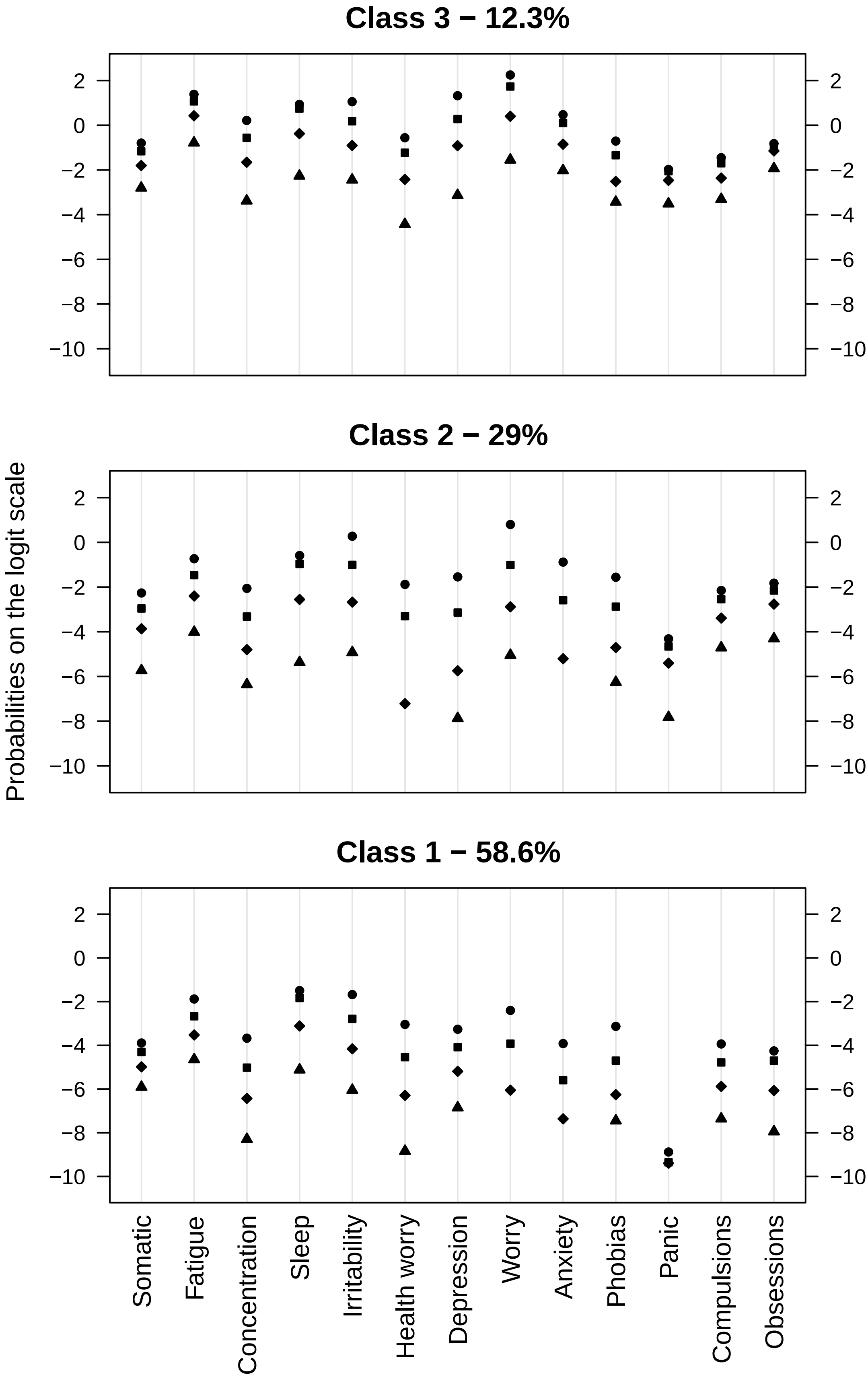

Figure 7.4: Intercept parameters for each of the classes in the 3 class factor mixture model. The mean latent trait score in each class is fixed at 0.

7.2.3 No easily interpretable classes

To sum up, neither the 2 or 3 class factor mixture models appear to show easily interpretable latent classes. In both models, the distinctions between the classes appear to be subtle. For the 2 class model, the larger class seems to allow for a greater prevalence of mild scores on symptoms such as worry and anxiety. A similar effect seems to be present in the larger two classes of the 3 class model, while there are also some modest differences in the relative prevalences of some of the symptoms. While the latent classes do not appear to represent easily interpretable 'disorder classes', they may still reflect real groups in the population. As suggested earlier, it is possible that highly symptomatic individuals with relatively mild levels of symptoms such as worry and anxiety may have conditions with a different longitudinal course or aetiology from those who tend to have a symptom severely or not at all. Alternatively, these mild symptoms in the larger classes might reflect relatively stable traits rather than symptoms arising from an acute episode of illness. Such speculations could be used to develop testable hypotheses (for example, that highly symptomatic individuals in Class 1 and Class 2 show much greater temporal stability in their symptoms than those in Class 3).

However, when it comes to trying to interpret the meanings of latent classes generated through exploratory analysis, it is relatively easy to generate plausible 'post-hoc' explanations for the groups that classes may represent. As such, it would be dangerous to accept such interpretations uncritically, particularly when it is possible that the latent classes may not in fact reflect the presence of real groups at all. The next section will therefore look for any evidence that the latent classes are actually performing some kind of indirect role.

7.3 Indirect roles for the classes?

Bauer & Curran (2004) identify three different scenarios that may result in the extraction of spurious latent classes when using hybrid models that combine continuous and discrete latent variables. They are paraphrased here as follows:

1. Non-normality of the latent trait distribution(s)
2. Misspecification of the model structure
3. Non-linear relationships between observed items and/or latent variables that cannot be accommodated by the chosen within-class model.

Each one of these scenarios might result in a factor mixture model being selected as the 'optimally fitting model' in the absence of true groups; the classes would then be playing an indirect rather than a direct role. The following paragraphs will describe how each of these scenarios could apply to the models being used in this project to analyse the structure of common mental disorders.

The first scenario is relatively straightforward, and refers to the situation where the latent trait in a standard factor model has a non-normal distribution (which could be similar to the situation illustrated in Figure 5.1 on page 85). Under this scenario, a model in which the latent trait is described by a mixture of normal distributions may fit the data better than a model with a single normal distribution for the latent trait, particularly if the actual distribution of the trait is highly skewed. However, this does not imply that the population is actually composed of two or more groups — the classes here are playing an indirect role.

The second scenario would refer to some form of misspecification of the measurement structure in the standard factor model. Some forms that this misspecification might take include: specifying the factor structure to be unidimensional when there are in fact two or more factors; fixing factor loadings for particular items to be zero inappropriately; or inappropriately constraining factors to be uncorrelated. In each of these situations, two or more latent classes may be able to account for the misspecification; an example of this is illustrated in Figure 6.8 on page 128, where 5 latent classes with uncorrelated factors are used to approximate the missing correlation between the two factors. Again, these classes are playing an indirect role, and it would be misleading to interpret them as if they represent true underlying groups in the population.

With regard to the final scenario of non-linearity, the situation is slightly different for unidimensional models estimated with binary or ordinal data than for the models estimated on continuous data that were considered by Bauer and Curran. For binary or ordinal data, the relevant consideration will be whether or not the logistic function is appropriate to describe the relationships between symptom probabilities and scores on the latent trait, rather than whether the relationship is linear. If the S-shaped logistic response function is inappropriate, this will result in a violation of the linearity assumption *on the logit scale*, the scale in which the factor model parameters themselves are actually estimated. (As was illustrated in Figure 6.1 on page 108 and Figure 6.2 on page 109, the factor model describes a *linear* relationship between the latent trait and the probability of response *on the logit scale*, but this implies the characteristic S-shaped logistic curve for the actual probabilities when the inverse logit transformation is applied.) If the latent classes were simply accommodating the inappropriateness of the logistic function, it would again be misleading to interpret them as if they represented true groups in the population.

(It may be noted that both non-normality and inappropriateness of the logistic function are themselves forms of model misspecification, as in the second scenario described above — as such, the division into three scenarios may seem artificial. Nonetheless, it will be useful to maintain this distinction, since the means to identify and/or remedy each of the three types of misspecification are different, as will be seen later in this chapter.)

In some research settings, a researcher may consciously aim to use latent classes in one of their indirect roles when they wish to relax one of the assumptions of the standard factor model. However, many applications in the field of psychiatry are likely to be searching for hidden groups in the population; indeed, a goal of the present thesis is to see whether latent classes can identify data-driven 'disorder' groups on the basis of individuals' symptoms of anxiety and depression. Where the goal of a researcher is to identify latent classes playing this kind of direct role, Bauer and Curran suggest that the burden of proof rests on the researcher to demonstrate that the classes represent true groups in the population, and are not just accommodating some form of model misspecification. While it may not be possible to prove this conclusively, a researcher's interpretation of the classes as representing true groups may be given more credence if they can show that they have attempted to rule out an indirect role for the latent classes resulting from one or more of the scenarios described above.

For this reason, the rest of this chapter will go through each of the three types of misspecification in turn, in each case looking for evidence that the classes from the factor mixture models may in fact be playing the indirect role of accommodating this type of misspecification. Unfortunately, Bauer and Curran gave little indication of how a researcher with ordinal observed data might go about investigating any of these three types of misspecification (the few specific suggestions they do make apply only to models with continuous observed data). Furthermore, as mentioned in the introduction to this chapter, none of the applications of factor mixture models that have been published so far have made any systematic attempt to investigate whether the classes in a factor mixture model may be playing an indirect role. The rest of this chapter is therefore moving into uncharted territory — the approaches presented are the best that the author was able to devise for investigating each type of indirect role. Future work may well identify easier or more effective methods for carrying out such an investigation.

7.3.1 Indirect role 1: Non-normality

Non-normality of the latent trait is actually the easiest form of model misspecification to investigate within the factor mixture model framework. This is because some particular specifications of the factor mixture model are perhaps best suited to modelling non-normality of the latent trait, rather than looking for classes playing a direct role. This was discussed in Sections 6.3.1 and 6.3.2 of the previous chapter, where these particular model specifications were described as 'semi-parametric factor analysis' and 'latent class factor analysis'. These models impose strong measurement invariance on the parameters of the class-specific factor models, i.e. intercepts and loadings are constrained to be the same in all classes. This means that these models involve the estimation of far fewer parameters than the factor mixture models described at the start of this chapter.

As a result, if the classes in the factor model are simply accommodating non-normality, the more parsimonious semi-parametric and latent class factor models should be able to demonstrate the same improvement in bivariate residuals that was seen for the factor mixture models. (This would imply that the observed improvement in bivariate residuals for the factor mixture models relative to the standard factor model was due solely to the ability of these models to relax the assumption of a normally distributed latent trait.) Furthermore, fit indices that penalise the model log-likelihood according to the number of estimated parameters (such as the Bayesian information criterion) would be expected to favour the relatively parsimonious semi-parametric and latent class factor models over the more complex factor mixture models.

The fit statistics for these models were reported in Table 6.1 on page 130. This table makes it clear that the semi-parametric and latent class factor models do not yield any improvement in their bivariate residuals over the standard factor model, and their Bayesian information criteria are worse than those for the both standard model and the factor mixture models (whose fit statistics were reported in Table 6.3 on page 142). Furthermore, it was also shown in the previous chapter that the estimated factor scores and factor loadings for these models were extremely similar to those from the standard factor model, giving little cause for concern about bias in the parameters resulting from the assumption of a normally distributed latent trait. These results suggest that the classes in the factor mixture model must be playing some other role than accounting for non-normality in the latent trait. The next section will consider whether the classes may be accounting for some form of misspecification in the measurement model.

7.3.2 Indirect role 2: Misspecification of the measurement model

All of the mixture models that were described in the previous section use a single dimension or latent trait to account for the correlations between symptoms within each class. This seemed a reasonable choice, since the examination of eigenvalues of the sample correlation matrix in Table 4.2 on page 71 suggested the presence of one dominant factor in the data. Given that all 13 symptoms are allowed to load on this factor, it is not possible for any of the factor loadings to be inappropriately constrained to zero. Moreover, since there is only one dimension, there can be no difficulty arising from the factors being inappropriately specified as uncorrelated. However, the modified parallel analysis reported in Table 4.3 on page 73 did suggest the presence of at least one other minor factor, and Section 4.3.3 reported that factor models with up to three dimensions produced results that were stable in both random halves of the data. Since one of the assumptions of the unidimensional factor model is that symptoms are uncorrelated given individuals' scores on the latent trait (the assumption of conditional independence), the existence of these minor factors represents a violation of this assumption, even if these factors represent only relatively unimportant residual correlations between a few items rather than substantively important dimensions of mental health. It is therefore possible that the classes in the factor mixture model are simply accommodating this misspecification.

One way to test whether the classes in the factor mixture model are actually playing this role would be to fit the factor mixture models incorporating two or three factors to model the correlations between symptoms within each class. If the same latent classes appear in factor mixture models with two or three dimensions as in the factor mixture models with a single dimension, this would suggest that the classes are not just accommodating the presence of these minor factors. However, if the classes are completely different or if models with two or three classes are not identified, this may provide evidence that the classes are in fact accommodating the misspecification of the factor structure.

Given that inappropriately fixing factor loadings to zero could make any model misspecification worse rather than better, these additional dimensions were incorporated into the factor mixture analysis in the most flexible way possible: the within-class factor models were specified using the 'exploratory factor analysis within the confirmatory factor analysis' framework (E/CFA) described in section 4.3.2 on page 76, which imposes the minimum number of constraints necessary to identify the model. For these models, the same across-class constraints were imposed as in the factor mixture models described in the first half of this chapter, i.e. factor loadings were constrained to be equal across classes, while intercepts were allowed to vary. The scales of the three latent variables were fixed by constraining the loadings for the anchor item on each factor (fatigue, worry and phobias — see Section 4.3.2) to be equal to 1, while the mean of each latent trait was fixed at 0 in each class. Factor variances and covariances were freely estimated and allowed to differ between classes.

The fit statistics for factor mixture models with one, two and three dimensions are shown in Table 7.9, along with the fit statistics of the standard factor models with one, two and three dimensions for comparison. (See Section 3.2.7 on page 56 for details of the various fit statistics that can be used for comparing models.) The Bayesian information criterion favours the model with three dimensions and two classes. However, this criterion cannot tell us what role the classes are actually playing or how much difference the additional dimensions make to the classes of the factor mixture models. Similarly, the overall bivariate Pearson chi-square and bivariate residuals indicate that the best fitting model is the one with three dimensions and three classes. However, given the large number of parameters in this model (206), it is not really surprising that this model provides the closest match to the observed data.

Table 7.9: Model comparison table for factor mixture models with 1, 2 and 3 dimensions — in all multi-class models factor loadings are constrained to be equal but intercepts are allowed to vary across classes

Model	# par	LL	BIC	Relative entropy	LMR aLRT statistic	LMR aLRT p value	Smallest class size (proportion)	Validates in split half	Overall bivariate Pearson chi-square	Pairs with significant* lack of fit (out of 78)	Bivariate standardised residuals > 3 (out of 1950)
1 dimensional models											
1f 1c	65	-90,962	182,531	-	-	-	-	Y	3,974	35	95
1f 2c	119	-90,453	182,015	0.66	1,017	0.0778	2,134 (0.19)	Y	2,135	8	17
1f 3c	173	-90,141	181,895	0.58	622	0.0251	1,385 (0.12)	Y	1,667	4	9
2 dimensional models											
2f 1c	77	-90,630	181,979	-	-	-	-	Y	3,588	34	69
2f 2c	133	-90,147	181,535	0.67	964	0.0000	2,098 (0.19)	Y	1,777	4	8
2f 3c	189	-89,984	181,730	0.55	326	0.0081	1,525 (0.14)	N	1,367	2	3
3 dimensional models											
3f 1c	88	-90,473	181,767	-	-	-	-	Y	3,258	27	55
3f 2c	147	-90,003	**181,378**	0.68	937	0.0000	1,971 (0.18)	Y	1,507	1	1
3f 3c	206	-89,897	181,715	0.49	213	0.0620	1,590 (0.14)	N	1,245	0	1

Abbreviations: # par, number of parameters estimated in the model; LL, log-likelihood; BIC, Bayesian information criterion; LMR aLRT, Lo-Mendell-Rubin adjusted likelihood ratio test; f, factor; c, class.
All models fitted to the same random half of the data. n = 11,230. *P value used for significance cut-off = 0.01.

What is of more interest is whether additional latent classes provide similar improvements in fit for models with two or three dimensions as in the unidimensional models. In each case, the bivariate fit statistics show a considerable improvement with the addition of the second class, along with a more modest improvement with the addition of the third class (although this improvement becomes smaller as the number of dimensions increases). Similarly, the Lo-Mendell-Rubin adjusted likelihood ratio tests for k versus $k-1$ classes support the need for the second and third classes in the models with two dimensions, while in the models with three dimensions they support the need for two classes but are less clear about the third.

The pattern of smallest class sizes is relatively stable across the models with one, two and three dimensions. Moreover, the make up of the classes appears to be fairly stable, even though there are differences in the estimated prevalences of the classes. This can be seen from the fact that the estimated proportions of scores endorsed for each symptom in the corresponding classes appear remarkably similar; these are shown for the three class model with *two* dimensions in Figure 7.5 and for the three class model with *three* dimensions in Figure 7.6. (These figures can be compared with the estimated score proportions from the corresponding unidimensional factor mixture model in Figure 6.17 on page 145; similarly stable estimated score proportions occur for the two class models, but those figures are omitted here to avoid too much repetition.) It appears that roughly the same classes are extracted however many dimensions are included.

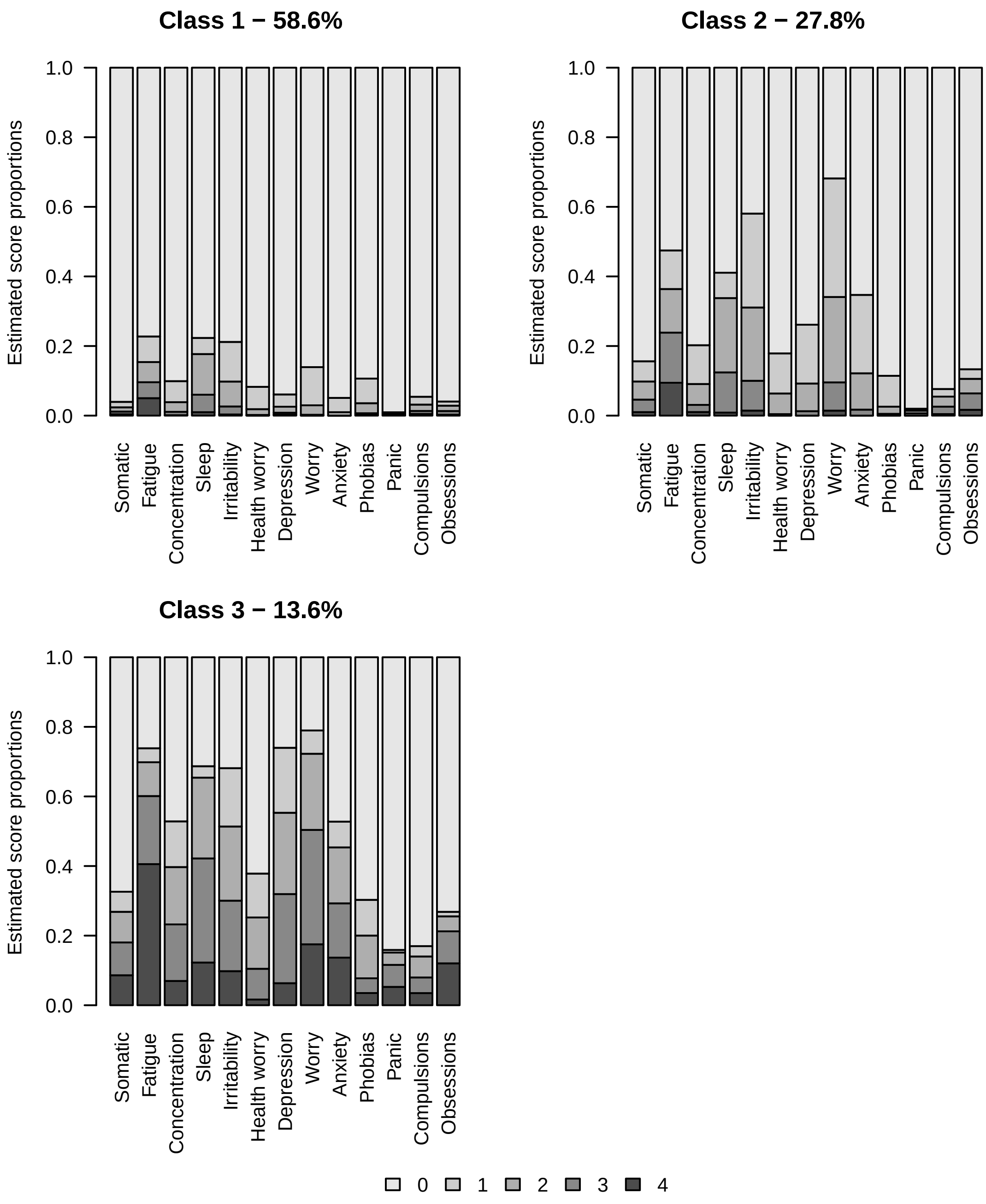

Figure 7.5: Estimated score proportions in each class of the factor mixture model with 2 dimensions and 3 classes

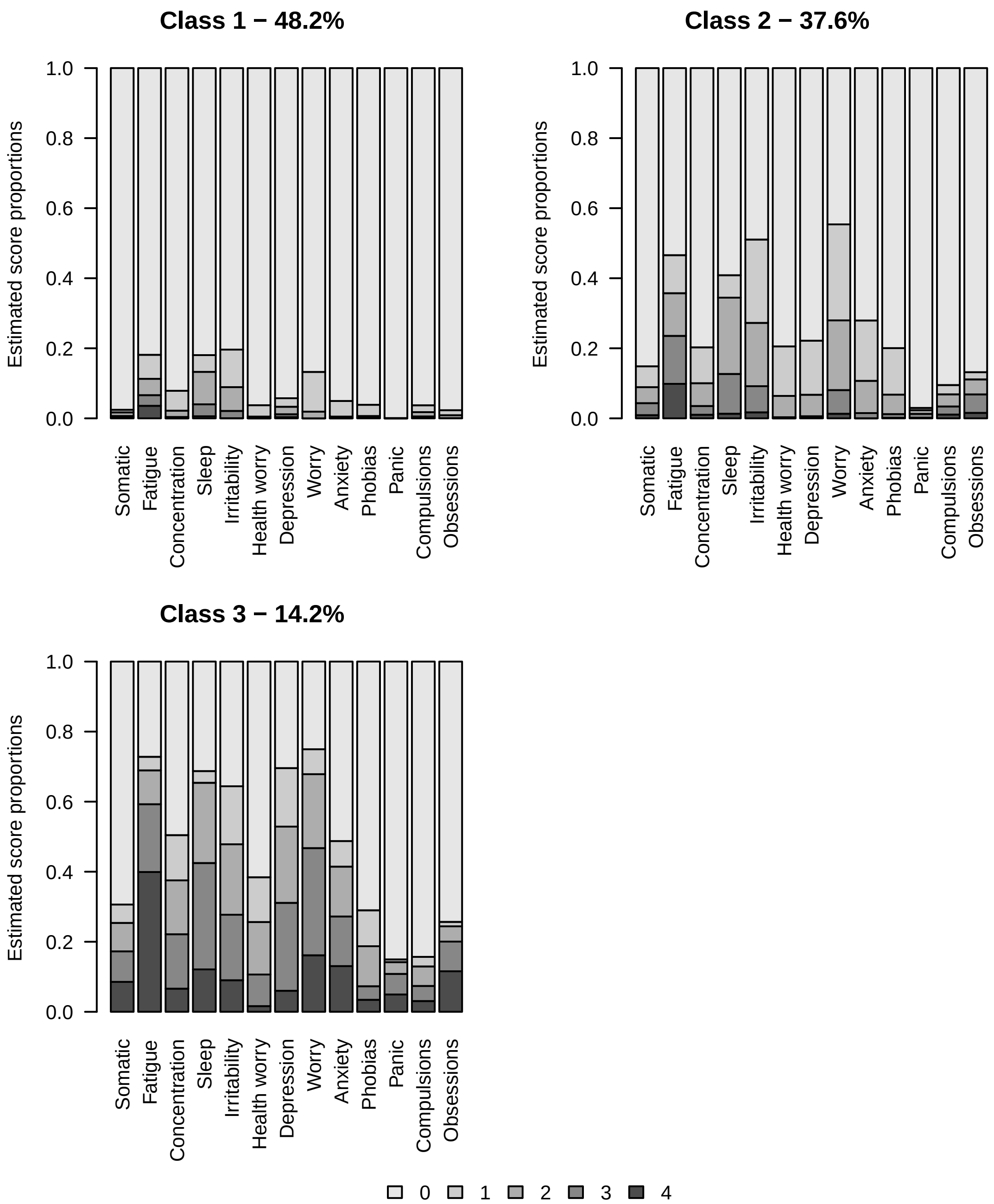

Figure 7.6: Estimated score proportions in each class of the factor mixture model with 3 dimensions and 3 classes.

In addition to looking at summaries of the score proportions in each class, it may also be useful to examine the allocations of individuals to classes in each of the models, and to see how stable these are with the additions of the second and third dimensions. The scatterplot matrix in Figure 7.7 compares the probabilities of allocation to the small 'severe' class for every individual under the models with one dimension, two dimensions and three dimensions. The class membership probabilities appear remarkably similar in each of the three models, so the two class model appears very stable.

Stability is not so good for the three class models, particularly for the middle 'intermediate' class. A scatterplot matrix comparing class membership probabilities for the middle class in each of the three class models is shown in Figure 7.8. Some individuals show quite large changes in their probabilities of belonging to a particular class as the number of dimensions changes, although a substantial core of individuals still receive similar class membership probabilities in each pairwise comparison. Furthermore, the models with three classes and two/three dimensions do not replicate well in the second split half of the data — the 'intermediate' classes become much less prevalent and include more severe symptoms, particularly for fatigue (see the bar charts for Figure E.10 on page 302 and Figure E.11). These differences between the stability of the two and three class models may indicate that the three class model is more sensitive to the existence of the 'minor factors' than the two class model. Nonetheless, given the similarities between the extracted classes in the one-dimensional, two-dimensional and three-dimensional three class models in the first half of the data, it does not seem as if these 'minor factors' are driving the need for latent classes. Therefore, it appears that we can discount accommodating misspecification of the factor structure as the major role of the classes in the factor mixture models.

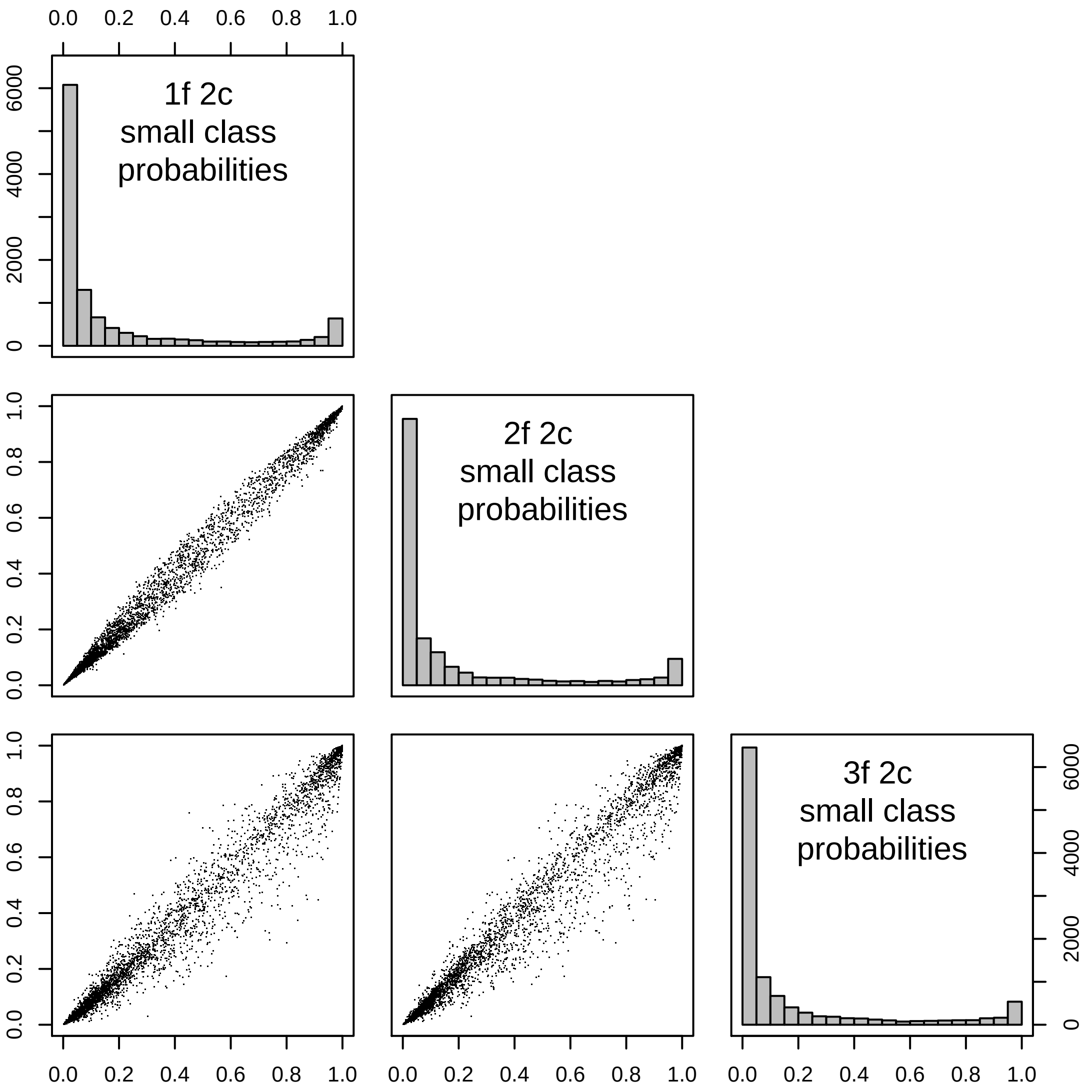

Figure 7.7: Scatterplot matrix showing individuals' class allocation probabilities for the small 'highly symptomatic' class in the 2 class factor mixture models with one dimension, two dimensions and three dimensions

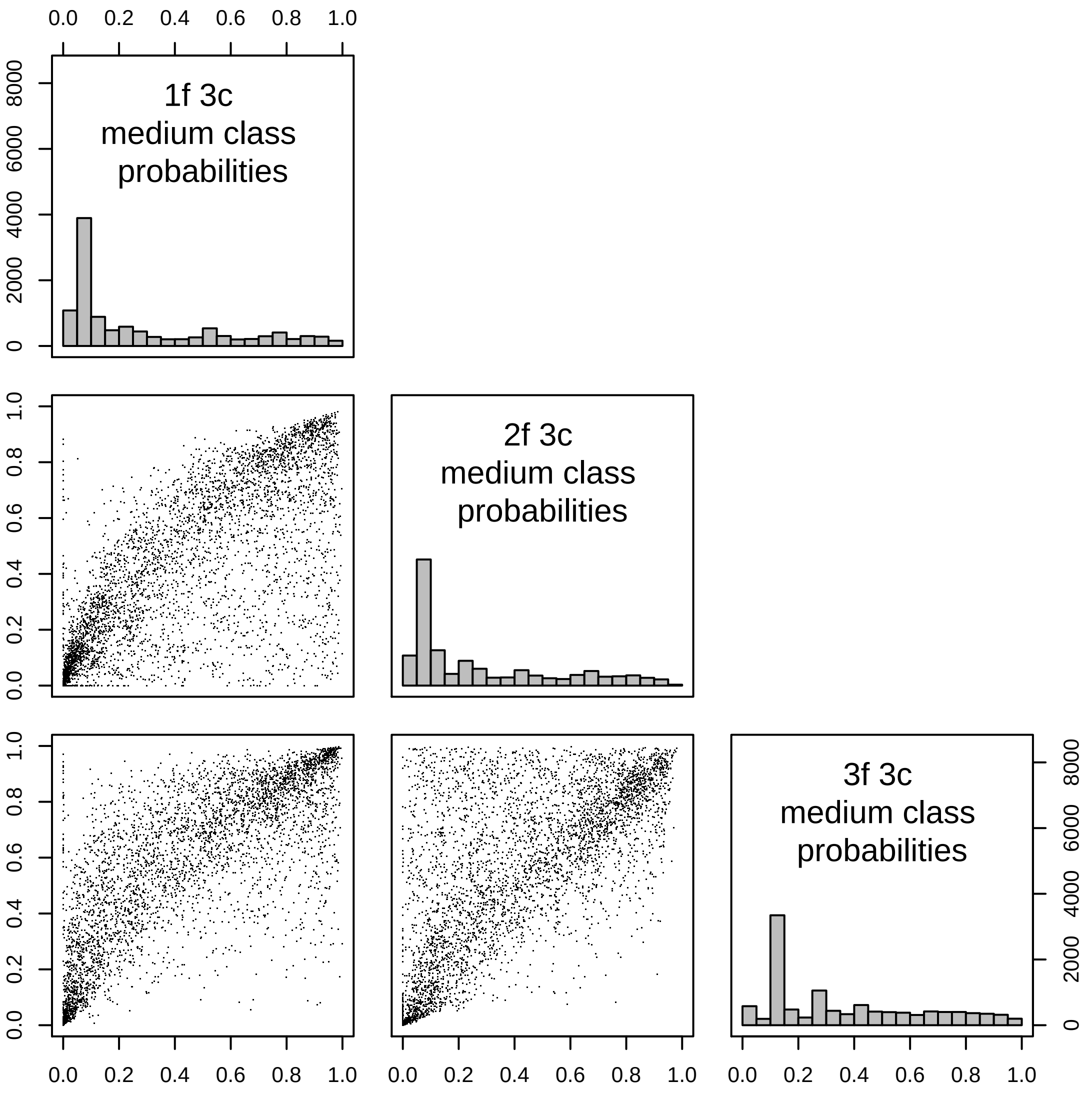

Figure 7.8: Scatterplot matrix showing individuals' class allocation probabilities for the middle 'intermediate' class in the 3 class factor mixture models with one dimension, two dimensions and three dimensions

7.3.3 Indirect role 3: When the logistic function is inappropriate

The third scenario in which a model with latent classes may appear to describe the data better than a standard factor model *in the absence of true groups* is when the logistic function is not appropriate for describing the relationship between the latent trait and the probability of experiencing a symptom. (As described earlier in this chapter, the implication of this is that the relationship is non-linear on the logit scale.)

The logistic function assumed by the ordinal factor model has a characteristic S-shaped curve that is illustrated in Figure 7.9; it possesses a number of key properties:

1. The upper asymptote of the curve occurs at a y value of 1. (The upper asymptote is where the top half of the curve flattens out and runs horizontally to the right.)

2. The lower asymptote of the curve occurs at a y value of 0. (This is the corresponding feature where the lower half of the curve flattens out and runs horizontally to the left.)

3. The curve has rotational symmetry around its inflection point. (The inflection point is the point at which the slope of the curve switches from becoming increasingly steep to becoming increasingly shallow. For the curves shown in Figure 7.9 the inflection points occur at probabilities of 0.5 and are marked by dots). Roughly speaking, this rotational symmetry means that the curve should approach the upper asymptote of 1 in the same fashion as it departs from the lower asymptote of 0; this means that the curve should not rise steeply in the lower half of the plot but then flatten off very gradually in the upper half of the plot (and vice versa).

4. Although this is not a requirement for the use of the logistic function *per se*, in the ordinal factor model the logistic curves are assumed to be the *same shape* for all response categories of a particular symptom. This is illustrated in Figure 7.9: the curves for higher categories are simply the logistic curve for the first category shifted to the right.

These characteristics of the logistic function suggest a corresponding set of requirements that should be present in (or at least not contradicted by) the symptom data, if the standard factor model is to fit well:

1. *The upper asymptote should be at 1:* A hypothetical individual who is at the

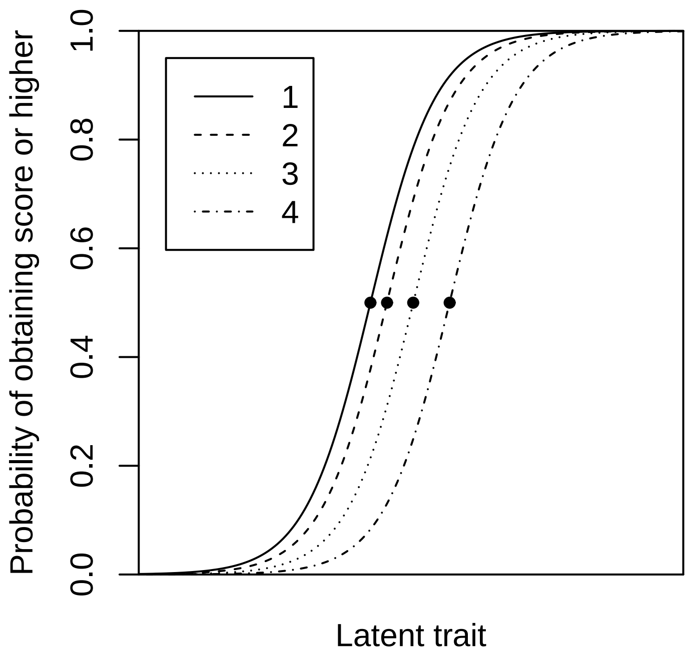

Figure 7.9: Illustration of the shape of the logistic curves assumed in the factor model for ordinal data. The dots indicate the point of inflection for each curve.

very highest point on the latent trait (and who therefore has the most severe symptoms possible) should be certain to have *every* symptom. Regardless of whether there actually are any individuals this sick in the sample, it should be *possible* for an individual to be this sick. Importantly, there should be no key symptoms that are frequently *notable by their absence* in very sick individuals.

2. *The lower asymptote should be at 0:* A hypothetical individual who is at the lowest point on the latent trait (and is therefore the most well/least distressed possible) should be certain to have *none* of the symptoms. This means that there should be no symptoms that occur at background levels quite independently of an individual's position on the latent trait.

3. *The curve should have rotational symmetry:* Symptoms should be equally discriminating at all levels of the latent trait — this is because the factor loading/discrimination parameter is actually estimated on the logit scale, and is therefore assumed to be constant at all values of the latent trait. This means there should be no symptoms that are very good at discriminating between those with mild and moderate trait severity, but that are poor at discriminating between those with moderate and severe trait severity (and vice versa): if one of these patterns occurs, the curve will not show rotational symmetry.

4. *The curves for ordered categories should be the same shape:* Mild, moderate and severe scores on a particular symptom should be equally discriminating, since their factor loadings are assumed to be the same. (Nonetheless, mild levels of a symptom may still be more *useful* for distinguishing between those who are in the lower part the latent trait, while severe levels of a symptom may be more useful for distinguishing between those who are in the upper part of the latent trait; for example, if *every individual* towards the upper end of the latent trait reports at least mild levels of a particular symptom, then mild levels of that symptom can tell you nothing about differences in trait scores between those individuals).

Where any of these requirements are contradicted by the data, additional latent classes may be extracted to accommodate the fact that the assumed logistic function is not appropriate for the data. Unfortunately, it is very difficult to test whether the classes in the factor mixture models are actually playing this indirect role. This is because there is no straightforward way to fit a factor model that relaxes the assumptions implied by use of the logistic function about the shape of the relationship between symptoms and the latent trait, and against which the fit of the mixture model might be compared. However, it may still be useful to examine whether any of the above requirements are unlikely to hold in the data. Therefore, the next section will use a non-parametric approach to investigate the true shape of the relationship.

Exploring the true shape of the relationship: non-parametric analysis

Non-parametric approaches attempt to describe the true shape of the relationship between two variables *without making any assumptions* about the underlying nature of this relationship. The approach to be used here is based on local logistic regression (as used in the 'irtoys' R package of Partchev, 2012); this applies a regression model to a moving window of data points centred around a particular value on the x axis. (Within this moving window, data points closer to the central x value are given more weight than data points that are further away). At each chosen x value a 'smoothed' estimate of y is generated based on the regression model 'local' to that point, and these estimates can be joined up to provide a smooth curve describing the shape of the underlying relationship. This approach is described in detail by Bowman & Azzalini (1997). (An alternative approach would be to take a simpler weighted average of the y values at each position of the moving window — this is the type of approach used by Ramsay (2000) for non-parametric analysis of questionnaire data in the TestGraf program.)

The main barrier to using a non-parametric approach for examining the shape of the relationship with the latent trait is that individuals' values on the latent trait are unmeasured, and therefore cannot simply be plotted along the x axis. While estimated trait scores could be obtained from a factor model and then used for a non-parametric analysis, that factor model would assume that the relationship between symptoms and the latent trait followed a logistic function. If this assumption is inappropriate, the estimated trait scores may be severely biased as a result of being distorted to make the logistic model fit; this distortion might actually conceal any evidence that the logistic function is inappropriate.

It would therefore be preferable to find some measure of the latent trait that does not depend on the assumptions of any particular model. The most obvious candidate is to use the sum of the 13 scores from the symptoms that are used to measure the latent trait; this provides a ranking of individuals that approximates their ranking by the underlying latent trait scores. Nonetheless, the summed score (or 'sumscore') is not a perfect substitute for the real latent trait scores, since it ignores the fact that some symptoms may be more closely related to latent trait severity than others. However, this is preferable to relying on latent trait estimates that depend on the very assumptions they are needed to test.

One potential problem with using the summed score in a non-parametric analysis is that the probability of obtaining a symptom score of 1, 2, 3, or 4 will be zero *by definition* at summed scores of 0, 1, 2 or 3 respectively (along with the corresponding phenomenon at the upper end of the scale). For example, the only way to have a sumscore of 0 is to score 0 on every contributing symptom — therefore, the probability of scoring 1, 2, 3 or 4 must be 0 at the sumscore of 0. This may result in artefacts in the shapes of the plotted non-parametric response curves, since they will be constrained to start from a lowest probability of 0 and to approach a highest probability of 1 at the upper end of the scale. This would negate much of the benefit of using a non-parametric approach. For this reason, it is preferable to use what is known as the 'restscore' in place of the summed score, where the symptom that is being examined is omitted from the sumscore for that particular plot (Junker & Sijtsma, 2000).

In order to provide the clearest comparison between results from the non-parametric analysis and the requirements of the standard factor model, it would be helpful to transform individuals' restscore rankings to have as similar a distribution as possible to that of the actual latent trait scores. Since the latent trait is frequently assumed to be normally distributed with a mean of 0 and a standard deviation of 1, the rankings are transformed to follow this distribution. Where many individuals share the same restscore, those individuals are ranked randomly, as in Partchev (2012) and

Ramsay (2000). This procedure makes the implicit assumption that everyone with a restscore of 0 is identical in terms of their true latent trait scores, and so it does not matter how they are ordered. Under the standard normal scaling, approximately 68% of individuals will have scores between -1 and 1, 95% will have scores between -2 and 2, and 99.7% will have scores between -3 and 3; correspondingly, fewer than half a percent of individuals will have scores outside the range of -3 to 3.

Following this approach, the non-parametric logistic regression curves for each of the CIS-R symptoms are plotted in Figures 7.10, 7.11, 7.12 and 7.13. In order to give the best possible description of the underlying shapes of the relationships, these plots combine both halves of the data (n = 22,574). The R code that was used to create these plots is included in Appendix F in Section F.1 on page 309. (The four curves representing the four cumulative probabilities on each plot are estimated independently; therefore the curves may cross inappropriately in regions where the data are sparse.)

We can now examine these plots focusing in turn on each of the four characteristics described above that should be present in the data, if a standard model based on the logistic function is to be appropriate for the CIS-R data:

1. *The upper asymptote should be at 1:* For most symptoms, an upper asymptote of 1 does not appear inconsistent with the data. However, Figure 7.12 strongly suggests that the upper asymptote for anxiety may be lower than 1: the highest curve (corresponding to anxiety scores of 1 or more) appears to flatten out at a probability of approximately 0.9. There is also some suggestion that the upper asymptote for irritability may be lower than 1.

2. *The lower asymptote should be at 0:* The non-parametric plots for fatigue, sleep, irritability and worry appear to suggest lower asymptotes that are substantially higher than 0 for these symptoms. However, when looking at these plots it is important to note that below latent trait scores of approximately -0.5 all individuals will have restscores of 0, and that these individuals have been ranked at random.

   Nonetheless, it is quite possible that there may be a 'floor effect' in the CIS-R questionnaire; this would occur if there are in fact differences in trait severity among individuals with restscores of 0 but there are insufficient items in the CIS-R interview to measure this region of the latent trait. If this were the case, it may be that probabilities for these symptoms do actually fall to 0 at lower values of the latent trait. Since the standard factor model is able to accommodate floor effects in the measurement instrument (so long as the assumption

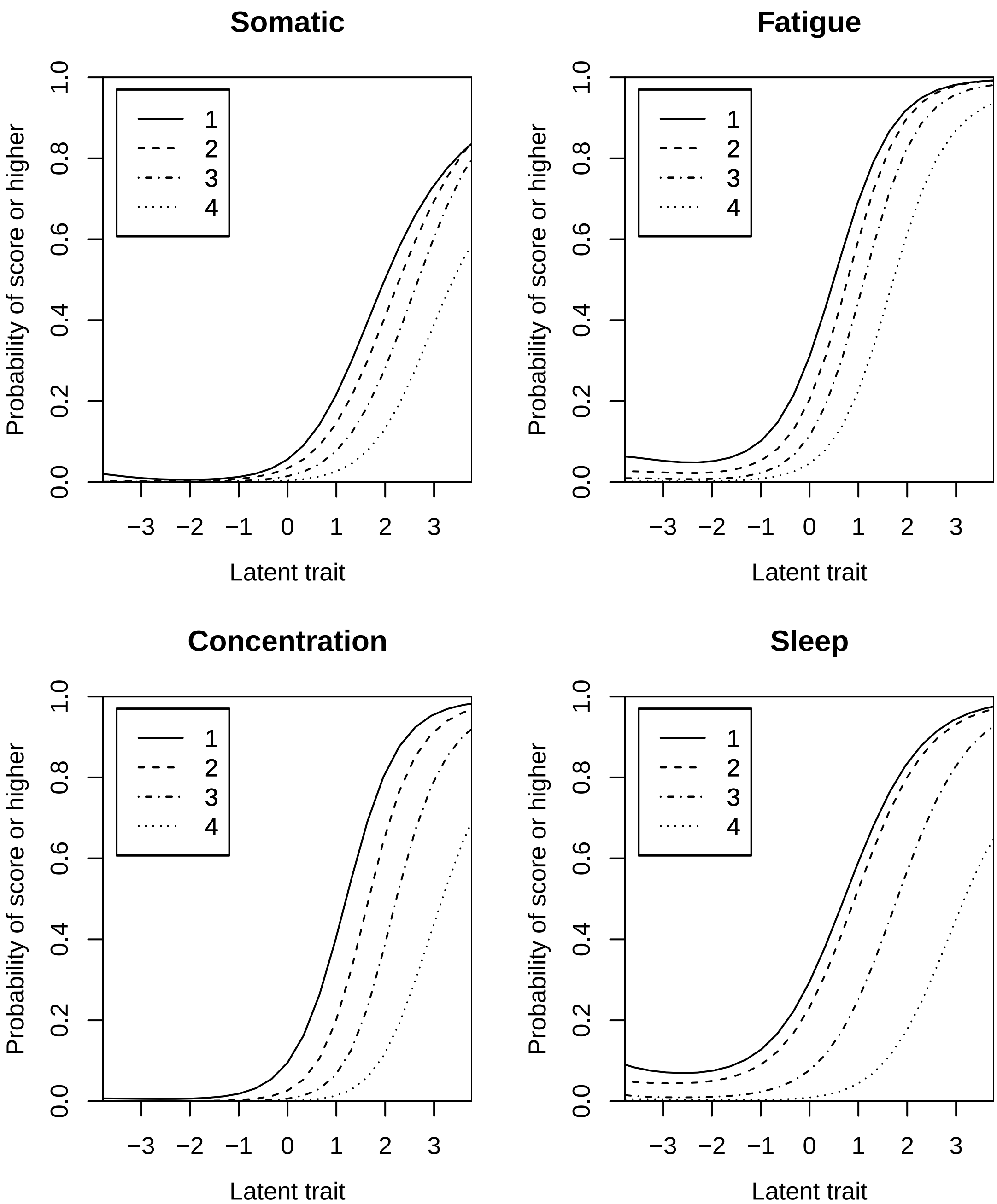

Figure 7.10: Non-parametric response functions describing the relationship between scores on the latent trait and the cumulative probability of endorsing each score. Value of smoothing parameter $h = 1$. All latent trait scores lower than ~ -0.5 correspond to individuals with restscores of 0; these individuals have been ranked at random.

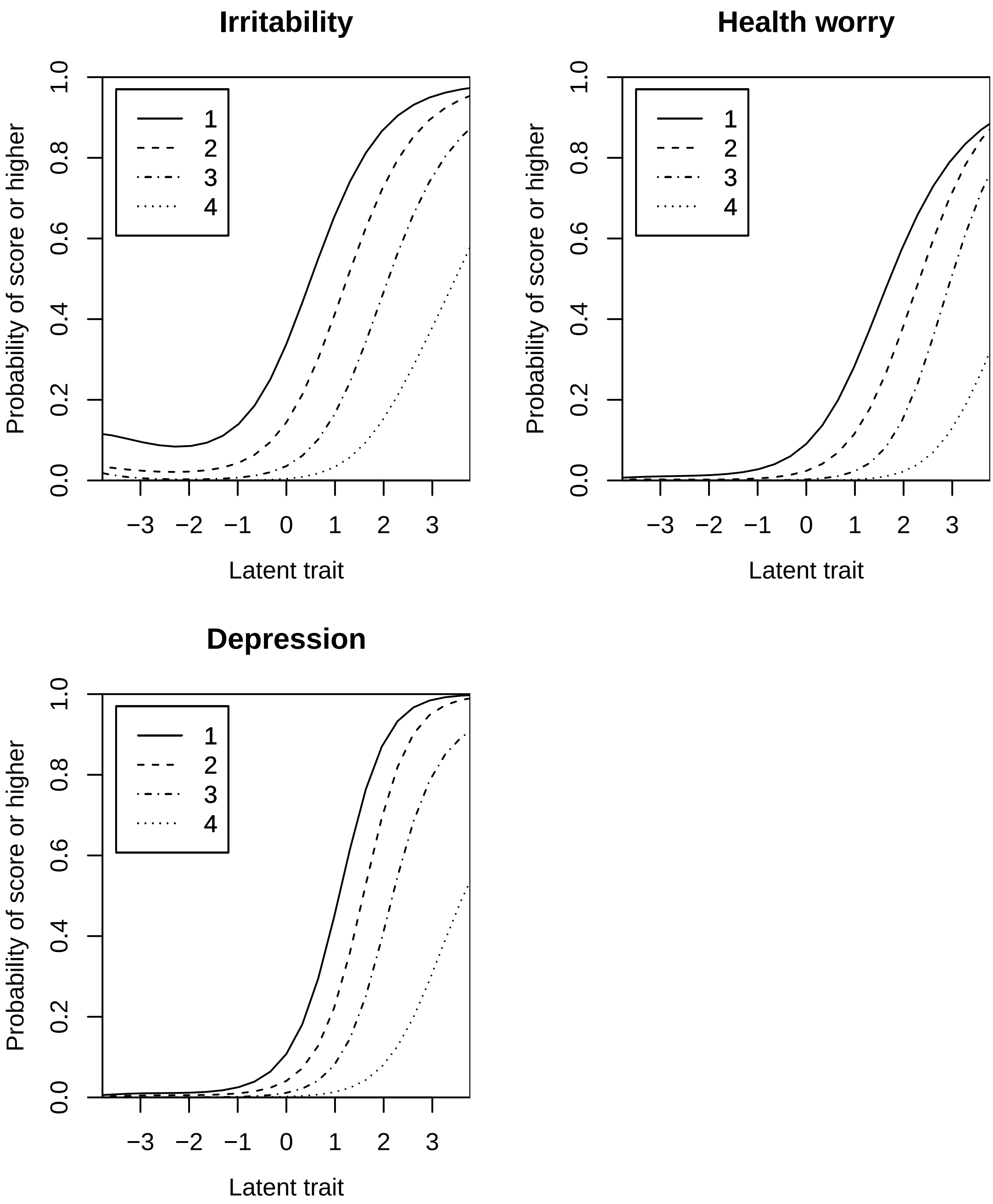

Figure 7.11: Non-parametric response functions describing the relationship between scores on the latent trait and the cumulative probability of endorsing each score. Value of smoothing parameter $h = 1$. All latent trait scores lower than ~ -0.5 correspond to individuals with restscores of 0; these individuals have been ranked at random.

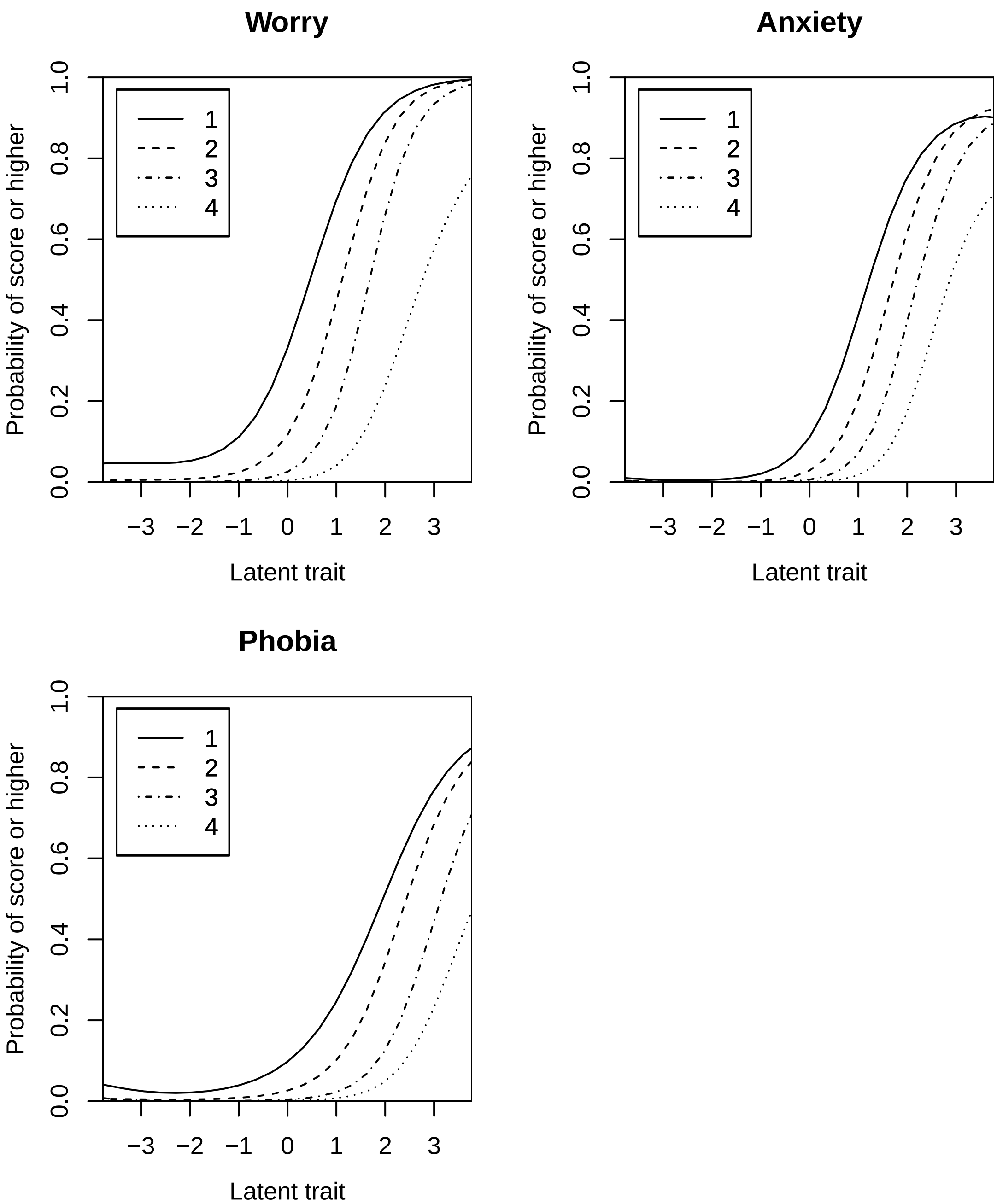

Figure 7.12: Non-parametric response functions describing the relationship between scores on the latent trait and the cumulative probability of endorsing each score. Value of smoothing parameter $h = 1$. All latent trait scores lower than ~ -0.5 correspond to individuals with restscores of 0; these individuals have been ranked at random.

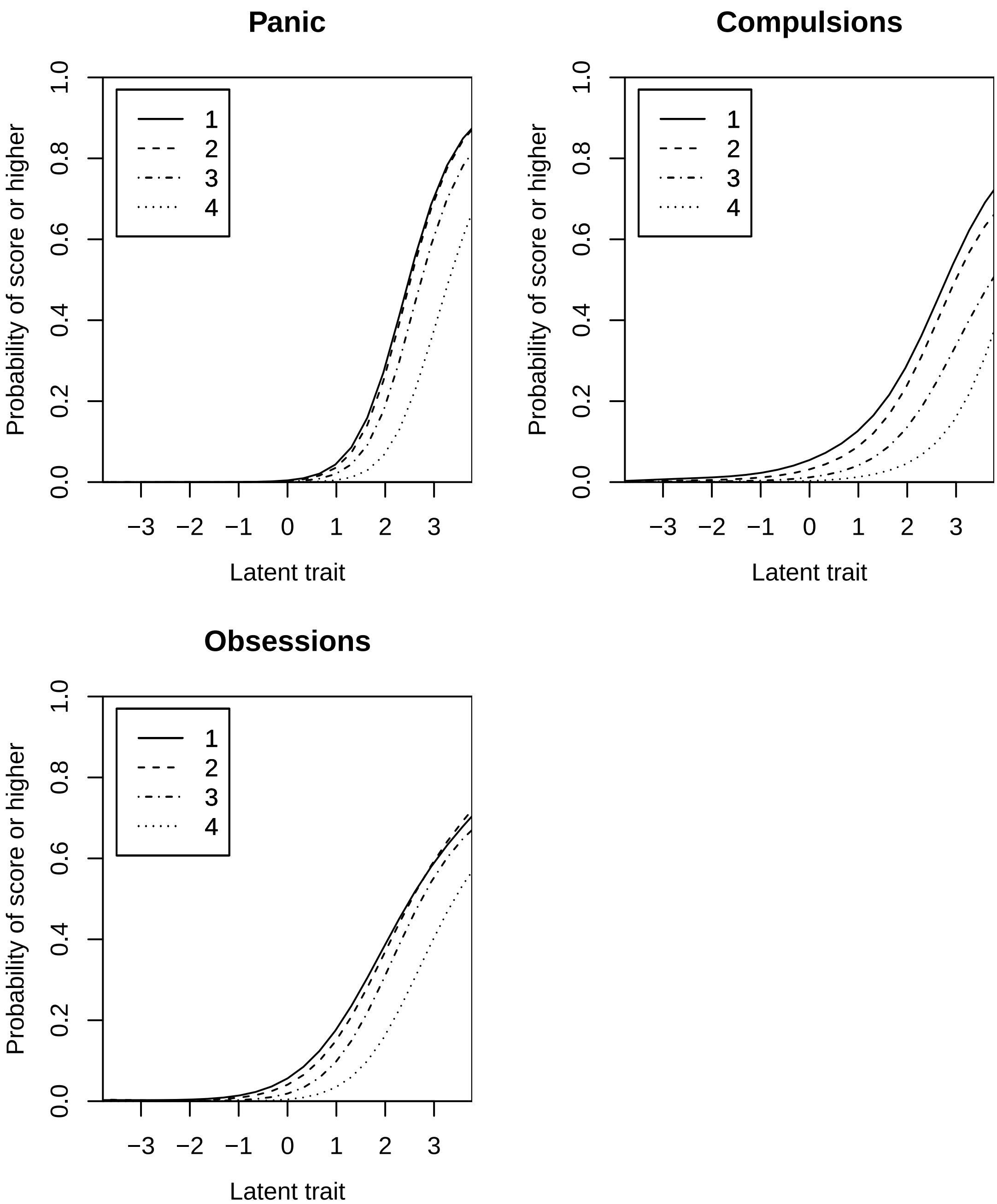

Figure 7.13: Non-parametric response functions describing the relationship between scores on the latent trait and the cumulative probability of endorsing each score. Value of smoothing parameter $h = 1$. All latent trait scores lower than ~ -0.5 correspond to individuals with restscores of 0; these individuals have been ranked at random.

of a normally distributed latent trait is valid), these non-0 lower asymptotes in the parametric plots do not necessarily indicate that the logistic function is inappropriate for these symptoms. It may be impossible to determine from the data alone whether there really is a floor effect in the CIS-R interview, or whether these symptoms do have non-0 lower asymptotes.

3. *The curve should have rotational symmetry:* It is difficult to assess from the non-parametric plots whether the curves really do have rotational symmetry around their inflection points, particularly where there is uncertainty about the shape of the lower section of the curve for the symptoms of fatigue, sleep irritability and worry. However, the upper half of the top curve appears to flatten off for a number of symptoms (particularly for somatic symptoms, health worry, worry, anxiety and obsessions). This also appears to happen for the fatigue and sleep curves: it is clearest from the fact that the lines for the highest two curves are much closer together in the top section of the plot than in the middle section. Perhaps this flattening off reflects the fact that mild versions of a symptom are less discriminating at high levels of the latent trait than they are at lower trait levels.

4. *The curves for ordered categories should be the same shape:* As mentioned in the previous point, it would appear that the curves for mild scores on some symptoms flatten off, while the corresponding curves for the more severe scores do not — this would imply that the curves are not the same shape for all ordered response categories. In addition, the non-parametric plots suggest that for some symptoms the most severe scores have curves that are less steep than the curves for more moderate scores; this is particularly noticeable for worry and depression.

To sum up, the non-parametric plots strongly suggest that some characteristics of the data may be incompatible with the use of the logistic function to model these items: most notably, the upper asymptote for anxiety appears to be substantially below 1; it also appears that the curves for mild scores on some symptoms may flatten off in their upper regions, suggesting that these symptoms become less discriminating at high levels of the latent trait. The plots also indicate that there may be differences in the steepness of the curves for the most severe scores of some symptoms. These discrepancies, if real and not just artefacts of the method adopted for the non-parametric analysis, could lead to the extraction of spurious latent classes to account for the misfit of the logistic curve imposed by the standard factor model.

A more flexible parametric factor model?

The results from the non-parametric analysis suggest that the logistic function may indeed be inappropriate for at least some of the symptoms in the CIS-R data. However, the fact that this analysis had to make do with restscores as a proxy for the true latent trait scores means that it is difficult to draw firm conclusions from the examination of these plots. The restscores make no use of the fact that some symptoms are more strongly correlated with the latent trait than others, and they allow for no differences between individuals with the same restscore. Therefore, it would be desirable to be able to confirm these results in some form of modified parametric factor model that could allow for asymptotes to be non-0 or non-1 (as well as ideally relaxing the assumption of rotational symmetry and allowing different shaped curves for each of the ordered response categories).

For binary response data there are in fact modified versions of standard models that allow the values of the asymptotes to be estimated from the data. The impetus for their development stemmed from the fact that binary factor models (more generally know as item response theory models) are heavily used in the field of standardised educational testing, where it has long been recognised that an individual with very low ability might still pick the correct answer to a test by guessing; candidates who are presented with a set of alternatives to choose from would never be expected to have a truly 0 probability of picking the correct response. An item response theory model known as the '3-parameter' model incorporates an additional set of parameters to estimate the lower asymptote for each item.

Although no standard software has been produced that can estimate a '4-parameter' model (which would allow both the upper and lower asymptotes to be estimated for any particular item), the presence of a non-1 upper asymptote can be investigated by reverse scoring the scale items and then re-estimating the 3-parameter model. (Reverse scoring would mean allocating points for items that are answered *incorrectly* or for symptoms that are *absent*, rather than for items answered correctly or for symptoms that are present as is traditionally done.) Indeed, during just such a reverse-scored analysis of binary psychopathology data Reise & Waller (2003) found that a substantial proportion of items showed a non-1 upper asymptote: they recommended that both normal and reverse scored 3-parameter models should be applied routinely to binary psychopathology data to identify any items with non-0 or non-1 asymptotes, as well as calling for the development of software that allowed items to have either a non-0 lower asymptote or a non-1 upper asymptote in the same model.

Unfortunately, the situation becomes more complicated for ordinal rather than binary response data. While several models have been suggested that would allow non-0 lower asymptotes for ordinal or nominal response data, these models are very complex both conceptually and in terms of estimation, and only the model known as the ‘multiple choice’ model (Thissen & Steinberg, 1984) has made its way into standard software. As its name suggests, this model was developed to handle the responses to multiple choice tests, and it used to be available in a piece of commercial item response theory software called MULTILOG. Regrettably, MULTILOG has now been replaced by newer programs (such as IRTPRO) that no longer offer this capability, and MULTILOG is no longer available for purchase.

However, Thissen *et al.* (2010) record a comment by D.B. Rubin that gives a tantalising suggestion of how non-0 or non-1 asymptotes could be accommodated without the use of such complex models. On hearing about the multiple choice model, Rubin apparently suggested that an alternative approach would be to extend a simpler item response theory model known as the ‘nominal’ model by using a quadratic rather than linear function. (The equation for the model would now include coefficients for θ_i^2 alongside the familiar slope coefficients for θ_i — to see the full equation for this model, see Equation F.3 on page 315 in Appendix F.) The nominal model is in fact one of the few item response theory models for polytomous items that does *not* assume equal slopes/discrimination parameters for all categories of an item (Thissen *et al.*, 2010); it is closely related to the multinomial logistic regression model, and as with multinomial regression, it can be applied to items with ordered or unordered categories. (The application to ordered categorical data may be particularly appropriate where the researcher does not want to assume equal coefficients for all ordered response categories — for an example of this applied to psychopathology data, see Preston *et al.*, 2011.)

Incorporating a quadratic term into the nominal model would allow relationships between items and the latent trait to be non-linear *on the logit scale*, and hence might allow either a non-0 lower asymptote or a non-1 upper asymptote to be accommodated. (A more flexible model would be required to allow *both* non-0 and non-1 asymptotes *for the same item*, but Figures 7.10 to 7.13 suggest that the CIS-R items would only need one or the other, and a similar pattern was noted by Reise & Waller (2003) in their psychopathology data.) At the same time, the use of the nominal model framework would allow the shapes of the response curves to be different for each of the response categories. It therefore seems that fitting a nominal model with quadratic terms would be an ideal way to confirm the findings of the non-parametric analysis above, given that this approach would use actual estimates of individuals’ positions on the latent trait (rather than relying on total scores or restscores as a proxy for actual trait scores). The next section considers whether it is possible to apply such a model to the CIS-R data.

The nominal model incorporating quadratic terms

While the nominal model is available in standard software (such as IRTPRO), there are no software packages that allow a quadratic term to be incorporated into the standard nominal model. However, the nominal model is relatively straightforward to specify in the general purpose WinBUGS or OpenBUGS software, which can estimate a wide range of models using Markov chain Monte Carlo (MCMC) methods. (OpenBUGS is an open-source version of the WinBUGS software package, and all future software development will focus on OpenBUGS rather than WinBUGS.) It is also quite straightforward to extend this nominal model to include quadratic terms. (Example code for estimating a number of item response theory models in WinBUGS, including the 'generalised partial credit model' that is a close cousin of the nominal model, is given in McKay Curtis, 2010.)

Nonetheless, while these models may be straightforward to specify, estimation remains rather more problematic. The main difficulty for models such as these involving large numbers of fixed and random effects (the item parameters and latent trait estimates for each individual respectively) is that they run very slowly in general purpose MCMC software. This is exacerbated by the fact that in these models the location and scale of the latent trait can drift slightly over successive iterations of the estimation algorithm, whatever the estimation method — this is handled with relative ease by most software packages (which can adjust the parameter estimates at every iteration if necessary to keep the scale and the location of the latent trait precisely fixed). However, in OpenBUGS such adjustments are needed for both the item parameters and individuals' latent trait estimates (for this particular model at least), and tend to make the estimation unfeasibly slow for large sample sizes. As long as the necessary identification constraints have been made to fix the location and scale of the latent trait, the slight parameter drift 'averages out' over long enough runs of the MCMC chains, and so is not of itself a problem. Nevertheless, this slight parameter drift leads to high correlations between successive iterations of the MCMC chains, and this means that the model estimation may need to be run for much longer than would otherwise be necessary.

As a result, while it is possible to obtain apparently stable estimates of the parameters of the nominal model incorporating quadratic terms in OpenBUGS, it is simply not feasible to run this model for long enough to obtain high quality information about the variability of these parameter estimates. (The model would probably need to be run for several months to provide complete descriptions of what are known as the 'posterior distributions' for each parameter.) Furthermore, it is not really feasible to perform the kinds of sensitivity analysis that are generally recommended for any MCMC analysis. Therefore, estimating this model in OpenBUGS is not recommended as a general strategy for modelling such data.

Nevertheless, given the limitations of the non-parametric approach described above and the lack of any other known model-based approaches to investigate the potential inappropriateness of the logistic function for the CIS-R data, even 'rough' estimates of the parameters from this model may be very informative. Therefore, the results of this model are presented in this report with the caveat that they are intended to be *indicative* rather than definitive. (It is also worth bearing in mind that quadratic functions are quite limited in the shapes of the curves that they can describe. An attempt was made to specify an even more flexible model in OpenBUGS using splines, but this was abandoned as futile — it ran so slowly that even after running for several weeks it was not clear whether the model as specified was actually identified.)

The general approach to estimating this model in OpenBUGS was to start off with the simpler nominal model, whose results could be validated against those from the software package IRTPRO using marginal maximum likelihood estimation. Full details of the nominal model, the OpenBUGS code and its results are included in Section F.2 on page 311 of Appendix F. The parameter estimates from IRTPRO and OpenBUGS were actually extremely similar. The OpenBUGS model was then extended to incorporate quadratic terms. Full details of the extended model, its OpenBUGS code and its results are included in Section F.3 on page 315 of Appendix F.

The estimated response functions from the extended model are shown with solid lines in Figures 7.14, 7.15, 7.16 and 7.17. While the nominal model is not itself a cumulative probability model, the predicted probabilities are presented as cumulative probabilities to allow easy comparison with the response functions from the standard factor model (indicated by the dotted lines). The equivalent plots estimated on the second half of the data are presented in Appendix E, starting with Figure E.12 on page 305.

Looking at these plots, one thing that is noticeable is the similarity of the curves from the extended nominal model and the standard factor model for the lower sections of each curve. The curves diverge more at higher latent trait values, possibly indicating that in this region of the latent trait the more complex model is accommodating some form of deviation from the assumed logistic functions in the standard factor model.

In order to get the most benefit from the provisional results of this model, we will

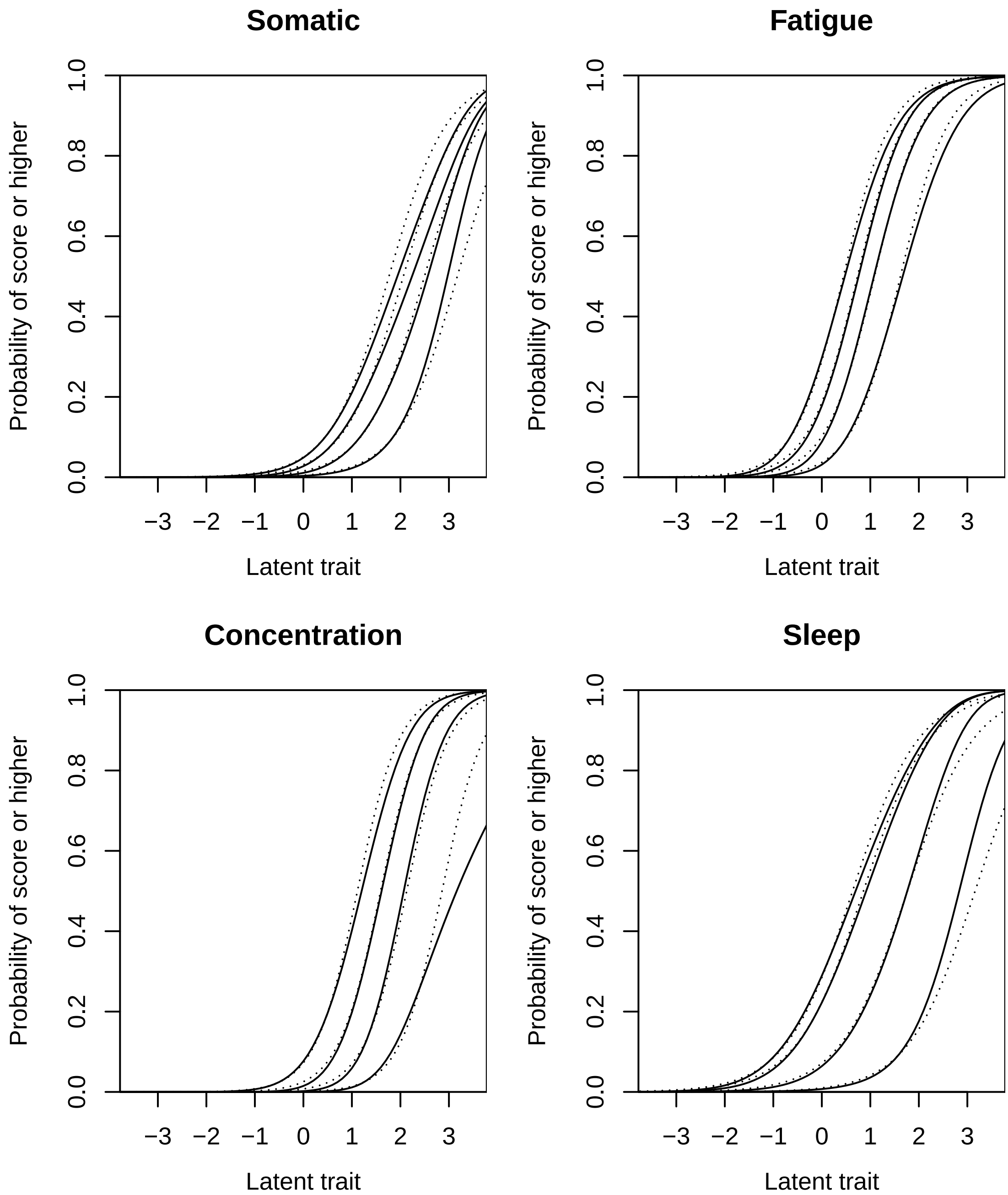

Figure 7.14: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

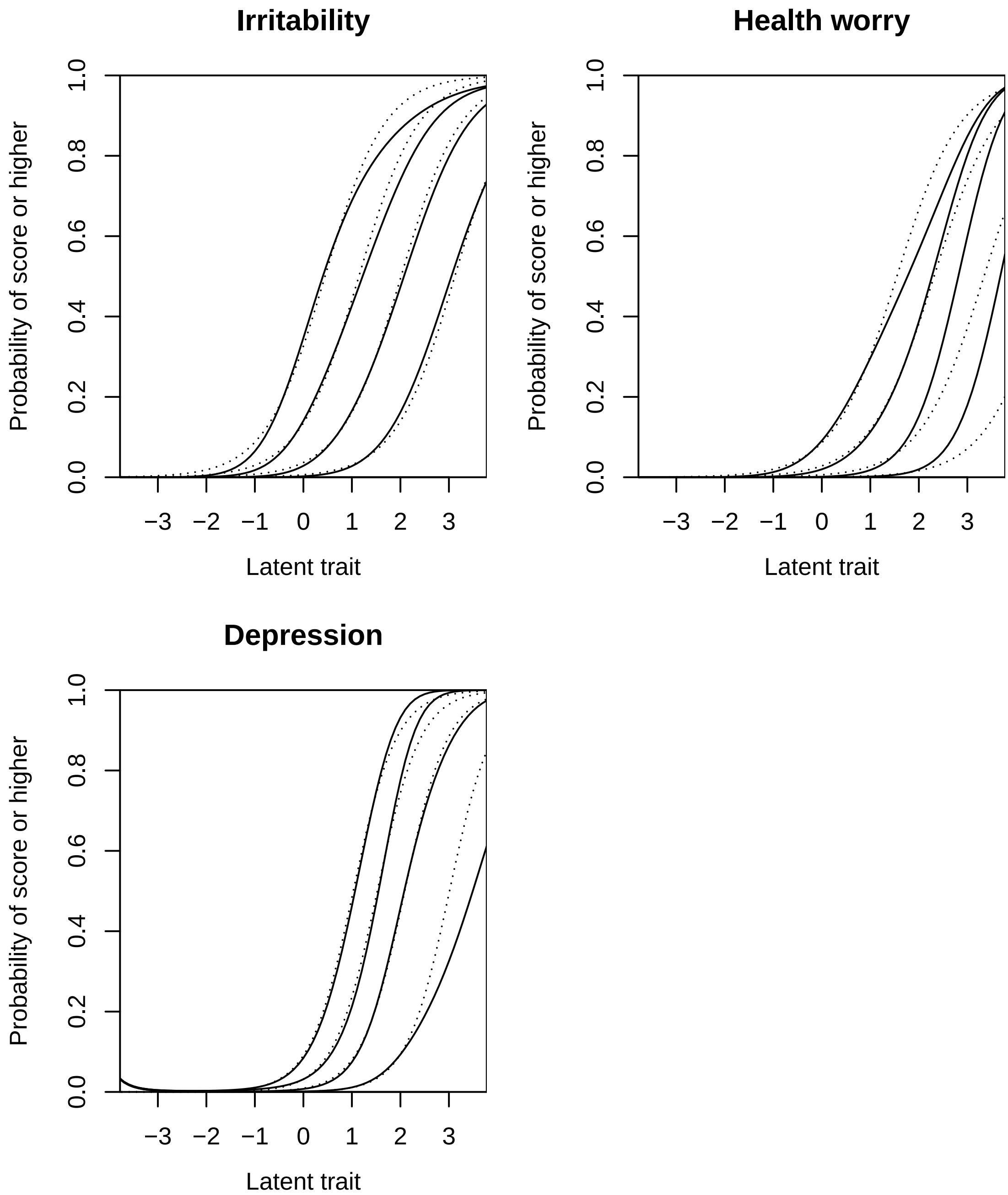

Figure 7.15: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

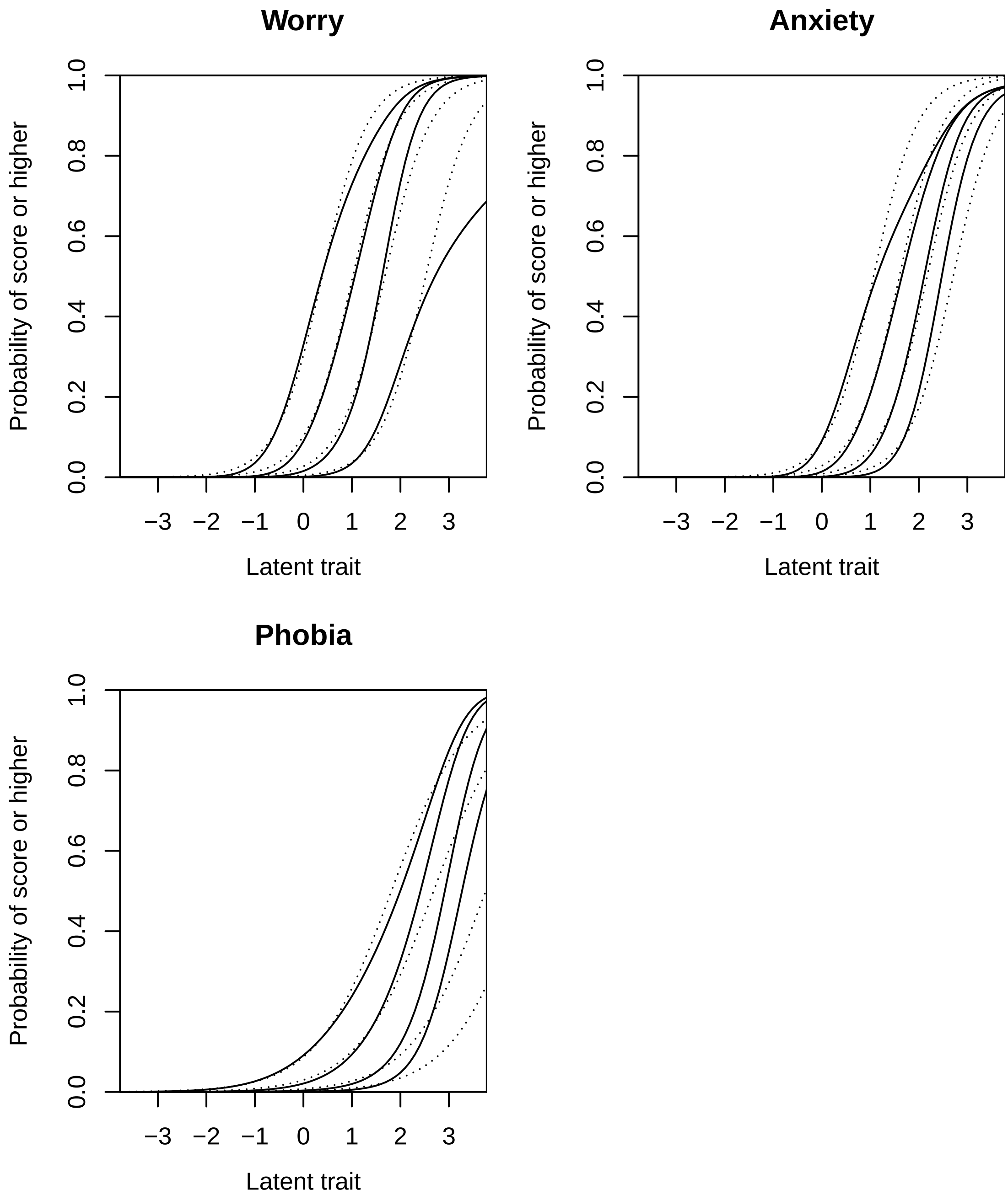

Figure 7.16: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

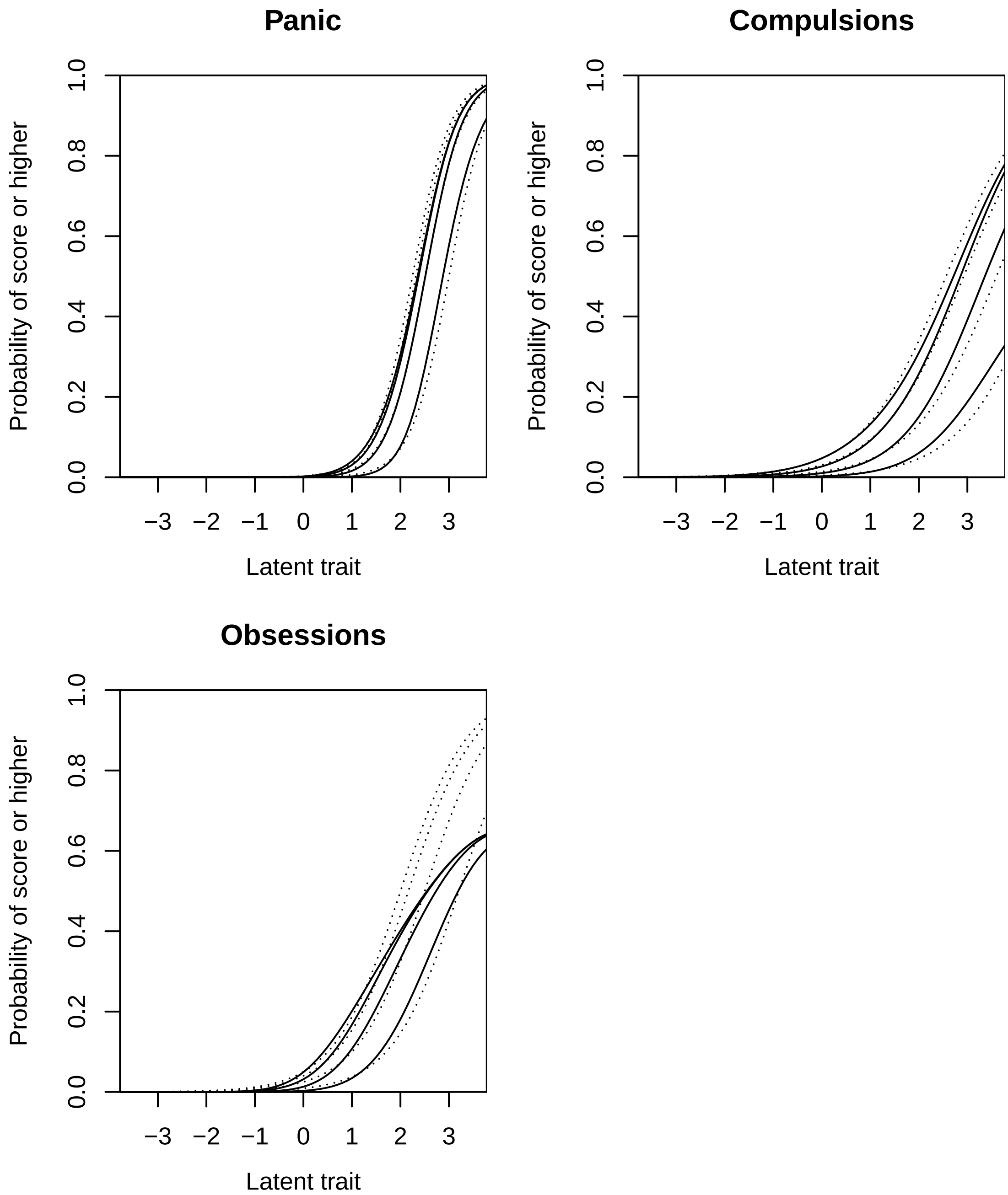

Figure 7.17: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

again consider each of the four characteristics identified earlier that an item should have if the logistic function is to be appropriate for that item; for each characteristic, we will first check whether these results confirm those from the non-parametric approach, before looking to see whether there is any additional evidence in these plots that particular symptoms do not meet the requirements of the standard factor model.

1. *The upper asymptote should be at 1:* The non-parametric plot in Figure 7.12 suggested that the upper asymptote for anxiety was lower than 1. This is confirmed by the plot from the nominal model incorporating quadratic terms in Figure 7.16. In addition, the flexible parametric plots suggest that there may also be non-1 upper asymptotes for the symptoms of irritability and obsessions; these are replicated in the plots from the model fitted to the second half of the data.

2. *The lower asymptote should be at 0:* The non-parametric plots suggested the existence of possible non-0 lower asymptotes for fatigue, sleep, irritability and worry. However, there is no evidence for this in the plots of the flexible parametric model. Nonetheless, given the lack of information in the data at trait scores below approximately -0.5 (individuals below this point scored 0 for every single symptom), this doesn't necessarily prove that these symptoms do have their lower asymptote at 0; it may simply indicate that an asymptote of 0 is *not inconsistent* with the logistic function relied upon by the standard factor model (although this finding may be closely bound up with the assumption that the latent trait is normally distributed in the population).

   Alternatively, it might be that a quadratic function is simply not flexible enough to describe the true shapes of the lower sections of these curves; one of the drawbacks of quadratic functions is that the curves they describe are actually parabolas. A hint of this limitation can be seen in Figure 7.15 in the response functions for depression: at the bottom far left of the plot, the probability of depression actually starts to increase at the very lowest values on the latent trait.

   It may be that any quadratic function which could accommodate a non-0 lower asymptote for symptoms such as fatigue and sleep problems would also lead to a much more noticeable (and inappropriate) increase in probabilities of endorsing the symptom at very low values of the latent trait. Furthermore and as noted earlier, a quadratic function will only be able to accommodate flattening off *either* at the top *or* at the bottom of a curve (since the quadratic curve is essentially shaped like a U or upside down U). Given that mild scores for irritability and worry both show flattening off in the upper regions of their curves, a quadratic function would never be able to simultaneously accommodate a non-0 lower asymptote on these two symptoms; this indicates that a more flexible non-linear function would be preferable if there were any means to fit it.

3. *The curve should have rotational symmetry:* The non-parametric plots suggested that the curves for *low* scores on some symptoms *flatten off* in their upper halves; this was noted for somatic symptoms, fatigue, sleep, health worry, worry, anxiety and obsessions. This flattening off is confirmed by the plots from the flexible parametric model.

4. *The curves for ordered categories should be the same shape:* In addition to some within-item differences implied by the previous point, the non-parametric plots suggested that in some symptoms the curves for scores of 4 are substantially *less steep* than the curves for milder scores; this was specifically noted for depression and worry. This pattern is confirmed in the flexible parametric plots. Additionally, these plots suggest that for other symptoms the curves for more severe scores may actually be *steeper*; this is most noticeable for sleep and phobias, and is replicated in the second half of the data.

To sum up, the results of both the non-parametric approach and the flexible parametric model provide clear evidence that some of the characteristics of the CIS-R data contradict the requirements of the logistic function assumed by the standard factor model. *At least* one of the symptoms has a non-1 upper asymptote; furthermore, the curves for some symptoms do not show rotational symmetry, while for several symptoms the curves of the four ordered response categories have different shapes. It is therefore quite possible that the latent classes in the two and three class factor mixture models are simply accommodating the fact that the assumed logistic function is *inappropriate* for these data.

7.3.4 Latent classes to accommodate misfit of the logistic function?

We have now looked in turn at each of the three indirect roles that the classes in the two and three class factor mixture models may be playing. There is no evidence that the classes are accommodating non-normality of the latent trait, or that they are accommodating misspecification of the measurement structure of the model. However, there is clear evidence that the CIS-R data have characteristics which are inconsistent with the logistic function incorporated in the standard factor model.

This may explain why the bivariate fit statistics and residuals for the standard factor model are so poor (as was described in Table 4.9 of Chapter 4). Nonetheless, these observations do not directly answer the question of whether or not the latent classes in the factor mixture models really are simply accounting for the misfit of the assumed logistic function. It may therefore be helpful to examine whether the descriptions of the latent classes are consistent with the characteristics of the CIS-R data that appear to contradict the requirements of the logistic function.

In fact, the non-1 upper asymptote of anxiety illustrated in Figure 7.16 is strikingly consistent with the patterns of symptoms described for the small 'severe' latent classes in both the two and three class factor mixture models. In Figure 7.16, the gap between the top curve (corresponding to a score of 1 *or higher*) and the second curve (corresponding to a score of 2 *or higher*) disappears at latent trait values greater than 2; since the probability of scoring exactly 1 at any value of the latent trait is given by the vertical distance between these two curves, this corresponds to a zero probability of scoring exactly 1 at the highest latent trait values, even though the non-1 upper asymptote for anxiety means that there is still a modest probability of a score of zero. The gap between the next two curves similarly diminishes at even higher levels of the latent trait, corresponding to a dwindling probability of scoring exactly 2.

As Tables 7.3 and 7.8 describe, one of the most striking characteristics of the 'severe' latent classes is that those assigned with high probability to these classes tend to have either scores of 3 or 4 for anxiety or scores of 0; there are no individuals with mild anxiety in the 'severe' classes, whereas such mild symptoms appear to predominate in the less severe classes. It therefore seems highly likely that the 'severe' latent class is accommodating this characteristic of the CIS-R data that could not be handled by the standard factor model. (The 'severe' class would also accommodate other less dramatic features of the data, such as the non-1 upper asymptote for irritability and the flattening off seen in the curves for mild scores of symptoms such as fatigue, sleep, health worry and non-health worry.)

But what about the additional class extracted in the three class factor mixture model? It seems probable that this class is simply allowing a more fine-grained approximation of the correct shapes of the response curves. Nevertheless, the 'mild' latent class in the three class model may also be related to the presence of non-0 lower asymptotes for some symptoms. Given the difficulties in fitting a *fully flexible* parametric model to the CIS-R data (the nominal quadratic model will allow *either* the upper asymptote of a curve to be non-1 *or* the lower asymptote to be non-0, but not both), it is hard to be sure whether the observed distributions of any symptoms are problematic under the standard factor model at the lower end of the latent trait.

In summary, the observed characteristics of the latent classes in the two and three class factor mixture models appear consistent with the features of the data that have been identified as contradicting the use of the logistic function. This provides support for the suggestion that the latent classes are playing an *indirect role* of accommodating the misfit of the logistic function, and do not therefore represent true groups in the population. Nonetheless, it is not possible to *prove* from an investigation such as this that the data come from a single homogenous population and that the classes are in fact playing an indirect role.

7.3.5 Checking the plausibility of this interpretation

While it is not possible to identify with certainty the role that the classes are playing in the CIS-R data, it is at least possible to check whether such deviations from the assumed logistic functional form *could in principle* lead to the extraction of additional latent classes in the absence of true groups. This can be done by simulating data in a way that replicates the violations of the logistic functional form seen in the CIS-R data, but where the simulated individuals are *known* to come from a single population (rather than from two or more discrete groups). Factor mixture models and standard factor models can then be estimated on the simulated data: if factor mixture models with two or more classes fit much better than the standard factor model, this would provide support for the interpretation that the latent classes in the real data are actually playing this indirect role.

In order to carry out this check, a data set was simulated that attempted to replicate the observed relationships between items and the latent trait seen in the CIS-R data. To do this, the nominal quadratic model described in Section 7.3.3 was specified as the 'true' population model used to generate the data, and the OpenBUGS parameter estimates of this model (listed in Table F.2 on page 318) were specified as the 'true' population parameters. Responses to the 13 ordered categorical items (each with five levels) were simulated for 10,000 individuals; for this simulation, the latent trait was assumed to be normally distributed and the responses to the 13 items were generated to be conditionally independent, given the latent trait score of each simulated individual. In this way, it is known that the simulated data only deviate from the assumptions of the standard factor model in the shapes of the relationships between items and the latent trait. The R code used for simulating the data is given in Section F.4 on page 319 of Appendix F.

The simulated data were then exported to Mplus where they were used for estimating three models: a standard single factor model, a unidimensional factor mixture model with two latent classes and a unidimensional factor mixture model with three latent classes. The factor mixture models were specified in the same way as those that formed the focus of the present chapter; i.e., factor loadings were constrained to be equal, but intercepts and factor variances were allowed to vary across the classes. For comparison, a second data set was also simulated based on the results of fitting the standard single factor model to the real CIS-R data (summarised in Section 4.2.1). The same Mplus models were also considered for this second data set.

Model fit statistics for these three models applied to the simulated data are shown in Table 7.10, along with the fit statistics obtained when these models were applied to the real CIS-R data as previously recorded in Chapter 6. (For the descriptions of these fit statistics, see Section 3.2.7 on page 56.) When the data were simulated under the nominal quadratic model, the standard factor model appeared to fit poorly (in much the same way as when this model was applied to the real CIS-R data): this can be seen from the relatively large overall bivariate Pearson chi-square statistic, as well as from the number of item pairs showing significant lack of fit and the number of large bivariate standardised residuals. However, the fit for the factor mixture model with two classes appears to be much improved; furthermore, the Bayesian information criterion supports the two class factor mixture model over the standard model, while the Lo-Mendell-Rubin adjusted likelihood ratio test also supports the need for at least two classes. These results confirm that violations of the assumption of logistic functional form can lead to the extraction of latent classes *in the absence of true groups.*

The two class factor mixture model for the simulated data actually yields very similar results to the same model applied to the real data. The two classes for the simulated data have prevalences of 81.7% and 18.3%, compared to 81.0% and 19.0% for the real data. Meanwhile, the proportions of individuals within the classes reporting each symptom score (shown in Figure 7.18) are also very similar to those found in the real data (which were shown in Figure 6.16 on page 143).

Table 7.10: Comparing fit statistics for factor mixture models applied to simulated and real data

Model	# par	LL	BIC	Relative entropy	LMR aLRT statistic	LMR aLRT p value	Smallest class size (proportion)	Overall bivariate Pearson chi-square	Pairs with significant* lack of fit (out of 78)	Bivariate standardised residuals > 3 (out of 1950)
Real CIS-R data (n = 11,230)										
1f 1c	65	-90,962	182,531	-	-	-	-	3,974	35	95
1f 2c	119	-90,453	182,015	0.66	1,017	0.0778	2,134 (0.19)	2,135	8	17
1f 3c	173	-90,141	**181,895**	0.58	622	0.0251	1,385 (0.12)	1,667	4	9
Data simulated under nominal quadratic model (n = 10,000)										
1f 1c	65	-80,051	160,701	-	-	-	-	3,173	39	35
1f 2c	119	-79,748	**160,593**	0.66	605	0.0000	1,834 (0.18)	1,489	0	7
1f 3c	173	-79,612	160,817	0.68	273	0.0053	686 (0.07)	1,131	0	1
Data simulated under standard one factor model (n = 10,000)										
1f 1c	65	-82,149	164,898	-	-	-	-	1,276	0	2
1f 2c	Not identified									
1f 3c	Not identified									

Abbreviations: # par, number of parameters estimated in the model; LL, log-likelihood; BIC, Bayesian information criterion; LMR aLRT, Lo-Mendell-Rubin adjusted likelihood ratio test; f, factor; c, class.

*P value used for significance cut-off = 0.01.

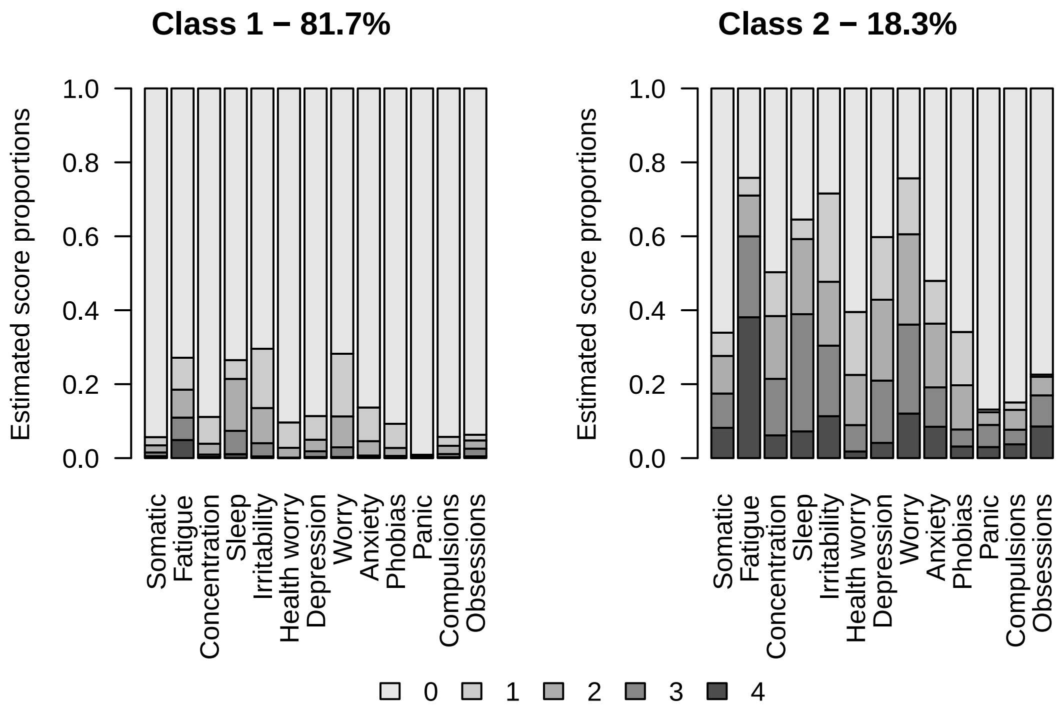

Figure 7.18: Estimated score proportions in the 2 class factor mixture model fitted to data simulated from the nominal quadratic model.

The three class factor mixture model also showed some modest improvement in its bivariate fit statistics (although Table 7.10 suggests that the two class model fitted so well that there was rather limited room for improvement in its bivariate fit). However, the Bayesian information criterion does not support the three class model over the two class model, reflecting the fact that the modest improvement from the additional latent class required a large number of additional parameters.

The three class factor mixture model for the simulated data also yields slightly different latent classes from the same model applied to the real data. While the three classes still appear to be ordered in terms of severity (see Figure 7.19), the prevalences are different (67.3%, 25.8% and 6.9% compared to 58.6%, 29.0% and 12.3% for the real data). The proportions of individuals reporting some of the symptoms (particularly severe levels of fatigue) also show modest differences from those that were seen for the real data in Figure 6.17 on page 145. However, this may simply reflect the fact that the approximate boundaries for the three ordered classes along the latent trait (the 'cutpoints' dividing the continuous dimension up into groups) are at lower percentiles of the distribution for the simulated data than for the real data.

These minor discrepancies between the results of the factor mixture models fitted to

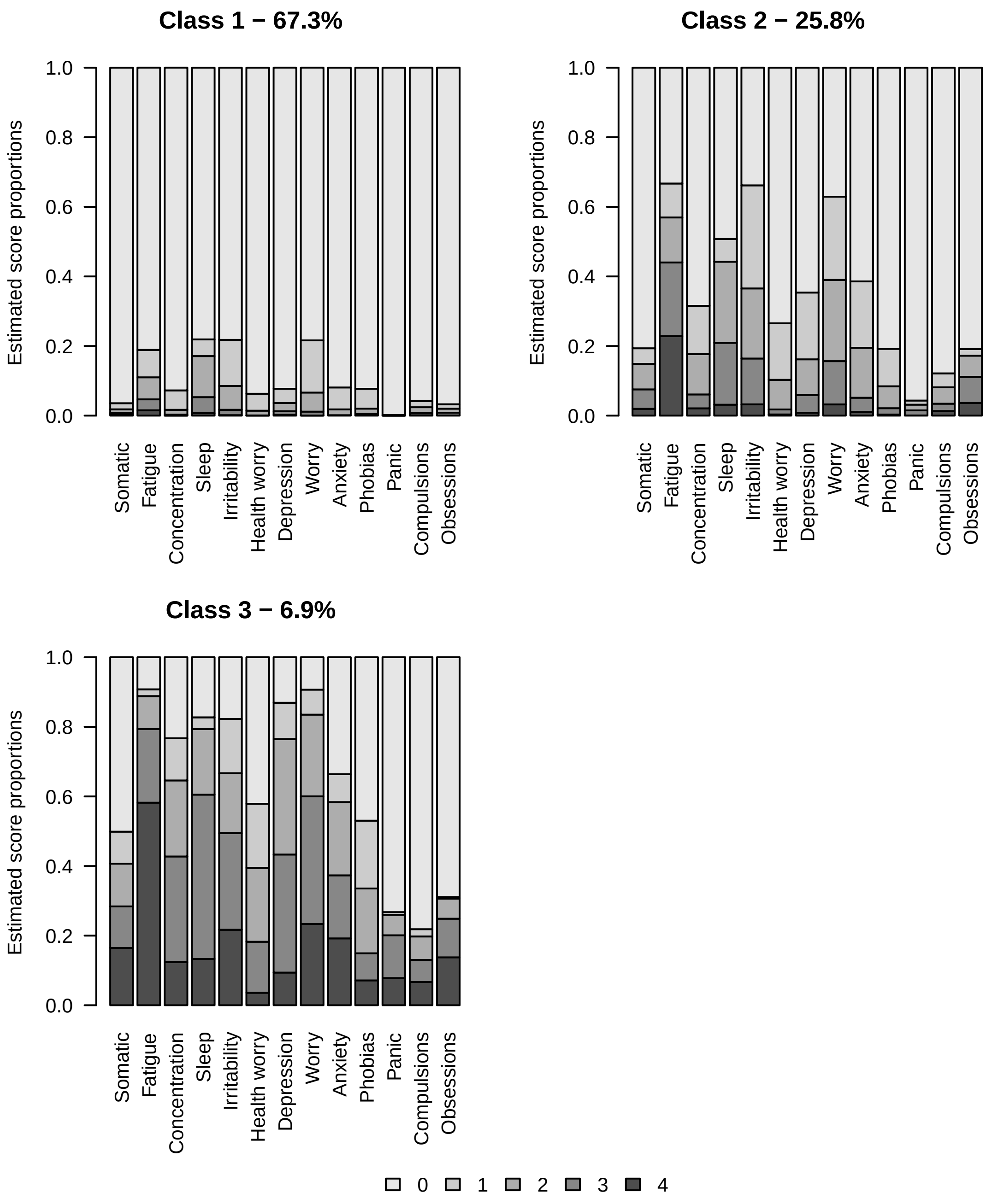

Figure 7.19: Estimated score proportions in the 3 class factor mixture model fitted to data simulated from the nominal quadratic model.

the real and simulated data suggest that the simulation process was not totally successful in its attempts to replicate the characteristics of the real data. A comparison of the relationships between items and the latent trait in the real and simulated data hints at a possible explanation: this may be because the nominal quadratic model is simply not flexible enough to accommodate the full degree of flattening off seen in the upper tails of the response curves for some symptoms (particularly somatic symptoms and anxiety). For the interested reader, this is illustrated and discussed in Section F.4.2 on page 323 of Appendix F.

For reference, the standard factor model fits much better when it is applied to data simulated using the standard factor model than when it is applied to either the real data or the data simulated using the nominal quadratic model (as reflected by the good bivariate fit statistics at the bottom of Table 7.10). Indeed, the fit statistics for this model give an indication of the degree of bivariate misfit that might be expected through sampling error alone if the standard factor model actually fitted the real data. However, the two and three class factor mixture models were not identified when applied to these data; this probably reflects the fact that there was simply no unexplained systematic covariation left for the additional classes to explain. This suggests that 'spurious' latent classes are unlikely to be extracted in the absence of true groups *without some form of factor model misfit*.

To sum up, this simulation exercise confirms that the violations of the logistic functional form seen in the CIS-R data are sufficient to drive the extraction of latent classes *in the absence of true groups*. Furthermore, we have seen that the inclusion of these spurious latent classes in a factor mixture model can lead to a substantial improvement in model fit. These findings appear to provide strong support for the conclusion that the classes of the factor mixture models described at the start of this chapter are actually playing this indirect role.

7.4 Conclusion

This chapter started with a careful examination of the make-up of the latent classes in the two and three class factor mixture models. It found that the classes had no straightforward 'direct' interpretation as clinically meaningful groups in the population. Nonetheless, this investigation did indicate that the 'mild' and 'moderate' classes appeared to show relatively high levels of mild scores for some symptoms, while the 'mild' class in the three class model also included individuals with levels of fatigue that were higher than their levels of other symptoms.

The chapter then moved on to look for evidence of the three types of model misspecification that may result in a model with latent classes appearing to fit better *in the absence of true groups*. There was no evidence that non-normality of the latent trait or misspecification of the measurement model were driving the need for the latent classes. However, there was clear evidence that some characteristics of the CIS-R data contradict the properties of the assumed logistic function. Furthermore, the descriptions of the latent classes appear to be consistent with these characteristics, suggesting that the latent classes may well be playing the indirect role of accommodating the misfit of the logistic function. A simulation study using data generated to follow a non-logistic functional form supported this view.

It is important to note that the evidence for the latent class playing this indirect role remains somewhat 'circumstantial'. We have seen evidence that the CIS-R data contradicts certain properties of the assumed logistic function. It is also clear that the standard factor model fits these data poorly, and latent classes provide a way to accommodate model misspecifications without needing to build a more flexible factor model that can incorporate the problematic features of the data; the characteristics of the observed classes appear to support this observation, and simulation confirms that this can indeed occur for the type of misspecification observed here. Nevertheless, this does not constitute proof; indeed, it is not obvious whether it would ever be possible to provide unambiguous proof in this type of situation.

However, there is very weak support for the view that the classes represent true groups in the population: the substantive interpretation of these 'would-be' groups is far from clear, and the alternative explanation in terms of an indirect role is highly plausible. The author therefore concludes that the classes are probably playing an indirect role. Nonetheless, one of the reasons why this chapter has included so much detail on the characteristics of the extracted classes and the results of the investigations into possible indirect roles is so as to allow readers to form their own judgements.

Given the conclusion that the classes are probably playing an indirect role, this project will not make any attempts to examine the predictors or consequences of class membership. It is not clear what the meaning of any observed differences between classes would be in such a situation, and the results of such analyses could easily be misinterpreted. This is not to say that latent classes playing an indirect role would never be useful when looking at relationships with other variables — when the latent classes are quite homogeneous (as in standard latent class models or the latent class factor models described in the previous chapter), these classes may provide a useful way to examine relationships, particularly if they are non-linear. However, when there are large variations between individuals within each class as well as substantial areas of overlap between classes, this type of approach seems harder to justify. Building from a recommendation of Bauer and Curran, it seems better in these situations to reserve inferences for regression models focussed on the aggregate population, rather than looking at relationships with the latent classes themselves.

Chapter 8

Discussion

8.1 Implications of findings for thesis aims

The general aim of this project was to investigate what latent variable modelling could tell us about the structure of common mental disorders. This was broken down into two more specific aims, as described in the Introduction:

1. To investigate whether factor mixture models could be used to suggest a data-driven classification of common mental disorders.

2. To look for evidence of whether depression frequently occurs as a 'pure' condition without symptoms of anxiety, or whether it would actually be more appropriate to consider depression as part of a broader condition.

The following sections will consider how the results of the latent variable modelling described in the previous chapters contribute to each of these two aims in turn.

8.1.1 Suggesting a data-driven classification

The results of all three of the latent variable modelling approaches used in this project indicate that the symptoms of common mental disorders measured by the CIS-R interview can largely be accounted for by a single underlying latent dimension. This dimension may be interpreted as the overall severity of symptoms or as the severity of common mental distress. Even where models including latent classes have appeared to fit the data better than models without latent classes, the classes still seem to be approximating a single underlying dimension of severity.

This failure to find *any* clinically interpretable latent classes could be seen as support for calls to focus on a broader 'neurotic syndrome' (e.g., Tyrer, 1985; Tyrer *et al.*, 2003) rather than attempting to further refine the present classification that divides the anxiety and depressive disorders into a large number of discrete (but often highly comorbid) diagnoses. Meanwhile, the consistency with which the various types of latent variable models have all supported dimensional interpretations might be seen as providing support for calls to focus more on dimensional rather than categorical aspects of diagnosis and measurement (e.g., Goldberg & Huxley, 1992; Brugha, 2002; Widiger & Samuel, 2005).

However, it is important to note that the findings of this project cannot be taken as evidence for the *absence* of discrete categorical disorders. To start with, it is possible that the classes found by either the two or three class factor mixture models do in fact reflect the presence of real groups. Perhaps some subtle distinctions between classes are not apparent in the way the results were presented; alternatively, perhaps the classes contain some 'core of truth' but allocations of individuals to classes (and hence the interpretations of the classes) are distorted by some form of inflexibility in the model.

Furthermore, there are a number of possible reasons why 'real' disorder categories may have been missed by this study:

- *Noise from factor model misfit.* The 'signal' from the disorders may have been drowned out in the factor mixture models by 'noise' from the violations of the standard factor model. Common sense suggests this might be more likely to occur where the prevalence of a disorder category is very low, or where individuals within this category are only subtly different from some who do not belong to this category. However, this has yet to be investigated by simulation studies.

- *Inability of latent variable models to distinguish between categories and dimensions.* Where the disorder categories differ solely in terms of the overall *severity* of their symptoms, it may simply be impossible to distinguish empirically between a dimension comprising real ordered severity categories and a truly continuous underlying dimension. This point was discussed in detail in Section 5.1.2 on page 84 and Section 6.3.1 on page 122.

- *Omission of important characteristics.* This project may have been limited in its ability to detect real disorder categories by the omission of one or more key symptoms or characteristics from the study data set. For example, a disorder category may be defined by a specific biological characteristic or a particular environmental trigger, or it may be distinguished by the presence of a specific cognitive process. A particular limitation of *this* study may relate to the fact that the symptom score for depressive ideas had to be excluded to avoid the possible extraction of spurious classes related to the structure of the questionnaire (as discussed in Appendix C); the result is that the presence of feelings of guilt, shame, hopelessness and thoughts of suicide could not be incorporated into the latent variable models, which may have been important distinguishing symptoms. Ideally, any future community mental health surveys that use the CIS-R interview would ask *all* respondents about these symptoms. This would allow a test of the assumption inherent in the present form of the interview that these symptoms only occur alongside depressed mood, and would further allow for the results of the present study to be tested for their sensitivity to the inclusion of depressive ideas.

- *Cross-sectional data.* Another important limitation of this analysis is that it only made use of cross-sectional data — the symptom scores were entirely based on the symptoms that the respondents had experienced in the week prior to interview. However, it may well be that some discrete disorders are only distinguished from other more general conditions by their longitudinal course, rather than by having a unique constellation of symptoms. Factor mixture models do not presently allow for the incorporation of longitudinal information alongside high-dimension cross-sectional symptom data; models may be developed in the future that can incorporate longitudinal information, but given the difficulties encountered during this project in understanding and interpreting the models based solely on cross-sectional data, this would perhaps be a step too far in terms of complexity.

- *Lack of power.* A final reason why discrete disorder categories may have been missed is that the study may simply have lacked power to detect them. For a rare disorder whose symptoms overlap with those found in individuals without the disorder, it may be that the sample size of 11,230 (the size of the first split half of the data) used in the main exploratory analysis was inadequate to have a reasonable chance of detecting the class corresponding to this disorder. Splitting the data in half will have reduced the available power, however the large number of models that failed to replicate in the second split half appears to justify the decision to do this. These complex hybrid models have very large numbers of parameters, and appear prone to overfitting; it therefore seems vital to check whether a similar pattern of classes is found in a reserved portion of the data.

  Furthermore, even if an extremely large study did identify what appeared to be a real and stable 'disorder' class, it would still be highly desirable to try and confirm this finding using studies with different measurement instruments and samples. Unfortunately, any results that require enormous sample sizes may in practice be impossible to check in other studies of a similar size; therefore, spurious 'disorder' classes that simply reflect artefacts of the measurement instrument or study design might go unchallenged by future studies.

A further potential problem with the use of extremely large samples to boost power is that factor mixture models applied to these large data sets may be highly sensitive to quite trivial misfit of the standard factor model; even very small departures from the model assumptions might result in the extraction of additional latent classes to accommodate this misfit. This could obscure any true disorders, and it may be extremely difficult to identify that the classes are in fact playing this indirect role when the factor model misfit is so minor. As a result, simply recommending that researchers use larger sample sizes in order to detect rare conditions may be futile. However, since there have been no simulation studies as yet which have investigated the sensitivity of factor mixture models to factor model misspecification under various conditions, this remains a matter of speculation.

In summary, while the factor mixture models used in this project have given no clear evidence of the existence of any discrete disorder groups, there are several reasons why real disorder categories may have been missed. This limitation is in addition to the risk of 'false positive' disorder classes resulting from the apparent sensitivity of factor mixture models to misfit of the standard factor model (which goes hand in hand with the extreme difficulty of identifying for certain what role the classes in the mixture model are actually playing). As such, the results of this analysis are limited in the contribution they can make towards attempts to formulate a data-driven system of classification for the common mental disorders.

Limitations related to the inability of latent variable models to distinguish between dimensions and categories, the use of cross-sectional data and power will be common to most (if not all) studies of this type. Furthermore, while issues surrounding the coverage of key symptoms and the misfit of the standard factor model in the CIS-R data may be specific to the measurement instrument used in this study, similar issues may be common in the data from community surveys that are widely used to address this type of research question. This is because the screening tools used in epidemiological surveys are designed to focus on symptoms covered by the current diagnostic criteria, and are not usually designed either on the basis of their psychometric properties or their utility for exploratory analysis.

Furthermore, there is a certain circularity in using epidemiological questionnaires which have been designed to screen for existing psychiatric disorders as the basis for new attempts to define disorders: if key symptoms have been overlooked in the present system of psychiatric classification, it is unlikely that these symptoms will be measured by most epidemiological surveys, and so data on these symptoms will not be available to researchers. This is clearly problematic for those hoping to use factor mixture models to clarify issues of psychiatric classification. Perhaps cross-cultural studies of the experience of mental distress may be able to suggest ideas for additional symptoms to include in measurement instruments beyond those that are traditionally incorporated. A later section will consider the question of whether factor mixture models are ever likely to be useful for this sort of analysis.

8.1.2 Looking for evidence of 'pure' depression

The second key aim of this project was to look for evidence on whether depression frequently occurs as a 'pure' condition without symptoms of anxiety, or whether it would actually be more appropriate to consider depression as part of a broader condition comprising symptoms of anxiety and depression. The assumption made implicitly at the start of this study was that if there exists a discrete group of individuals whose depression occurs without any accompanying symptoms of anxiety, then this group would emerge as a separate latent class. As we have seen, no such class emerged, although as was noted in the previous section, this cannot be taken as evidence that such a discrete group does not exist. The exclusion of the depressive ideas symptom score may have been particularly problematic here because of its coverage of key symptoms of major depressive disorder such as inappropriate guilt, feelings of worthlessness, hopelessness and suicidal thoughts.

While the results of models including latent classes shed little light on the matter, the results of some of the dimensional models highlighted a feature of the data that requires further consideration. In the previous chapter, it was noted that both the non-parametric item response plots and the nominal quadratic model suggested an upper asymptote for anxiety (at any severity of the symptom) that was lower than 1. This was illustrated in Figure 7.12 on page 191 and Figure 7.16 on page 200 (and was replicated in the second half of the data). Indeed, this was identified as one of the model violations that the classes in the factor mixture models appear to be accommodating. Conversely, the symptom of depression *does* appear to have an upper asymptote of 1 (at least for mild and moderate levels of the symptom), so does not violate this requirement of the standard factor model.

The implication of this is that among the most severely ill individuals (those with very high total scores or high scores on the latent trait) some degree of depression is almost certain. However, the same does not appear to apply to anxiety; anxiety may be notable by its absence in some very sick individuals. (Figures 7.10 to 7.13 suggest that the same may also apply to irritability, and possibly some other symptoms.) It is possible that the non-1 upper asymptote for anxiety reflects the existence of a distinct group of individuals with severe symptoms of depression but no anxiety. To examine whether this may be the case, Table 8.1 presents the symptom profiles of individuals with total scores of 30 or more who had a 0 symptom score for anxiety. (This time the table combines the data from both split halves, since the data are sparse among those with very high total scores.)

In fact, the response profiles in Table 8.1 suggest it would be highly inappropriate to consider the non-1 upper asymptote of anxiety as reflecting a group of individuals with ‘pure depression’. While these respondents may have responded negatively to the questions about non-specific anxiety (anxiety that was *not* brought on by ‘some specific situation or thing’), virtually all of the respondents in this table reported symptoms of phobias or panic (which both relate to feelings of anxiety or nervousness). This is perhaps not that surprising, since it will be difficult to obtain a high total score (and therefore a high score on the latent trait) while scoring 0 on all of these symptoms. If there is a group of individuals with ‘pure’ depression, these individuals are likely to be hidden among the respondents with lower total scores.

The non-1 upper asymptote for anxiety appears simply to reflect the fact that even among those with very high total scores ‘non-specific’ anxiety is not universal. It does not reflect a group of individuals without any symptoms of anxiety. Therefore, it seems that these dimensional models add nothing further to the consideration of whether or not there is a discrete group of individuals with ‘pure’ symptoms of depression.

8.2 Wider implications

While the results of the latent variable modelling have not been as useful as hoped at the start of this project for helping to resolve uncertainties about the classification of the common mental disorders, the results do have important implications for researchers who use latent variable models. These implications relate to two general areas: the application of dimensional latent variable models to the symptoms of mental health, and the use of factor mixture models generally.

Table 8.1: The response profiles of those with 0 scores for anxiety who have total scores of 30 or more — both halves of the data combined

Somatic	Fatigue	Concentration	Sleep	Irritability	Health worry	Depression	Worry	Anxiety	Phobia	Panic	Compulsions	Obsessions	Total score
2	4	4	4	3	3	4	4	0	4	4	4	4	44
3	3	4	4	4	3	3	3	0	4	4	4	4	43
4	4	4	4	3	3	4	3	0	4	0	4	4	41
4	4	2	4	4	3	4	4	0	4	4	0	4	41
0	4	4	4	3	4	4	4	0	3	4	3	3	40
4	4	4	4	2	4	3	4	0	3	3	0	4	39
4	3	3	4	4	2	3	3	0	4	0	3	4	37
4	4	1	4	0	3	2	3	0	4	4	3	4	36
3	4	2	3	2	2	4	4	0	3	4	1	4	36
4	4	3	3	3	2	3	4	0	1	4	0	4	35
4	4	4	3	4	3	3	3	0	3	4	0	0	35
4	4	3	3	4	3	4	4	0	3	3	0	0	35
4	4	4	4	4	3	4	3	0	2	3	0	0	35
4	4	4	4	3	3	2	3	0	0	0	3	4	34
4	2	0	4	0	3	4	3	0	4	4	3	3	34
4	4	0	4	3	2	2	3	0	4	4	0	3	33
3	3	3	4	3	1	2	3	0	1	4	2	4	33
3	4	2	2	3	4	3	4	0	4	4	0	0	33
3	4	4	3	3	2	3	0	0	4	3	0	4	33
4	4	4	3	3	2	3	4	0	3	0	3	0	33
4	4	1	4	3	2	2	3	0	4	0	2	3	32
0	4	0	3	2	3	4	3	0	3	3	3	4	32
3	4	3	4	2	2	4	4	0	0	3	3	0	32
0	3	4	2	4	0	0	4	0	4	4	4	2	31
4	4	0	4	3	3	2	2	0	4	4	1	0	31
0	4	4	3	0	4	3	4	0	2	0	3	4	31
4	3	2	3	2	3	3	4	0	4	0	3	0	31
2	4	3	2	4	0	3	3	0	3	0	3	4	31
0	4	4	4	4	0	1	2	0	3	0	4	4	30
4	4	2	2	3	0	2	2	0	3	3	1	4	30
0	4	3	4	2	4	3	3	0	3	0	0	4	30
4	4	3	4	3	1	3	2	0	3	3	0	0	30
0	4	2	2	4	2	3	2	0	4	4	3	0	30
3	0	3	3	4	4	3	4	0	1	0	2	3	30
4	4	3	4	4	3	3	1	0	3	0	1	0	30
0	4	0	4	0	1	4	3	0	3	3	4	4	30

Note: Symptom scores of 3 and 4 have been highlighted to facilitate interpretation.

8.2.1 Dimensional modelling of mental health symptoms

The factor mixture models used in this project may have failed to uncover any clinically meaningful classes, but they did highlight violations in the assumptions of the standard factor model for categorical data that may have been difficult to spot by other methods. There are no tools for checking the assumption of a logistic functional form in standard software for modelling with latent variables such as Mplus (or for checking the assumption that the relationship is the same shape for all ordered categories). It is therefore unlikely that the violations of this assumption identified in the previous chapter would have been spotted without the detailed investigation that was required to work out what role the classes in the factor mixture model were playing.

This is important because violations of the assumption of functional form may be common in mental health. The assumption that the response curves describing the relationship between symptoms and the latent trait follow a logistic functional form is actually very restrictive. For example, the standard factor model for categorical data implies that *all* severely ill individuals at the top end of the latent trait should endorse relatively common symptoms that are strongly related to the latent trait (have high factor loadings). However, there may be many reasons why a very sick individual may not experience or report a particular symptom, even though that symptom may be highly discriminating among those who are less unwell. For example, someone who was extremely distressed in the week prior to their interview may have found it much harder to answer the following question in the anxiety section of the CIS-R interview than someone with only mild symptoms:

> [J5] In the past month, when you felt anxious/nervous/tense, was this always brought on by the phobia about some **specific** situation or thing or did you sometimes feel **generally** anxious/nervous/tense?

This question was used in the interview to clarify whether an individual who had reported feeling anxious about specific things or situations (classified in the CIS-R as phobia symptoms) was also suffering from general anxiety. However, it may be much easier to answer for someone who has only felt anxious once or twice in the past month than for someone who has felt anxious every day. Therefore, some of these very sick individuals may have failed to report general anxiety that they did actually experience. It may also be that the distinction between general and phobic anxiety is much less informative in those who are severely distressed than among those with mild symptoms: those who are anxious about specific things most or all of the time may have little opportunity to experience general anxiety. These factors may contribute to the flattening off and non-1 upper asymptote in the response curves for anxiety seen in Figure 7.12 on page 191.

Similar speculation could be made about the reason for the non-1 upper asymptote for irritability seen in Figure 7.11 on page 190. The standard factor model assumes that all severely ill individuals at the top of the latent trait should endorse the symptom of irritability. However, the CIS-R questions about irritability are framed in terms of becoming irritable or having arguments with 'those around you'; individuals who live alone and who are so sick that they struggle to go to work or carry out social activities may have had few opportunities for becoming irritable with those around them in the previous seven days. There may also be stable personal characteristics such as personality or ethnicity that modify the expression of a particular symptom in individuals who are severely ill, or that influence whether a symptom that is experienced is actually reported. Characteristics such as these may explain the flattening off seen in some response curves.

It is for reasons such as these that Reise & Waller (2003) argue for the development and provision of models allowing both non-0 lower asymptotes and non-1 upper asymptotes in standard software for modelling psychopathology data. For the time being, researchers with binary data may follow their recommendation to try '3-parameter' models that allow either the upper or lower asymptote to be freely estimated (with standard and reverse-keying of responses). However, most researchers with ordinal data have no option to use such models. Furthermore, even '3-parameter' models may fail to capture more general forms of asymmetry in the response curves. Unfortunately, the use of models which inappropriately assume a logistic response function may have unfortunate consequences in several of the applications for which these models are widely used:

- *Investigating the properties of items and tests.* Many researchers use item response theory models to investigate the properties of items as well as to estimate the precision of latent trait estimates at each point along the latent dimension, a measure know as 'test information' (for an example, see Sharp *et al.*, 2006). However, when the assumed functional form is inappropriate estimates of item parameters may be biased or misleading. For example, Reise & Waller (2003) found that when a '3-parameter' model was needed to model a non-0 lower asymptote, discrimination parameters were lower when estimated under the standard model than under the 3-parameter model. They also found that the shape of the test information function depended on which model was used.

- *Investigating differential item functioning.* Another common use of factor models and item response theory models in mental health is to investigate whether measurement/diagnostic instruments have the same properties in different groups (for an example, see Gallo *et al.*, 1994). Researchers and clinicians are often particularly interested in whether individuals from different ethnic groups or countries respond to items or experience symptoms in the same way; if they do not, this differential item functioning may bias cross-cultural comparisons or even lead to inappropriate treatment.

However, for models based on *continuous* data, Bauer (2005) demonstrated that even small violations of the assumption of a linear functional form can lead to false positives when testing for differential item functioning. These will not occur when the groups being compared have very similar distributions of scores on the latent trait, but may be highly likely to occur when there are large differences between the mean scores of the groups (which may well happen when comparing young and old, men and women, etc.) This is because, under the standard assumption of linearity, straight lines are being used to approximate a relationship that is curved; if the groups are centred on different portions of the curve, it is likely that the lines fitted for each group will have different slopes and intercepts. This will give the impression that item parameters vary across groups, when in fact the overall shape of the curve may be identical in each group.

For models with *binary or ordinal* data, violations of the logistic functional form imply non-linearity *on the logit scale* in which the model is actually estimated. For this reason, it seems likely that Bauer's finding will generalise to these models also. Therefore, it is possible that some of the reports of differential item functioning in the psychiatric literature may simply be artefacts resulting from ill-fitting factor models.

- *Exploratory factor analysis.* Factor models are also commonly used in mental health to explore the dimensionality of measurement instruments; this may be to check that a set of items form a unidimensional scale as expected (as in Chapter 4), or to confirm whether a set of symptoms conform to some hypothesised multi-factor model (for an example, see Elhai *et al.*, 2012).

  However, for exploratory factor analysis with *continuous* data, McDonald (1965) established that violations of the assumption of a linear functional form could result in the extraction of what have confusingly come to be known as 'difficulty factors'. These are additional factors that do not represent content-related dimensions in the data but are instead artefacts of the non-linearity in the relationships between items and the latent trait. Such 'difficulty factors' may give the misleading impression that a unidimensional measure actually comprises two or more content-related dimensions.

As before, since in models for binary and ordinal data any violations of the assumed logistic functional form correspond to non-linearity on the logit scale, it seems likely that difficulty factors may also afflict models for binary or ordinal data. Indeed this may have affected the exploratory factor analysis described in Chapter 4. It is therefore possible that ill-fitting factor models may have resulted in misleading conclusions in the literature about the dimensionality of measurement instruments or specific groups of symptoms. This may be particularly problematic for psychiatry since investigations of dimensionality are often used to argue for new ways to conceptualise the structure of psychopathology (for example, see Markon, 2010).

Taken together, these three areas of concern highlight the potential importance of violations of the functional form assumption in models for categorical data applied during psychiatric research. However, it is rare to see this assumption checked in any way, perhaps as a result of the perceived lack of tools by which to do this (particularly for those who don't have access to the specialised commercial software designed for estimating and scoring item response theory models). Nonetheless, the type of non-parametric approach used in the previous chapter may provide a useful way to assess the validity of this assumption.

Indeed, Meijer & Baneke (2004) recommended that researchers analysing psychopathology data *routinely* apply this type of non-parametric modelling *before* fitting the standard factor or item response models. The results of this project reiterate the importance of this recommendation. Researchers who are happy to use the statistical computing package R (R Core Team, 2012) can use code similar to that given in Section F.1 of Appendix F, while those who prefer a 'point and click' approach can try the TESTGRAF package (Ramsay, 2000). (TESTGRAF may be less useful for researchers with ordinal data, since it presents plots of probabilities for the ordered response categories themselves rather than the *cumulative* probabilities that are actually modelled in the factor model for ordinal data.).

The present focus of much of the methodological work on latent variable modelling seems to be on developing software for (and encouraging the application of) more and more complex models, for example, 'structural equation mixture models' (Tueller & Lubke, 2010) and 'multilevel mixture item response theory models' (Vermunt, 2007). However, the difficulties encountered during the interpretation of the factor mixture models used in this project indicate a more pressing need to focus on methods that can relax the assumptions of simpler but more widely used models.

8.2.2 Use of factor mixture models

The second broad area where the results of this project may have implications is in the area of the general use of factor mixture models. It has already been mentioned that these models do not appear to provide any useful information in relation to the research questions of the current project. However, it is worth considering whether these models may be more useful when applied to other data.

The key problem with factor mixture models is that their results are so ambiguous. If a model including latent classes appears to describe the data better than a purely dimensional model, these classes *may* represent true groups. However, the classes might instead be accommodating some form of misspecification of the factor model. It may be possible to rule out some forms of misspecification as potential explanations, but it might be impossible to demonstrate conclusively that they are playing a particular role. Moreover, the attempt to identify the role the classes are playing could be difficult and very time consuming.

If dimensional models appear to describe the data just as well as models including classes, it may be true that there are no discrete groups in this population. However, there are also a number of reasons why real groups may not be detected, such as the omission of key differentiating characteristics, or the fact that it may simply be impossible to distinguish empirically between underlying categories and dimensions (as was discussed in Section 5.1.2).

For such a time consuming exercise, the results of factor mixture models may not be particularly informative. They are certainly not a magic bullet for identifying the true latent structure, and if latent classes that are playing an indirect role are misinterpreted as representing real groups, the models may do more harm than good. The author would go as far as suggesting that factor mixture models are *not useful* for resolving uncertainties over psychiatric classification, and that researchers would be better off spending the many months that are required to do justice to such an investigation on other research questions. While uncertainties over classification are clearly important issues to address, it may be that progress cannot truly be made without the development of a better understanding of the biological and cognitive mechanisms underlying these disorders. Once (or if) these mechanisms are understood, it is likely that disorders with different disease processes or aetiologies will be much easier to distinguish.

Can factor mixture models ever be useful?

Although factor mixture models may not be useful for identifying the true latent structure of mental disorders, it is worth considering whether they may be useful in other research applications. It is important to remember here that the phrase 'factor mixture model' is only used in this project to refer to models in which item intercepts at least are allowed to vary between classes, as described in Chapter 6. Models in which factor loadings and intercepts were constrained to be equal across the classes are described as 'semi-parametric factor models' (where factor scores vary within each class) and 'latent class factor models' (where factor scores have zero variance within each class).

As was noted in Chapter 6, semi-parametric and latent class factor models are most useful in their *indirect* role, since they are able to accommodate violations of the assumption of a normally distributed latent trait in the standard factor model (or the assumption of multivariate normality for models with two or more factors). Latent class factor models are particularly useful in this indirect role, since they are extremely flexible in the shapes of the distributions that they can accommodate, and they are much quicker to estimate than semi-parametric factor models and factor mixture models (since they do not require numerical integration during model estimation). Latent class factor models could also make it relatively straightforward to model non-linear relationships between predictor variables and a latent trait in a structural equation model, since the latent trait could be modelled as classes rather than as a continuous variable with minimal loss of information.

Unfortunately, factor mixture models are not nearly so useful in an indirect role. The main reason for this is that the scale and/or location of the latent trait must be fixed separately for each class, and so factor scores from individuals in different classes cannot be pooled after model estimation; for example, a factor score of 2 may mean low trait severity in one class but high trait severity in another class, so ordering the entire population by factor scores would be meaningless. Furthermore, the classes themselves are limited in their usefulness as an approximation of the latent trait, since they will be few in number and the ranges of factor scores that they cover may overlap substantially (as in Figure 7.3 on page 165).

Factor mixture models could be used to identify violations of assumptions in the standard factor model, as they have done in this project. However, they are not really an ideal method for this, since they are very slow and problematic to estimate, and it may be difficult and time consuming to work out what role the classes are actually playing. Other methods for exploring possible violations of assumptions are likely to be much quicker to carry out and easier to interpret.

In summary, the author sees no real benefits from the use of factor mixture models, and in fact would specifically recommend not using them. This is because there is a strong danger that classes playing an indirect role may be misinterpreted as representing true groups. However, these models are becoming increasingly popular as knowledge of them spreads, and despite the limitations listed here, some researchers may remain convinced of their worth. (As justification for this last statement, the author points to the example of growth mixture models, which are still being widely used in attempts to identify meaningful latent trajectory classes 9 years after the Bauer & Curran (2003) article highlighting the difficulties with interpreting these longitudinal models.) Given that these models will continue to be used, the next chapter will consider how the risk of misinterpretation can be minimised and how the amount learned from these models can be maximised.

Chapter 9

Minimising the risk of misinterpretation of factor mixture models

Since some applied readers will never have even heard of factor mixture models before they encounter an article that uses them (and will need guidance on how the results should be interpreted), the responsibility for the interpretation of the results of a factor mixture model must lie with the authors writing up the results for a journal article. Therefore, the key issue in minimising the risk of misinterpretation is how the results are reported. This chapter will consider what standards of reporting are needed to help minimise the risk of misinterpretation, and whether these standards are currently being met.

9.1 Basic reporting requirements for factor mixture models

Since it may be impossible for the researcher to identify with certainty which role the classes are playing, the bare minimum requirement for any article in which models with latent classes are estimated should be that it is made explicit that classes can play both direct and indirect roles. While the authors of some key methodological papers describing factor mixture models do mention that latent classes may play different roles (such as Lubke & Muthen, 2005), many readers of an applied research article will not have seen such papers, so they are dependent on the author of the article to highlight this limitation of the models. Furthermore, it is not sufficient to mention solely that mixture models may accommodate non-normality in the latent trait: it should also be mentioned that the classes may accommodate other forms of model misspecification, such as misspecification of the factor structure, non-linear relationships between pairs of factors and violations of the assumed functional form linking indicators to the latent trait.

Beyond this, authors really need to provide evidence to support their chosen interpretation. If an author wishes to interpret the latent classes that they have found as representing real groups, their interpretation would be far more convincing if they could show that they had checked for the plausibility of other explanations (for example, by showing that they had checked for the validity of the assumption of a logistic functional form). If the author is unsure which role the classes are playing, it might aid subsequent research if they could highlight this in their discussion section, rather than simply avoiding saying whether or not they believe the classes represent real groups (although it has to be acknowledged that this may make the model results harder to publish). At least this would reduce the risk of researchers who are unfamiliar with these models assuming that the classes must represent real groups, and treating them as clear evidence of such.

In order to assess whether existing papers that have used factor mixture models to investigate the latent structure of mental disorders may have misinterpreted their findings, it is important to find out whether existing papers tend to meet any of these requirements. Therefore, a brief literature review will be carried out to investigate the current standard for reporting for these models.

9.2 Literature review

9.2.1 Search strategy

In order to identify papers that have used factor mixture models to investigate the latent structure of mental disorders, a search was carried out on the Web of Science (date of search: 21st August 2012) for articles where the 'Topic' contained the phrases 'factor mixture', 'item response mixture', 'IRT mixture' , 'Rasch mixture' or 'mixed Rasch': the search covered any publication year and included all available databases. This was not intended to be an exhaustive systematic review, but simply to provide a 'snapshot' of the current state of reporting of these models. Therefore, it is possible that some relevant papers were missed. The lack of a standardised terminology to describe these models presents one possible reason for missing papers, since papers that have described their models in other ways may have been overlooked. Nonetheless, it is hoped that the majority of the relevant papers have been identified; the omission of a small number of papers will not affect the conclusions of this narrative review.

The search returned 149 results of varying degrees of relevance. These are summarised below:

- 19 papers were completely unrelated. Many of these related to biochemistry and focussed on mixtures of growth factors.
- 58 papers were methodological and focussed on the methodology rather than the application of factor mixture models. Some of these papers were largely theoretical, some used simulation studies, while some aimed to describe and illustrate a novel methodology in a way that would be accessible to applied researchers (often using real data). As may be inferred from this last point, the distinction between methodological and applied papers was not always clear cut, particularly where the use of factor mixture models was illustrated with real data. Five of the papers identified here as methodological illustrated the models by applying them to real mental health data (Kim & Muthen, 2009; Lubke & Neale, 2008; Lubke & Tueller, 2010; Masyn *et al.*, 2010; Muthen & Asparouhov, 2006).
- 43 papers applied factor mixture models to answer substantive questions, but these questions did not cover the symptoms of mental disorders. These covered a wide range of topics such as intelligence (including IQ, reasoning and spatial ability), life satisfaction and physical functioning.
- 29 papers applied factor mixture models to investigate the symptoms of mental disorders. However, several of these papers did not report the results of new models: Bauermeister *et al.* (2010) is a systematic review of ADHD research, Bernstein *et al.* (2011) carried out secondary analysis using the model results from a previous paper, while Zheng & Baker (2007) is a conference abstract containing no model results. This leaves 26 papers that report the results of factor mixture models used to investigate the symptoms of mental disorders.

We will now examine more closely the 31 papers (5 methodological and 26 substantive) that have applied factor mixture models to mental health data in order to see how they have reported their findings.

9.2.2 Methodological papers illustrated with mental health data

The methodological papers that applied factor mixture models to mental health data will be considered separately for a couple of reasons. First, since much of the text in these papers is taken up with methodology, these papers often have very limited scope for going into detail about the interpretation and implications of the model results they present, and may not be able to go into the same level of detail as applied papers. For example, in Lubke & Neale (2008) only 2 and a half pages of the 29 page article are dedicated to the illustration using real data. Secondly, these papers often make it clear that their aim is solely to use the data to illustrate a new technique and that they will leave matters of interpretation to subject matter experts (e.g., Muthen & Asparouhov, 2006). Nonetheless, since many applied researchers will use these methodological papers as 'templates' for their own research with factor mixture models, it will be useful to consider how these papers report their results.

The five methodological papers that used mental health data are summarised in Table 9.1. In addition, one further paper is included that illustrates the use of hybrid models for distinguishing between dimensional and categorical conceptions of mental disorders: this paper was not identified during the literature search, but had previously been found on the Mplus web site (Muthen, 2006).

Only one of the six methodological papers (Kim & Muthen, 2009) makes no mention of the fact that latent classes in a mixture model may play indirect roles. This is also the paper that makes the strongest interpretation of the classes found during the illustrative analysis as representing real groups. The authors of this paper do not seem to have considered the possibility that their classes may be performing some form of indirect role.

Table 9.1: Summary of the reporting of factor mixture models applied to symptoms of mental disorders in methodological papers

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Kim & Muthen (2009)	Aggressive behaviour in children	Two-part FMM	No	Yes	"The proposed model revealed otherwise unobserved subpopulations among the children in the study in terms of both their tendency toward and their level of aggression." (p. 603)[1]
Lubke & Neale (2008)	Depression	SP-FA and FMM	Non-linearity and non-normality	Yes	"In this model, the loading estimates in the higher scoring class show much more variability across items than in the lower scoring class. On a conceptual level this would mean that in the class of the participants with higher levels of depression, the 10 items vary more with respect to how well they discriminate than in the class of unaffected participants." (p. 604)[1]
Lubke & Tueller (2010)	Depression	SP-FA and FMM	Non-normality	Yes	"More realistic is a model that imposes a structure reflecting mainly zero scores in the unaffected class, and a single-factor model in the class that contains subjects with severity differences in depression." (p. 615)[1]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]Taylor & Francis, [2]John Wiley and Sons, [3]Elsevier.

Continued on next page...

Table 9.1 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Masyn *et al.* (2010)	Delinquency	LCFA, SP-FA and FMM	Non-normality: identifies that LCFA and SP-FA can model a non-normal latent trait	No	NA
Muthen (2006)	Alcohol dependence and abuse	LCFA and FMM	Non-normality: describes LCFA to model a non-normal factor.	Yes	"The FMA version reported here is the one that focuses on a clustering of subjects, not a representation with measurement invariance and a single dimension for all individuals . . . A class with very low probabilities of endorsing items contains 81% of the subjects . . . The high 19% class contains individuals who have varying degrees of problematic alcohol involvement." (p. 14)[2]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]Taylor & Francis, [2]John Wiley and Sons, [3]Elsevier.

Continued on next page...

Table 9.1 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Muthen & Asparouhov (2006)	Tobacco dependence	FMM and LCFA	Non-normality: describes LCFA to model a non-normal factor.	Yes	"The factor mixture analysis does not suggest that the high class 1 should necessarily be seen as the critical group to be reported as tobacco dependent. Indeed, 58% of current smokers would seem a high number . . . The important point is that factor mixture modeling uncovers a heterogeneous latent variable structure that fits the data well and that sheds more light on the tobacco dependence phenotype. This is useful for guiding substantive experts in their understanding and decisions." (p. 1064)[3]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]Taylor & Francis, [2]John Wiley and Sons, [3]Elsevier.

All of the other five papers mention at least one of the indirect roles that the latent classes can play. Masyn *et al.* (2010), Muthen (2006) and Muthen & Asparouhov (2006) all mention that specific types of model from the broader family of factor mixture models (semi-parametric and latent class factor analysis) can be used to accommodate non-normality of the latent trait. However, this may leave the reader with the misleading impression that when a 'true' factor mixture model fits the data better than a semi-parametric or latent class factor model the classes must therefore represent real groups; it is not clear whether or not the authors of these papers hold this view. While Masyn *et al.* (2010) select a final model without latent classes (so class interpretation is not an issue), both Muthen (2006) and Muthen & Asparouhov (2006) are somewhat vague in their interpretation of the classes that they find. Muthen & Asparouhov (2006) note that their classes do not appear to correspond to a neat division between those who are tobacco dependent and those who are not:

> The factor mixture analysis does not suggest that the high class 1 should necessarily be seen as the critical group to be reported as tobacco dependent. Indeed, 58% of current smokers would seem a high number . . . [p.1064][1]

Nonetheless, it is not clear how the latent classes *should* be interpreted. Muthen & Asparouhov (2006) appear to believe that their discovery of latent classes is scientifically useful:

> The important point is that factor mixture modelling uncovers a heterogeneous latent variable structure that fits the data well and that sheds more light on the tobacco dependence phenotype. This is useful for guiding substantive experts in their understanding and decisions. [p. 1064][1]

However, it is far from clear *what* light has actually been shed on the tobacco dependence phenotype by the factor mixture model, and how these results really will help guide substantive experts in their understanding. If the authors are aware that the classes may not represent true groups, they do not mention this; nonetheless, it must be a distinct possibility. Unfortunately, the 'substantive experts' who are supposed to be guided by these results are likely to be unaware that the latent classes in this particular model could be playing an indirect role.

Lubke & Tueller (2010) are clearer that the ability of latent classes to play indirect roles may be problematic for the interpretation of factor mixture models; they actually highlight the ability of latent classes to accommodate non-normality as one

[1] Permission to reproduce these quotations has been granted by Elsevier.

of the limitations of their simulation study:

> When fitting mixture models to data that have within-class non-normality, the additional classes needed to approximate the distribution might not necessarily reflect true subgroups. [p. 624][2]

Nonetheless, this paper makes no mention of the fact that latent classes may also play other indirect roles.

Out of the six methodological papers to apply factor mixture models to mental health data, Lubke & Neale (2008) do the best job of highlighting the fact that latent classes may play indirect roles. As well as mentioning non-normality, they also highlight that latent classes may be extracted to accommodate non-linearity in the relationship between items and the latent trait:

> If group membership is unobserved, the interpretation is less clear due to the indirect application of mixture models. For instance, factor loadings that increase as a function of the underlying factor score can be approximated using a model with several classes and class-specific loadings. [p. 599][2]

Overall, while five out of the six methodological papers make at least some mention of the fact that the latent classes in factor mixture models may play an indirect role, it is concerning that only one of these papers mentions indirect roles other than non-normality. It appears that in several of these papers there is a risk that the latent classes may be misinterpreted (if not by the authors, then possibly by readers). However, since these are primarily methodological papers, their aim was to illustrate a novel methodology rather than to draw substantive conclusions — therefore, it may be less important if some of the latent classes are over-interpreted. It is perhaps more important to investigate how the results of factor mixture models have been interpreted in papers that have applied these models with the aim of drawing substantive conclusions. The next section considers how those papers have been reported.

[2]Permission to reproduce these quotations has been granted by Taylor & Francis.

9.2.3 Substantive papers applying factor mixture models to mental health data

The literature search identified 26 papers reporting the results of factor mixture models applied to the symptoms of mental disorders. These papers are shown in Table 9.2, as well as one further paper that was identified separately (Picardi *et al.*, 2012). This additional paper was, at the time of writing, only available through advance online publication, and therefore had not yet been indexed for inclusion in the electronic search databases; it was found by chance through a routine scan of the latest papers in key psychiatry journals. Three of the 27 papers in Table 9.2 actually focus on investigating the factor structure of the questionnaire (Kim *et al.*, 2010; Schultz-Larsen *et al.*, 2007; Wu & Huang, 2010) rather than investigating the latent structure of the symptoms in question. Nonetheless, these papers are included here because they are using the same models; they therefore face the same problems in deciding whether any latent classes found represent real groups who respond differently to the questionnaire or not.

Table 9.2: Summary of the reporting of factor mixture models applied to symptoms of mental disorders in applied papers

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Almansa *et al.* (2011)	DSM-IV *diagnoses* of internalising and externalising disorders	FMM and SP-FA	No	Yes	"To our knowledge, this is the first time that mentally healthy and ill subpopulation have been jointly analysed and explicitly modelled. . . . in so doing, the models approach closer to reality." (p. 127)[1]
Asmundson *et al.* (2012)	Health anxiety	Factor model then finite mixture model on saved factor scores	No	Yes	"Contrary to current conceptualizations and taxometric findings, the FMM results indicate the latent structure of health anxiety to be taxonic rather than continuous." (p. 246)[2]
Baillie & Teesson (2010)	Alcohol and cannabis abuse and dependence	Factor model with a symptom free 'zero class'	No	No	NA

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Bernstein *et al.* (2010)	Anxiety sensitivity	FMM and SP-FA	Non-normality	Yes	"FMM analyses showed that AS is taxonic and that each AS class has a multidimensional structure." (p. 523)[2]
Elhai *et al.* (2011)	PTSD	SP-FA?	No	Yes	"Results support taxometric work on PTSD by demonstrating that PTSD instruments scores are dimensional rather than taxonic, but that symptom presentations are best represented by multiple latent classes of individuals."(p. 442)[2]
Frazier *et al.* (2012)	Autism spectrum disorders	SP-FA	No	Yes	"The model included both a categorical distinction between youth with and without ASD and dimensional representations of social communication and interaction difficulties and restricted, repetitive behavior." (p. 36)[2]
Gillespie *et al.* (2011)	Cannabis abuse and dependence	FMM?	No	No	NA

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.
Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Gillespie *et al.* (2012)	Cannabis abuse and dependence	FMM?	No	No	NA
Hildebrandt *et al.* (2010)	Body image disturbance	SP-FA	No	Yes	"The results suggest that body image disturbance is more than a simple trait, but rather a mixture of subgroups that are more or less pathological (i.e., highly invested and distressed about appearance) and vary within their subgroup by level of severity." (p. 844)[2]
Hong & Min (2007)	Depression	FMM (Rasch)	No	Yes	"...these results are best interpreted as suggesting where the boundaries of 'subgroups' or various forms of depressive disorders are likely to be, if there are such boundaries." (p. 299)[3]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Kim *et al.* (2010)	Psychological functioning	SP-FA	No	Yes	"Class 1 in the factor mixture model can be interpreted as the psychologically vulnerable class, whereas Class 2 contains the psychologically invulnerable class." (p. 289)[3]
Kuo *et al.* (2008)	Alcohol dependence	FMM?	No	Yes	"...provides empirical evidence for the heterogeneity in AD diagnosis and suggests the reconsidering of the diagnostic system to classify according to, e.g., the 3-class system found here." (p. 112)[2]
Lubke *et al.* (2007)	ADHD	FMM?	No	Yes	"These best fitting models differentiate between the unaffected majority of the cohort and the potentially affected minority, ..." (p. 1591)[2]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Lubke *et al.* (2009)	ADHD	FMM?	No	Yes	"The analysis . . . shows that the samples consist of three latent classes that are located along correlated continua (severe-, moderate-, and low-scoring AP classes)." (p. 1091) ". . . these data argue for considering DSM-IV ADHD as existing on a severity continuum rather than as discrete diagnostic categories." (p. 1092)[2]
McBride *et al.* (2011)	Alcohol abuse and dependence	FMM	No	No	NA
Naifeh *et al.* (2010)	PTSD	SP-FA?	No	Yes	"The results of the current study, however, suggest that it is possible to identify heterogeneity within PTSD at a latent level using FMM. Thus, a more complete understanding of PTSD's structure may necessitate the identification of subgroups for which a particular factor model fits differently." (p. 672)[4]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.
Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Picardi *et al.* (2012)	Schizophrenia	SP-FA	No	Yes	"Our finding that a reliable factor structure for the BPRS could be obtained only after assuming population heterogeneity corroborates the view that schizophrenia is neither a single disease entity nor a circumscribed syndrome, but rather a conglomeration of phenomenologically similar disease entities and syndromes." (p. 392)[2]
Piper *et al.* (2008)	Nicotine dependence	SP-FA	Non-normality	Yes	" . . . the superiority of the fit criteria for the multiple-class FMAs in comparison with the single-class FMA also lends support to non-parametric factor distributions that emerged in the multiple-class FMA solutions as opposed to the bivariate normal factor distribution of the one-class solution." (p. 754)[4]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Ranby *et al.* (2012)	Adult ADHD	FMM	No	Yes	"The two latent classes divided people into a smaller affected class and a larger unaffected class." (p. 262)[5]
Roberson-Nay & Kendler (2011)	Panic disorder	FMM	No	Yes	"Results . . . suggest two panic subtypes, with one subtype characterized by a respiratory component and a second class typified by general somatic symptoms." (p. 2411)[6]
Schultz-Larsen *et al.* (2007)	Cognitive impairment and dementia	FMM (Rasch)	No	Yes	"The final analysis used mixed Rasch models because different Rasch models were expected to apply to the 2 groups of elderly (with and without cognitive impairments) ..." (p. 273)[2]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.
Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Shand *et al.* (2011)	Opioid abuse and dependence	FMM	No	Yes	"In an in-treatment, opioid-dependent sample, there appears to be two classes of individuals exhibiting distinct patterns of abuse and dependence criteria endorsement . . . This study provides preliminary evidence that the proposed DSM-V opioid use disorder distinction between moderate and severely dependent people is valid." (p. 590)[1]
Shevlin *et al.* (2007)	Psychosis-like symptoms	LCFA?	Non-normality	Yes	". . . this pattern of classes could be explained in terms of an underlying skewed continuous distribution of psychosis." (p. 778)[7]
Viroli (2012)	Dementia and cognitive impairment	SP-FA with covariates	No	Yes	"The results we have obtained indicate the actual existence of three clusters of individuals which mainly correspond to the classes of diagnosed normals, CIND [cognitive impairment but not demented], and dementia." (p. 2118)[1]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.

Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

Continued on next page...

Table 9.2 – continued from previous page

Author (year)	Symptom area	Type of model	Mentions alternative roles?	Chooses model with classes?	Authors' interpretation of the role of the latent classes
Walton *et al.* (2011)	Rule-breaking and aggression in adolescence	LCFA	No	No	NA
Wu & Huang (2010)	Depression	FMM (Rasch)	No	Yes	"A distinguishing property of this model is to identify the latent classes of individuals who differ in the use of the response scale, . . . " (p. 162) "These findings indicated that Class 2 used the response scales more adequately and had more of a preference for avoiding an extreme low response category than did Class 1." (p. 163)[3]
Wu *et al.* (2011)	Opioid abuse and dependence	FMM	No	Yes	"The results from the FMM suggested the presence of two groups of non prescribed opioid users, . . . " (p. 659)[2]

Abbreviations: FMM, factor mixture model; SP-FA, semi-parametric factor analysis; LCFA, latent class factor analysis.
Quotations reproduced with permission from the following sources: [1]John Wiley and Sons, [2]Elsevier, [3]SAGE Publications, [4]American Psychological Association, [5]Taylor & Francis, [6]Cambridge University Press, [7]Springer.

The first feature of Table 9.2 that needs highlighting is the range of model types that have been used in these papers. This emphasises the fact that there is no consensus on the terminology for these models, and that two researchers who both say they have used factor mixture models to investigate a particular symptom area may have used very different models. Further inspection of the types of models used reveals an even more pressing cause for concern: in many cases the authors failed to give sufficient information about the way they had specified their factor mixture models for their approach to be clearly classified into one of the model types identified in Chapter 6; in this case, a 'best guess' was made based on information such as reported numbers of parameters in the models, degrees of freedom for models or likelihood ratio tests, or from the types of results that were presented (for example, a chart presenting mean factor scores for each of the classes implies that a semi-parametric or latent class factor model has been used). This is clearly an undesirable situation since results cannot be replicated if other researchers are unable to identify how the original model was specified.

As can be seen in Table 9.2, very few of these papers mentioned the alternative roles that latent classes can play. Three of the 27 papers mention that latent classes can approximate a non-normal factor distribution (Shevlin *et al.*, 2007; Piper *et al.*, 2008; Bernstein *et al.*, 2010). For two of these three papers, this was probably adequate. Shevlin *et al.* (2007) used latent class factor analysis to model a single latent dimension; the classes in this model can only really play one possible indirect role of relaxing the assumption about the distribution of the factor, and the authors accept that their results probably reflect a skewed continuous distribution. Meanwhile, Piper *et al.* (2008) used semi-parametric factor analysis with two correlated latent dimensions. In this paper, it might also have been helpful to highlight the fact that latent classes can accommodate a non-linear relationship between two factors (although this may be considered as simply a specific form of the violation of multivariate normality, which was mentioned by the authors); a scatterplot showing factor scores from all classes would have clarified the shape of the relationship between the two factors. Nevertheless, this paper distinguishes itself by being the only one to explicitly acknowledge how difficult it is to distinguish empirically between categories and dimensions "...even with specialized statistical procedures ..." [p.759]

The third paper to mention the possibility that classes can represent non-normality used a 'true' factor mixture model (Bernstein *et al.*, 2010), so in this case there are a couple of other important indirect roles that the latent classes may be playing and that need to be mentioned. Furthermore, the authors appear to believe that it is the decision to impose parameter invariance across classes that can lead to the extraction of classes playing an indirect role:

> . . . in a mixture distribution, imposing too many restrictions on parameters can result in accepting too many components (i.e., classes). Lower fit indexes [better results in models with too many classes] may result from overly restrictive assumptions because such assumptions may limit the effect of departure from normality in the distribution of questionnaire items, rather than from a distinct subgroup of individuals . . . [p. 519][3]

Unfortunately, the authors seem to believe that allowing parameters to vary across classes eliminates the possibility that the classes can play an indirect role, as the following passage in the paper's discussion indicates:

> Importantly, in the event that [anxiety sensitivity] was indeed a 1-class dimensional variable, alternative FMM findings would have been observed, but were not. Had the 1-class 3-factor (multidimensional) model of AS demonstrated better fit relative to the 2-class 3-factor, then the results would be consistent with taxometric findings documenting AS dimensionality . . . The FMM findings, though, did not support the latter dimensional solution. [p. 524][3]

It therefore seems that the authors of this paper did not actually consider the possibility that the latent classes they found may have been playing an indirect role, nor did they alert readers to this possibility.

Most of the papers listed in Table 9.2 make no mention of the fact that latent classes may play indirect as well as direct roles. In fact, a number of the papers that found latent classes use these to make very strong claims about the 'taxonicity' of the construct being examined without ever questioning the reality of the classes. This is particularly concerning in cases such as Asmundson *et al.* (2012) where the authors admit that their results were not as expected:

> Identification of a large "anxious" class and small "nonanxious" class in this large non-clinical sample was not anticipated. Indeed, it was expected that the "anxious" class, if observed, would be proportionately close to past 12-month and lifetime prevalence rates of clinical forms of health anxiety . . . [p. 250][3]

Despite the pattern of extracted classes being unexpected, the authors of this paper still go on to interpret the smaller of their two classes as representing a real group:

> . . . a small but significant minority is characterized by a qualitatively

[3]Permission to reproduce these quotations has been granted by Elsevier.

> distinct form of low health anxiety denoted by little somatic focus, low levels of worry and fear, and little functional interference due to symptoms. This latter combination may be potentially dangerous in cases where somatic changes and related functional limitations signal actual health threats but are not acted upon (e.g., delaying visit to the doctor) or are dismissed as unimportant . . . [p. 250][4]

In cases such as these where the results were unexpected, failing to explore alternative explanations for the classes may result in a high risk of misleading conclusions.

While some of the papers appear to put *too much* faith in their findings of classes, other papers appear to *discount* their findings of classes without offering any explanation for why a model including classes fits better than purely dimensional models in the first place. For example, after finding that a factor mixture model with three classes was the 'best fitting' model, Lubke *et al.* (2009) claim

> These findings, using the FMM approach, advance the argument that the AP syndrome exists on a severity continuum, . . . [p. 1091][4]

While this may be a very reasonable conclusion, Lubke *et al.* (2009) are vague in their interpretation of the classes, and never say explicitly whether they think the classes imply real groups or not. The implication of their discussion is that *qualitatively distinct* classes resembling existing DSM-IV disorder subtypes would have been given credence, but that quantitatively ordered classes should not (which may be a sensible approach). However, they offer no suggestions as to why a model with classes fits better than a purely dimensional model if these classes do not represent real groups. This suggests that the authors may suspect that the classes are actually playing an indirect role, but that either they are unsure exactly what role this is or for some reason they prefer not to draw attention to this. It is left for the reader to guess what has been gained by fitting these complex hybrid models (and to make up their own minds about whether to interpret the latent classes as real groups).

9.2.4 Summary

To sum up, the reporting of the results of factor mixture models appears largely inadequate. Even in methodological papers where the authors are likely to be relatively knowledgeable about the models they are using, most papers make limited mention of the indirect roles that latent classes can play. In substantive papers

[4] Permission to reproduce these quotations has been granted by Elsevier.

where the final selected model includes latent classes, most authors don't even mention the possibility that latent classes may play an indirect role, let alone investigate whether this may actually be the case for their model. Therefore, there is a real danger that studies using factor mixture models may be misleading readers through the misinterpretation of their model results, and even setting back attempts to improve the system of psychiatric classification.

This low standard may stem from a couple of problem areas surrounding the use and reporting of these models. The first is that while some methodological papers have highlighted the difficulties of interpreting the results of factor mixture models and have recommended checking for alternative explanations (e.g., Bauer & Curran, 2004), no papers have provided any usable recommendations about how this should be done (particularly for binary or ordinal data). The second problem area is that journals may not have access to the necessary expertise to spot shortcomings in the reporting and interpretation of these models. In order to address both of these areas, the following section will suggest a set of clear recommendations on the use, interpretation and reporting of factor mixture models that, if followed, will hopefully minimise the risk of results being misinterpreted.

9.3 Recommendations

Recommendations for the use, interpretation and reporting of factor mixture models to explore latent structure:

1. *Allow sufficient time for interpretation of results.* Researchers should only commit to using factor mixture models to identify latent structure if they are prepared to dedicate considerable time to exploring alternative explanations of their results. If the best fitting model includes latent classes, it may take as long to identify the most likely role that the classes are playing as it did to set up and estimate the factor mixture models in the first place. Researchers who are not prepared to commit this time should stick to simpler approaches.

2. *Be explicit about alternative possible roles of latent classes.* Researchers must be explicit in any papers they write that latent classes in a factor mixture model may play two very different roles: they may represent discrete groups, or they may simply be accommodating violations in the assumptions of the standard factor model.

3. *Report model specifications clearly.* Researchers should describe clearly in the

methods section how their factor mixture models were specified. They should identify which parameters were held constant across classes, which were allowed to vary, and which were fixed at particular values (including any covariances fixed at zero). The total number of parameters estimated for each model should be reported alongside fit statistics in tables; these should tally with the description in the text of which parameters were fixed and which were estimated.

4. *Check for overfitting.* Models with large numbers of class-varying parameters may be prone to overfitting. If possible, the data should be split to check that the same class structure emerges in both halves. A random split provides the most rigorous check, since dividing the data by known characteristics such as gender or using samples from different studies invites the interpretation of discrepancies as relevant group differences. However, this does not eliminate the need for replicating key findings in independent samples and with other measurement instruments.

5. *Examine the response profiles of individuals allocated to each class.* If the final selected model includes classes, it is important that researchers examine the response profiles of individuals actually allocated to classes rather than just relying on class summary statistics when trying to assess whether the classes reflect real groups. Summary statistics may give a misleading impression of the characteristics of individuals in a particular class, and of whether their symptoms do indeed reflect the presence or absence of a particular form of psychopathology. There may be insufficient space to present this level of detail in a journal article, but it could be provided as supplementary material to be made available online if this is an option in the target journal.

6. *Explore alternative roles for the latent classes.* If the final selected model includes latent classes, researchers should check whether the classes appear to be accommodating violations of assumptions in the standard factor model before concluding that the classes represent discrete groups.

   (a) Latent class factor models and/or semi-parametric factor models may be used to explore whether the classes are purely accommodating a violation of the assumption of normality or multivariate normality; if these models appear to fit as well as factor mixture models, that may well be the case. If semi-parametric models do fit well, plots of the overall distribution implied by the mixture of normal distributions (as in Figure 6.9 on page 132) or scatterplots of factor scores for all the classes combined should be presented: this will reveal whether there is potential multi-modality (which *may* reflect the presence of discrete groups), or whether the relationship between factors is non-linear.

(b) Exploratory factor analysis may be used to check for possible misfit in the factor structure. Where there is any uncertainty, researchers may incorporate exploratory factor analysis in a confirmatory factor analysis framework (E/CFA) into the factor mixture model to see whether the class structure is sensitive to this addition.

(c) Non-parametric plots of item response functions may be used to check whether the assumed functional form for the relationship between items and the latent trait is appropriate.

7. *Interpret classes clearly and realistically.* If the final selected model includes classes, researchers need to identify in their discussion which role(s) they believe the classes are playing; it is not sufficient simply to report that models with classes fit better than purely dimensional models. If researchers are not sure which role the classes are playing, ideally they will mention this in their discussion.

If researchers follow these guidelines, the risk of factor mixture models being misinterpreted will hopefully be minimised. In addition, the work required to explore possible alternative roles of the latent classes may lead researchers to a much fuller understanding of the data they are analysing and their measurement instrument than they would otherwise have achieved.

Chapter 10

Summary and conclusions

10.1 Summary of findings

The previous chapters have described the results of a range of latent variable models applied to the symptoms of common mental disorders, as measured by the CIS-R interview. The results of both the factor modelling in Chapter 4 and the latent class modelling in Chapter 5 indicated a dimensional structure in the symptoms of common mental disorders. The data appeared to be primarily unidimensional, with all 13 symptoms reflecting a single underlying latent dimension — this latent variable may be thought of as representing a continuum of severity of common mental distress. However, exploratory factor analysis and latent class models with five or more classes also indicated the presence of stable 'minor' dimensions, although these might simply reflect the presence of correlated residuals.

While both the factor and latent class models suggested an underlying dimensional structure for the CIS-R data, the factor models with one, two and three dimensions all showed clear evidence of misfit in terms of the relationships they implied between pairs of symptoms (as was seen in the bivariate fit statistics and residuals). However, the two and three class factor mixture models described in Chapter 7 were noticeably better at replicating the relationships between pairs of symptoms (as was indicated by their much improved bivariate fit statistics and residuals). Since these factor mixture models seemed to provide a good fit to the data while requiring far fewer parameters than the latent class models with four or more classes (the only other models that could accurately recreate the relationships between all pairs of symptoms), the factor mixture models with two and three classes appeared to provide the best description of the CIS-R data.

Meanwhile, it was noted that latent classes in a factor mixture model may play two very different roles: they may represent true groups in the population (their direct role), or they may simply approximate a continuous underlying distribution that might be difficult to capture with a simpler model (their indirect role). While a clear clinical interpretation of the classes might have provided support for the claim that the extracted classes represented true groups, in this case there was no obvious clinical interpretation of the classes. Moreover, there *was* evidence that one of the key assumptions of the standard factor model for categorical data (the assumption of a logistic functional form) was inappropriate for several symptoms in the CIS-R data. Furthermore, the characteristics of the latent classes appeared to correspond to these violations in the factor model. Finally, data simulated to mimic the observed violations of the logistic functional form confirmed the extraction of 'spurious' latent classes in this scenario. The author therefore concluded that the latent classes in the factor mixture models were unlikely to represent real groups of individuals with distinct patterns of symptoms, but were instead accommodating misfit in the standard factor model.

The key aim of the thesis was to investigate whether factor mixture models could be used to suggest a data-driven classification of mental disorders. However, careful consideration of the results of the factor mixture modelling suggested that these results could tell us little about the true nature of common mental disorders. Although no discrete groups were identified, it was noted that 'absence of evidence is not evidence of absence': there were several key reasons why discrete disorder categories may have been missed, including the potentially problematic effect of noise from the factor model misfit. Similar considerations applied to the secondary thesis aim of looking for evidence of whether depression frequently occurs as a 'pure' condition without symptoms of anxiety.

Nonetheless, it was noted that the findings of this thesis have important implications for other researchers who use latent variable models to investigate the latent structure of mental disorders. In particular, the thesis highlighted the difficulties of interpreting the results of factor mixture models. Furthermore, the thesis revealed the real possibility that many existing papers that have used these models to investigate the structure of mental disorders may have misinterpreted their findings. A set of recommendations were provided on the reporting of factor mixture models that may help to minimise the risk of future misinterpretation.

10.2 Further work

Given the current popularity of factor mixture models for exploring the latent structure of symptoms of mental disorders, it is likely that stronger evidence will be required concerning the limitations of these models if researchers are to be persuaded to stop using them for this purpose. The brief simulation study reported at the end of Chapter 7 confirmed that violations of the assumed logistic functional form *can* result in the extraction of 'spurious' latent classes, and it seems likely that such a scenario is occurring in the real CIS-R data. However, further simulation studies are needed to investigate how severe such violations need to be to result in the extraction of additional classes, and how this would relate to other considerations such as sample size and the existence of real groups in the data.

In addition, further work is needed to develop the methods suggested here for investigating possible indirect roles of the latent classes. An important step will be to confirm in simulation studies that these methods can actually identify the model misspecification responsible when latent classes are playing an indirect role, and whether they can do this under all conditions (such as for the potentially challenging situation involving trivial model misfit but a very large sample size). Once the need for this work has been highlighted, other researchers may also be able to suggest easier or more sensitive methods for identifying when latent classes are playing an indirect role.

10.3 Conclusions

For this particular research project, factor mixture models turned out to be of little use for investigating the latent structure of symptoms of common mental disorders; therefore, these results provide little guidance on how the classification of common mental disorders might be revised. While it is possible that these models may be more effective when applied to other sets of data, they do not appear to be nearly so useful for resolving issues of psychiatric classification as many researchers seem to believe.

While the implications of this project with regard to clarifying the structure of common mental disorders may be minimal, it does have important ramifications for users of latent variable modelling in mental health research more generally. To start with, it is possible that the restrictive nature of popular dimensional models used for exploring the symptoms of mental ill health may be creating pitfalls for unwary researchers, particularly for those investigating the dimensionality or structure of measurement instruments, or for those comparing the properties of instruments across groups. Most importantly, it appears that poor reporting and naive interpretation of factor mixture models may have undermined many recent attempts to explore the latent structure of mental disorders, and may in fact be setting back attempts to revise the psychiatric classification. There is an urgent need for improved rigour in the way that these models are interpreted and reported; the recommendations made in this thesis could help to put scientific conclusions on a much sounder footing.

Appendix A

Issues with combining the CIS-R data from the three surveys

Although the UK adult psychiatric morbidity surveys in 1993, 2000 and 2007 were designed to be a series of three repeated cross-sectional surveys, there were many changes to the form and content of the questionnaires that could cause problems during a pooled analysis. Some parts of the questionnaire (such as the sections on smoking, alcohol consumption and illicit drug use) changed so dramatically between surveys that pooling the data from these sections is impractical. However, the questions covering the symptoms of common mental disorders (the CIS-R interview) remained largely unchanged across the three survey periods, allowing these data to be pooled.

Nonetheless, even in the CIS-R interview there were some minor changes over the course of the three surveys. While these changes were small, it is possible that they might influence the results of a pooled analysis, most notably if the analysis is looking for differences between the three survey periods. This appendix documents three ways in which the CIS-R interview changed over the course of the three surveys: some minor changes to individual questions; changes in the way the interview was administered (and the resulting effect on levels of missing data in the CIS-R interview); and changes in the coding of ICD-10 disorders from the raw CIS-R data. All of these areas may affect comparisons between the three surveys — the first two may affect any analyses of CIS-R data, while the third will only affect analyses using the derived variables relating to ICD-10 diagnoses.

The changes described in this appendix are likely to have had minimal impact on the latent variable modelling described in this thesis, particularly since no attempt was made to compare the latent structure across the three time periods. Nonetheless, the changes are recorded here in order to allow interested readers to form their own judgements about their importance, and in case they may be of use to subsequent researchers attempting to replicate or extend these analyses. Modifications made to eliminate discrepancies in the coding of derived disorder variables may lead to slight differences in the prevalences and counts of the common mental disorders for the pooled data set (which were presented in the summary statistics in Section 3.1.4 of Chapter 3) when compared to those obtained by other researchers who have created their own pooled data set but are unaware of these changes and discrepancies.

A.1 Changes to questions in the CIS-R interview

A.1.1 Section H: Questions H8 and H9 (suicidal thoughts)

In 1993, the two questions on suicidal thinking were included in Section H and were only asked of those with a score of at least one or more on Section G (depression). However, in 2000 and 2007, these questions were moved to form part of a longer section on suicidal thoughts, suicidal attempts and self-harm — this section was placed at the end of the CIS-R interview and was asked of all respondents. The two items contribute to the Section H symptom score (for those with a score of at least one or more on Section G). Of key importance here is that the wording of the questions and response options was changed in 2000, as shown below:

1993 Questions

```
H8.  In the past week have you felt that life isn't worth living?
            Yes
            (Spontaneous:   Yes, but not in the past week)
            No

H9. In the past week, have you thought of killing yourself?
            Yes
            (Spontaneous:   Yes, but not in the past week)
            No
```

2000 and 2007 Questions

```
DSH1. Have you ever felt that life was not worth living?
            Yes
            No
```

```
(If Yes)  Was this ...
            ... in the last week,
            ... in the last year,
            or at some other time?

DSH3. Have you ever thought of taking your life, even if you
would not really do it?
            Yes
            No

(If Yes)  Was this ...
            ... in the last week,
            ... in the last year,
            or at some other time?
```

The corresponding proportions who at H8 reported feeling in the past week that life was not worth living were 26% in 1993, 8% in 2000 and 12% in 2007 (even after excluding those in 2000 and 2007 who would not have been routed to these questions in 1993). The numbers on which these percentages are based are small, so some degree of fluctuation would be expected randomly between the three surveys. However, it seems likely that changes to the question wording may have influenced responses. In 1993, respondents were not given the response option "in the last year" — they were only asked whether or not they had felt this way in the past week. This may have made respondents more likely to give a 'yes' answer when they were actually unable to remember precisely when these feelings occurred — whether it was within the last 7 days, or, for example, 8 or 12 days ago.

Furthermore, the H9 question wording"have you thought of killing yourself?" that was used in 1993 may have been less likely to elicit an honest response than the wording used in 2000 and 2007: "have you ever thought of taking your life, even if you would not really do it?" There may also have been context effects — in 1993, these two questions were asked as part of a section on depressive symptoms and ideas. However, from 2000 onwards these questions were asked at the end of the CIS-R interview, so in between the depression section and these items respondents were also directed to think about anxiety, worry, phobias, obsessions, etc. For these reasons, changes in reported levels of suicidal thoughts between 1993 and 2007 may not represent changes in the prevalence of suicidal ideation over this period. Changes in the Section H symptom score (which includes items H8 and H9) would need to be interpreted with similar caution.

A.1.2 Section I: Worry

In 1993, respondents who were only worried about their physical health were not asked questions I6 to I10, and so could not have a symptom score greater than 0 for this section. However, in 2000 and 2007, those who only worried about their health were still allowed to answer this section. The question wording directed respondents to think about worries other than their physical health, so this should not have made too much difference. However, in the 2000 and 2007 data combined there were 6 cases of mixed anxiety and depressive disorder scoring close to the CIS-R cut-off score of 12 that would no longer have met the criterion for this disorder had the 1993 rule on worry been applied.

A.1.3 Section K: K3 and Phobia comorbidity

In 1993, interviewers were instructed to code one answer only for question K3 "Which of the situations or things listed made you the most anxious/nervous/tense in the past month?" Since this question is used in the algorithms for diagnosing ICD-10 disorders, an individual in 1993 could meet the criteria for, at most, one type of phobia. However, in 2000 and 2007, interviewers could code all items that applied. This means that respondents could meet the diagnostic criteria for two or more types of phobia — 40 individuals met the criteria for all three of social phobia, agoraphobia and specific phobia. This will not affect comparisons of the prevalence of phobias overall. However, in 1993 the three categories of phobia may appear to have a lower prevalence than in later years simply because the questionnaire did not permit more than one type of phobia for each individual.

A.1.4 Sections M and N: Compulsions and Obsessions introductions

In 2000, the introductory texts to questions M1 and N1 were changed. This was because some respondents in 1993 appeared to have misunderstood the questions about repetitive thoughts and actions (for example, respondents specified worrying thoughts rather than true obsessions), and it was thought that the prevalence of true obsessive and compulsive symptoms had been overestimated (Office for National Statistics, 2001, p. 62). The two forms of wording for questions M1 and N1 are shown below:

1993 wording

```
M1.  In the past month, did you find that you kept on doing
     things over and over again when you knew you had already
     done them, for instance checking things like taps or
     washing yourself when you had already done so?

N1.  In the past month, did you have any thoughts or ideas
     over and over again that you found unpleasant and would
     prefer not to think about, that still kept on coming into
     your mind?
```

2000 and 2007 wording

```
M1.  In the past month, did you find that you kept on doing
     things over and over again when you knew you had already
     done them? For example, making your bed or washing your
     hands over and over again.

N1.  In the past month did you have any thoughts or ideas over
     and over again that you found unpleasant and would prefer
     not to think about, that still kept on coming into your mind?
     For example, constantly thinking about death.
```

Table A.1 shows the distributions of symptom scores for sections M and N across the three survey periods. It seems likely that the changes in question wording may have contributed to the reductions in percentages with scores of 1 or more in 2000 and 2007.

Table A.1: Distributions of symptom scores for sections M and N across the three survey periods

	M: Compulsions			N: Obsessions		
Symptom score for section	1993 %	2000 %	2007 %	1993 %	2000 %	2007 %
0	89	96	95	88	93	94
1	4	1	1	2	1	1
2	4	1	2	3	2	2
3	2	2	1	4	2	2
4	1	1	1	3	2	2

A.2 Changes in interview administration and missing data

The method for administering interviews changed between 1993 and 2000 from paper and pencil interviewing (PAPI) to computer assisted personal interviewing (CAPI). Since many major surveys worldwide underwent similar changes in methodology around this time, experimental studies have been embedded within some of these surveys that have attempted to investigate the effects of this change. For example, Baker *et al.* (1995) compared responses on the National Longitudinal Survey for Labour Market Experience Youth Cohort (NLS/Y) where respondents were randomly allocated to receive the interview in PAPI or CAPI form. Although they did find some significant differences to response distributions on a number of questions, they state that "... what is especially remarkable about the overall pattern of results ... is the large number of critical NLS/Y variables that yielded small or insignificant CAPI-PAPI differences" (Baker *et al.*, 1995, p. 429). This suggests that a change from PAPI to CAPI does not have any dramatic general effect on how respondents answer.

Most of the places where they did find differences were on questions where PAPI interviewers were required to manually check answers to previous questions or cover-sheets in order to know which subsequent questions to ask, a process which is prone to error. In this case, they concluded that the CAPI results were likely to have greater validity, since the computer supplied the relevant information automatically. This is particularly relevant to the UK psychiatric morbidity surveys since the PAPI version of the CIS-R interview requires interviewers to do a considerable amount of cross-checking with previous questions and summary sheets, both when working out which questions should be asked and when tailoring the wording of follow-on questions to reflect the particular symptoms that were reported in earlier questions. There are many points in the questionnaire where the route that an interviewer should take depends on the respondent's answers to previous questions.

The patterns of missing data in the CIS-R sections of the UK psychiatric morbidity datasets suggest that interviewers in 1993 did have some difficulty working out which questions should be skipped and which should be asked — blocks of two or more missing items frequently tend to occur immediately following the passages with the most complicated routing instructions. Approximately 12% of individuals in the 1993 dataset have an inappropriate 'skip' in at least one section of the CIS-R interview. In structured interviews such as the CIS-R, a clear advantage of CAPI is that complex routing and skip patterns can be handled automatically by the computer (assuming that the computer software has been correctly programmed). This removes the scope for interviewer error: Baker *et al.* (1995, p. 417) report that CAPI "completely eliminated illegal skips, which account for most of the missing data in the NLS/Y".

However, while the use of CAPI may substantially reduce the problem of missing data resulting from interviewer errors and improve data quality, this difference in the amount of item-missing data introduces a minor discrepancy between datasets obtained with PAPI and CAPI. In the UK psychiatric morbidity survey datasets the diagnostic algorithms for common mental disorders implicitly assume that missing items reflect absent symptoms. Since most of the inappropriate skips meant that individuals who *had* experienced a particular symptom in the previous month were not asked relevant follow-up questions about those symptoms, respondents may have lost the opportunity to report symptoms contributing to section scores or diagnostic algorithms. Therefore, prevalences of these disorders may be slightly underestimated in 1993 compared to 2000 and 2007, since the use of CAPI in the later surveys prevented these inappropriate skips from occurring.

However, when an exploratory analysis was conducted using multiple imputation to 'fill in' the inappropriate skips in the 1993 data (in a way that respected the structure of the CIS-R interview), the changes in disorder prevalences were minimal. Therefore, it seems that the difference in data quality between the 1993 and the two later surveys relating to the mode of interview administration is of little practical importance.

A.3 Coding issues in the archived data sets

A.3.1 Incorrect calculation of depression criteria in 2000

The paper questionnaire used in the 1993 survey was designed in such a way that interviewers could easily calculate CIS-R symptom scores for each section by simply summing the scores of any shaded response codes that they had circled. For this reason, answer codes were not always presented in ascending numerical order. For example, on question G5 the response options were presented as follows:

```
G5. In the past week have you been able to enjoy or take an interest
in things as much as usual?
        Yes................................2
        No/no enjoyment or interest.........1
```

Here, it is the second response "No/no enjoyment or interest" that is indicative of possible depression, and it is this second response that contributes 1 point towards the section G depression symptom score.

However, from the 2000 survey onward the calculation of symptom scores was carried out by a computer, and the numbering of response options reverted to the standard ascending numerical sequence:

```
(1) Yes
(2) No/no enjoyment or interest
```

This means that the numeric coding on this particular item (and on G9 similarly) reversed between 1993 and 2000. It does not seem that this change was taken into account when the CIS-R algorithms were applied in 2000. (However, it seems that the change in G9 *was* taken into account.) I believe that this resulted in some of the depression criteria being incorrectly applied. When a correct version of the CIS-R algorithm is applied in Stata to the raw 2000 data from the data archive, the counts for cases of the following disorders do not match the counts of these disorders provided by the derived variables included in the archived data set: mild depression, moderate depression, severe depression, depressive episode (any severity), mixed anxiety and depressive disorder, and any neurotic disorder. When the CIS-R algorithms are deliberately made incorrect by replacing `g5==2` with `g5==1` in two places, the disorder counts obtained by the algorithm exactly match those from the derived variables in the archived data.

The CIS-R algorithms used in 2000 were not included in the documentation provided along with the data set in the UK Data Archive, so I have not been able to confirm the presence of these errors directly. However, the same pattern of discrepancies also occurred in the first edition of the 2007 data set that was deposited with the UK Data Archive; in this case, the presence of these two errors was confirmed in the CIS-R algorithms included in the documentation for the 2007 survey, and therefore it seems highly likely that the same errors were present in 2000. The data depositor for the 2007 survey (the National Centre for Social Research) was notified of and confirmed the presence of these errors, and the errors in the derived variables for the disorders listed above were corrected in the third edition of the archived data set. This correct version of the 2007 data set was used for the analyses in this project.

The UK Data Archive was also notified of the errors in the 2000 data set. (The 2000 survey was carried out and deposited in the archive by the Office for National Statistics, so staff from the National Centre for Social Research were unable to correct this data set.) Unfortunately, at the time of writing no corrected version of this data set has yet been released. Therefore, for the disorder tabulations given in this project, corrected versions of the CIS-R algorithms have been used (the same algorithms have been applied for all data sets after harmonising the coding of the raw data), and prevalences or case counts of the disorders listed above may differ slightly from those listed in existing publications using the 2000 survey.

A.3.2 Inconsistencies in the coding of derived mental disorder variables

When the 1993 survey was carried out, a hierarchy was created for the psychotic and neurotic disorders so that each individual could be given a primary diagnosis. As part of this, all indicators for neurotic disorders were reset to 0 ('disorder not present') if an individual met the diagnostic criteria for a psychotic disorder, and this remains the case in the archived 1993 data set. However, in the 2000 survey the approach moved towards examining comorbidities: respondents could be given diagnoses of both psychotic and neurotic disorders. This will have inflated counts of respondents with neurotic disorders in 2000 and 2007, compared to the 1993 survey. For example, in 2000 there were 42 individuals identified as having probable psychosis who also met the diagnostic criteria for at least one neurotic disorder — these individuals would have been excluded from analyses of those with common mental disorders in 1993. These inconsistencies between the three survey years have been eradicated for this project by rerunning the diagnostic algorithms *without the code for creating a hierarchy* on the entire combined dataset.

Appendix B

The CIS-R interview: symptoms contributing to the CIS-R score

1. **Somatic symptoms:** aches, pains or any sort of discomfort that were brought on or made worse because the respondent was feeling low, anxious or stressed.

    - Noticed ache or pain/discomfort for four days or more in the past seven days.
    - Ache or pain/discomfort lasted more than three hours on any day in the past week.
    - Ache or pain/discomfort was very unpleasant in the past week.
    - Ache or pain/discomfort bothered respondent when doing something interesting in the past week.

2. **Fatigue:** feeling tired or lacking in energy for any reason other than physical exercise.

    - Felt tired/lacking in energy on four days or more in the past seven days.
    - Felt tired for more than three hours in total on any day in past week.
    - Felt so tired/lacking in energy that had to push self to get things done on at least one occasion during the past week.
    - Felt tired/lacking in energy when doing things that respondent enjoyed (or used to enjoy) at least once during past week.

3. **Concentration/forgetfulness:** problems with concentrating on what they were doing or forgetting things.

- Noticed problems with concentration/memory for four days or more in the past week.
- Could not always concentrate on a TV programme, read a newspaper article or talk to someone without mind wandering in past week.
- Problems with concentration actually stopped respondent from getting on with things they used to do or would like to do.
- Forgot something important in the past seven days.

4. **Sleep:** problems with trying to get to sleep or with getting back to sleep, or sleeping more than usual.
    - Had problems with sleep for four nights or more out of the past seven.
    - Time spent trying to get to sleep or sleeping more than usual: at least a quarter of an hour but less than an hour (1 point), an hour or longer (2 points)
    - Spent three hours or more trying to get to sleep (or sleeping more than usual) on 4 or more of the past seven nights.
5. **Irritability:** feeling irritable or short tempered with those around them (over things that seemed trivial looking back on them).
    - Felt irritable or short tempered/angry on four days or more in the past seven days.
    - Felt irritable or short tempered/angry for more than one hour on any day in past week.
    - Felt so irritable or short tempered/angry that they wanted to shout at someone in the past week (even if they hadn't actually shouted).
    - Had arguments, rows or quarrels or lost their temper with someone in the past seven days and felt it was unjustified on at least one occasion.
6. **Worry about physical health:** worrying about their own physical health (all respondents) or worrying that they might have a serious physical illness (only respondents who didn't report a long-standing illness, disability or infirmity in physical health section of questionnaire).
    - Worried about physical health/serious physical illness on four days or more in the past seven days.

- Felt that they had been worrying too much, in view of their actual health.
- Worrying had been very unpleasant in the past week.
- Not able to take their mind off health worries at least once by doing something else in the past week.

7. **Depression:** feeling sad, miserable or depressed, or not being able to enjoy or take an interest in things as much as usual.
   - Unable to enjoy or take an interest in things as much as usual in the past week.
   - Felt sad, miserable or depressed/unable to enjoy or take an interest in things on four days or more in the past week.
   - Felt, sad, miserable or depressed/unable to enjoy or take an interest in things for more than three hours in total on any day in the past week.
   - When sad, miserable or depressed did not become happier when something nice happened, or when in company.
8. **Depressive ideas** (only asked of respondents who scored 1 or more in the previous Depression section).
   - Felt guilty or blamed self when things went wrong when it hadn't been their fault at least once in the past seven days.
   - Felt that they were not as good as other people during the past week.
   - Felt hopeless, for instance about their future, during the past seven days.
   - Felt that life wasn't worth living in the past week.
   - Thought of killing themselves in the past week.
9. **Worry:** worrying about anything other than their own physical health.
   - Had been worrying about things other than physical health on four or more days out of the past seven days.
   - Had been worrying too much in view of their circumstances.
   - Worrying had been very unpleasant in the past week.
   - Had worried for more than three hours in total on any of the past seven days.

10. **Anxiety:** feeling anxious or nervous, or finding their muscles tense or that they couldn't relax.

    - Felt generally anxious/nervous/tense on four or more of the past seven days.
    - Anxiety/nervousness/tension had been very unpleasant in the past week.
    - When anxious/nervous/tense, had one or more physical symptoms.
    - Felt anxious/nervous/tense for more than three hours in total on any of the past seven days.

11. **Phobias:** feeling anxious, nervous or tense about any specific things or situations when there was no real danger, or avoiding any situation or thing because it would have made them feel nervous or anxious, even though there was no real danger.

    - Felt nervous/anxious about a particular situation/thing four or more times in the past seven days.
    - On the occasions when they felt anxious/nervous/tense, had one or more of the following physical symptoms: heart racing or pounding; hands sweating or shaking; feeling dizzy; difficulty getting their breath; butterflies in stomach; dry mouth; nausea or feeling as though they wanted to vomit.
    - Avoided the situation/thing at least once (1 point) or more than four times (2 points) in the past seven days.

12. **Panic:** anxiety or tension getting so bad that they got in a panic (for example, feeling that they might collapse or lose control unless they did something about it).

    - Anxiety or tension got so bad that the respondent got in a panic once (1 point) or more than once (2 points) in the past week.
    - The feelings of panic were very unpleasant or unbearable in the past week.
    - The worst of these panics lasted longer than 10 minutes.

13. **Compulsions:** finding that they kept on doing things again and again when they knew they had already done them (for example, checking things like taps or washing themselves when they had already done so).

- Found themselves doing things over again (that they had already done) on four days or more in the past week
- Tried to stop doing these things over again during the past week.
- Doing these things over again made them upset or annoyed with themselves in the past week.
- Repeated any of these things three or more times during the past week.

14. **Obsessions:** having thoughts or ideas over and over again that the respondent found unpleasant and that they would have preferred not to think about, that still kept on coming into their mind.
    - Unpleasant thoughts or ideas kept coming into their mind on four days or more in the past week.
    - Tried to stop thinking any of these thoughts in the past week.
    - Became upset or annoyed with themselves when they had these thoughts in the past week.
    - Longest episode of having such thoughts was a quarter of an hour or longer.

Appendix C

Sensitivity analysis for the inclusion of depressive ideas

It is possible that the structure of the CIS-R interview might lead to the extraction of one or more spurious latent classes. The feature of the interview that is of particular concern is that those with a zero score for the depression section (primarily covering low mood/sadness and loss of interest) were not asked the subsequent section on depressive ideas — hence, they are given zero scores on this section. This makes the implicit assumption that those with no depression symptoms have a zero probability of any score on depressive ideas.

Figure C.1 shows an example of an analysis of the CIS-R data that may be being influenced by this implicit assumption. This is a latent class analysis specified to have five latent classes, and it is applied to the pooled CIS-R data from all adults aged 16-64 from all three years of the adult psychiatric morbidity surveys ($n = 22,574$). All 14 symptom scores are included from the 14 sections of the CIS-R interview, and the scores are specified as ordered categorical items with five ordered categories. (While the depressive ideas score actually has six ordered categories, the highest scores are very rare and so the top two categories have been combined into a single category.)

While most of the reported classes in Figure C.1 appear to reflect severity differences in a single pattern of symptoms, class 2 stands out because of its very low probability of symptoms for depression and depressive ideas. Such a distinctive class would clearly be of interest to those interested in the structure of mental disorders — it might be interpreted as support for 'pure' anxiety disorders that are quite distinct from disorders involving depression. However, it is possible that this distinctive class is in fact an artefact of the assumption that has been made about the missing

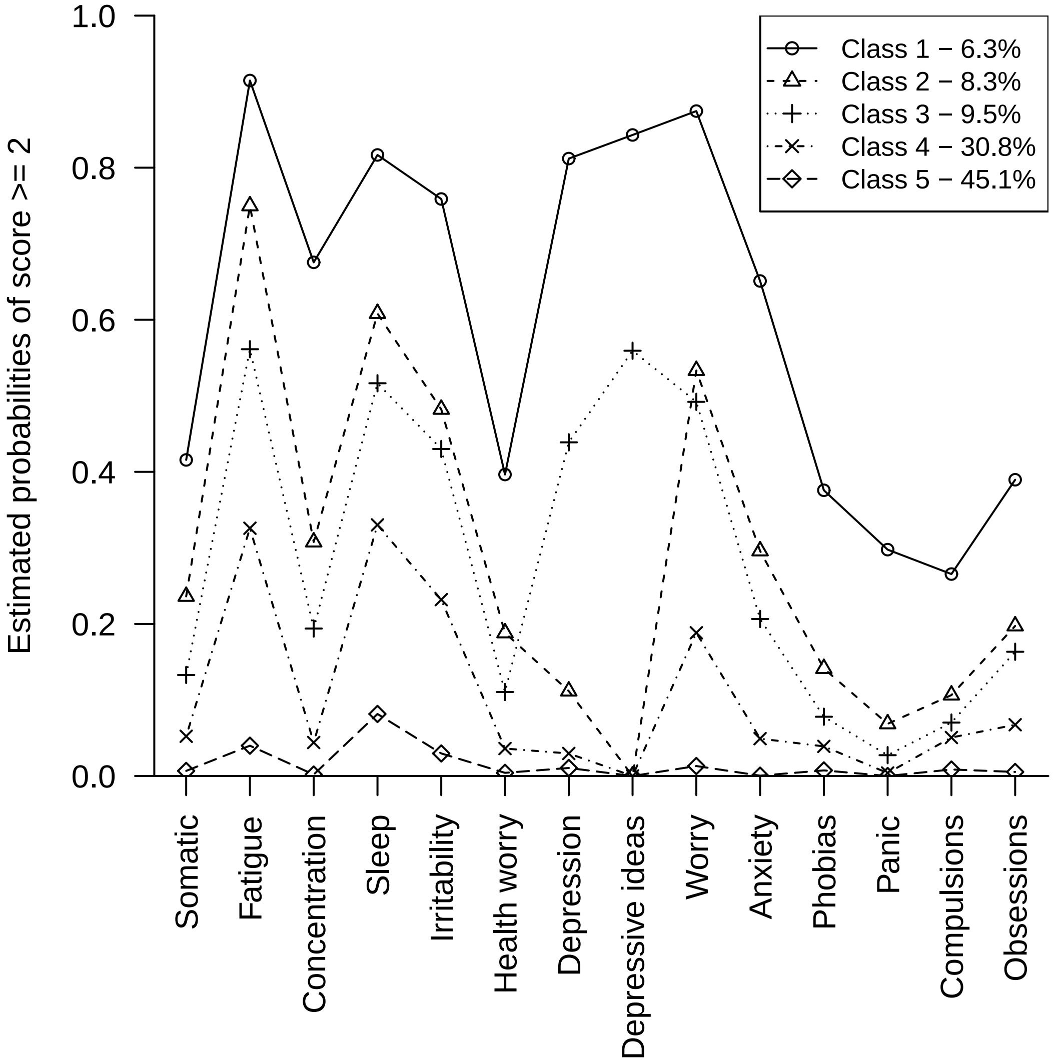

Figure C.1: Latent classes extracted by a five class model applied to the unmodified CIS-R data (assuming that the missing depressive ideas scores are all 0)

depressive ideas scores (particularly since this class follows the same pattern as the other classes for worry/anxiety and other symptoms).

The section on depressive ideas includes items on feelings of hopelessness, guilt and inferiority, as well as two items on suicidal ideation. It therefore seems plausible that individuals with a zero score for depression who reported worry and/or anxiety may still have experienced one or more of the listed symptoms (particularly the feelings of guilt and inferiority, which may be related to anxieties about past events). Therefore, the allocation of a default zero score for depressive ideas may be inappropriate for these particular individuals, and may distort the results of the latent class analysis.

In order to examine the potential importance of the depressive ideas scores allocated to those with a zero score for depression during latent variable modelling, an informal sensitivity analysis was carried out. This involved looking at the sensitivity of the 5 class solution illustrated in Figure C.1 to the addition of a proportion of non-zero scores among those individuals allocated zero scores by default on the depressive ideas section.

C.1 Sensitivity analysis A

Based on the premise that some of the depressive ideas symptoms listed above may be related to worry, the depressive ideas scores were generated randomly for those who were not asked this section based on individuals' observed scores for worry. For sensitivity analysis A, it was assumed that the bivariate relationship between depressive ideas and worry among those with depression scores of zero was the same as the fully observed bivariate relationship for those with depression scores of one. This bivariate relationship is described in Table C.1.

Table C.1: Proportions of each depressive ideas score that were randomly imputed for those not asked this section in sensitivity analysis A. This corresponds to the fully observed bivariate distribution of depressive ideas and worry scores among those with a depression score of 1 (adults aged 16-64 only, depressive ideas scores of 3, 4 and 5 combined into a single category)

	Observed worry score				
	0	1	2	3	4
Depressive ideas score (to be imputed)	%	%	%	%	%
0	54.8	32.5	22.5	13.4	5.9
1	24.2	28.7	24.7	23.4	19.3
2	13.5	22.3	27.9	26.2	34.5
3	7.6	16.5	24.9	37.1	40.3
Total	100.0	100.0	100.0	100.0	100.0

Figure C.2 shows the results of a latent class analysis with five classes applied to these data. As Figure C.2 shows, the 'pure' anxiety class that was distinguished by the absence of depression symptoms has disappeared. Alongside the four classes reflecting severity differences in a single pattern of symptoms, there is now a new class (class 3) that is characterised by fatigue and sleep problems in the absence of symptoms of depression or anxiety.

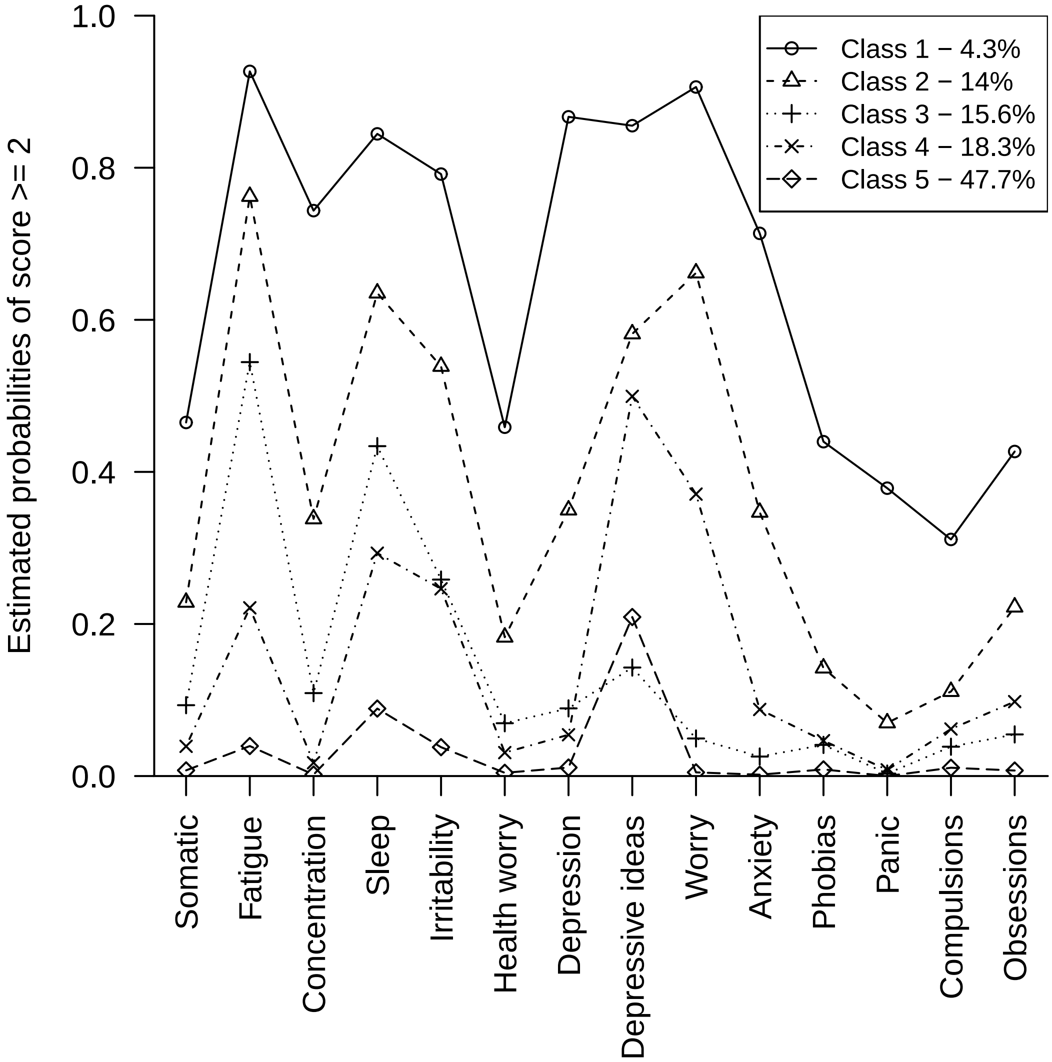

Figure C.2: Latent classes extracted by a five class model applied to the the CIS-R data for sensitivity analysis A (assuming that the missing depressive ideas scores follow the bivariate relationship with worry described in Table C.1)

C.2 Sensitivity analysis B

The assumption that the relationship between depressive ideas and worry is the same for those with depression scores of 0 and those with depression scores of 1 may be unrealistic — individuals reporting mild levels of depression may be more likely to report symptoms such as hopelessness and suicidal ideas than those not reporting any depression. Therefore, in sensitivity analysis B the proportions being randomly allocated each non-zero score for depressive ideas were approximately halved. The resulting bivariate relationship between depressive ideas and worry is shown in Table C.2. This aims to represent a more realistic scenario than that used in sensitivity analysis A.

Table C.2: Proportions of each depressive ideas score that were randomly imputed for those not asked this section in sensitivity analysis B

	Observed worry score				
Depressive ideas score (to be imputed)	0 %	1 %	2 %	3 %	4 %
0	77.3	66.1	61.1	56.6	52.8
1	12.1	14.4	12.4	11.7	9.7
2	6.8	11.2	14.0	13.1	17.3
3	3.8	8.3	12.5	18.6	20.2
Total	100.0	100.0	100.0	100.0	100.0

The latent class analysis for sensitivity analysis B shown in Figure C.3 again shows no evidence of the ‘pure’ anxiety class characterised by the absence of depression and depressive ideas seen in Figure C.1.

C.3 Conclusion

While there is no way to know what the true scores for depressive ideas would have been among those who were never asked this section, this sensitivity analysis has made it clear that the latent class solution is sensitive to whatever assumption is made about these scores. Furthermore, since the class from Figure C.1 that was particularly sensitive to the assumptions about the unknown depressive ideas scores (class 2) is also the class that was most distinctive (and therefore the most exciting from the point of view of interpretation), there is a clear risk of drawing spurious and misleading conclusions from the analysis.

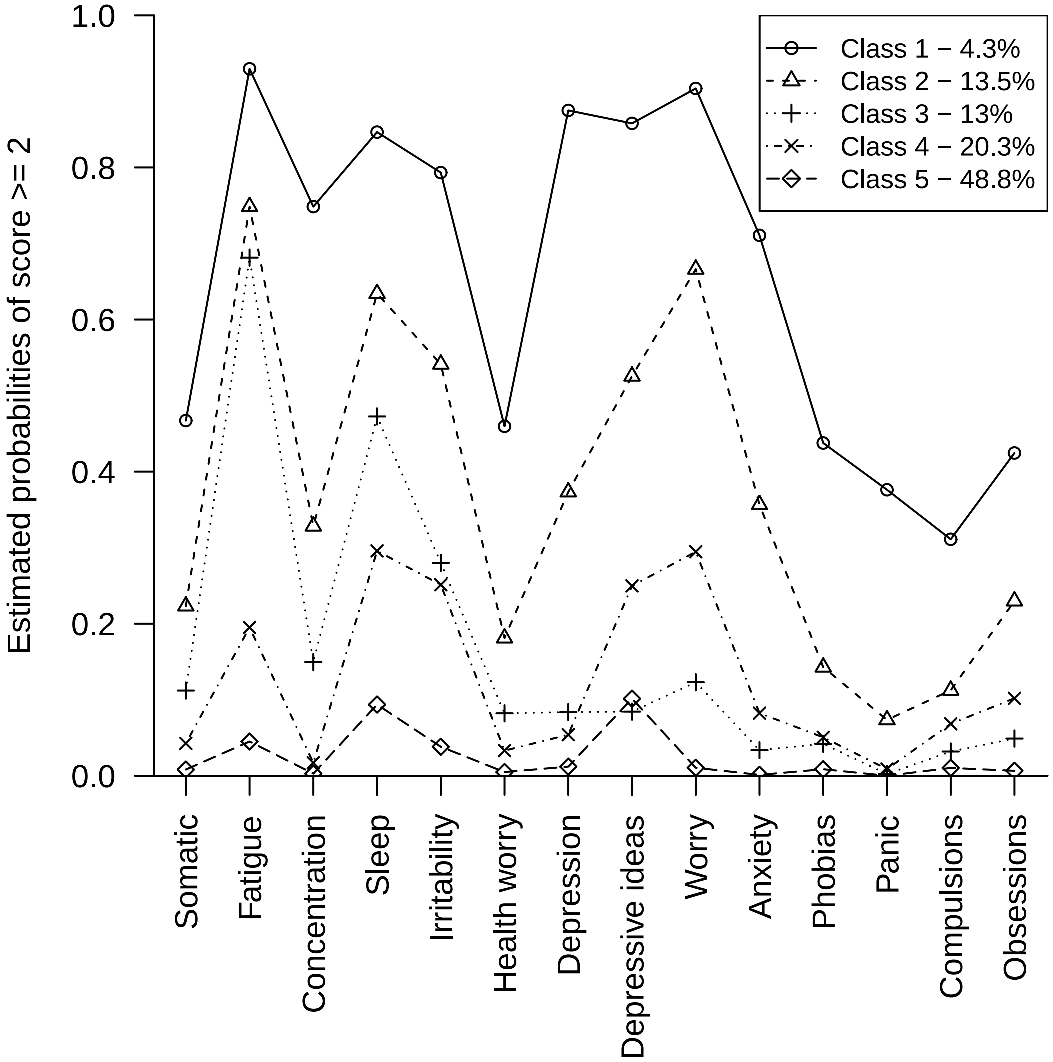

Figure C.3: Latent classes extracted by a five class model applied to the the CIS-R data for sensitivity analysis B (assuming that the missing depressive ideas scores follow the bivariate relationship with worry described in Table C.2)

For this reason, the depressive ideas variable will not be used in the latent variable modelling of this project. This preserves a particular strength of the CIS-R data — that (with the exception of the depressive ideas section) all respondents are asked about every type of symptom, without any filtering based on whether or not they are likely to meet the criteria for a particular disorder. This means that no assumptions are made 'a priori' about which symptoms should occur together and which are unrelated.

Appendix D

Sample Mplus code for the key latent variable models

This appendix shows examples of Mplus code that can be used for running each of the main types of latent variable model discussed in this project. For each type of model, one configuration only is given. However, this can be modified easily to include a different number of classes or factors. To save space, some code details that are unrelated to the specification of a particular type of model are omitted (such as file names and variable lists) — where such code has been omitted, this is indicated by "...".

In the 'Model' sections for the factor mixture models, some of the lines of code that are specified are not actually required since these lines correspond to Mplus defaults in Version 6.1. However, it is safer and clearer to write out the full specification regardless of defaults, particularly since the defaults vary in different versions of Mplus and for different model specifications.

Please note that the code given for these models below is not exactly the same as that used during the project analyses. Some details are omitted for readability or to save space. In particular, starting values for most parameters were specified in the model code for all factor mixture models, but they are not shown here. The original analysis code for all models is available from the author.

All analyses for this project were carried out using Mplus version 6.1 (Muthen & Muthen, 2010a). The Mixture add-on is required in addition to the Mplus base program for all models incorporating latent classes. For further help with setting up an Mplus model or understanding the functions of any of the commands listed below, consult the Mplus manual.

D.1 Standard single factor model

```
Title: Standard single factor model

Data:
  File is ... ;

Variable:
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Auxiliary are id ... ;

Analysis:
  Estimator = MLR;

Model:
  Factor BY somatic@1 fatigue concforg sleep irritab worrheal
     depress worry anxiety phobiasc panicsc compuls obsess;
  Factor*;

Output:
  Svalues; STD; STDY; Tech1; Tech8; Tech10;

Plot:
  Type = Plot3;
```

D.2 Exploratory factor analysis in a confirmatory factor analysis framework (E/CFA)

```
Title: Three factor E/CFA
Data:
  File is ... ;

Variable:
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;

Analysis:
  Estimator = MLR;

Model:
  Factor1 BY somatic* fatigue@1 concforg sleep irritab
     worrheal depress worry@0 anxiety phobiasc@0 panicsc
     compuls obsess;
  Factor1*;

  Factor2 BY somatic* fatigue@0 concforg sleep irritab
     worrheal depress worry@1 anxiety phobiasc@0 panicsc
     compuls obsess;
  Factor2*;

  Factor3 BY somatic* fatigue@0 concforg sleep irritab
     worrheal depress worry@0 anxiety phobiasc@1 panicsc
     compuls obsess;
  Factor3*;

  Factor1 WITH Factor2 Factor3;
  Factor2 WITH Factor3;
```

```
Output:
  Svalues; STDY; Tech1; Tech8; Tech10;

Plot:
  Type = Plot3;
```

D.3 Latent class model

```
Title: Four class latent class model
Data:
  File is ... ;

Variable:
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Classes = c(4);

Analysis:
  Estimator = MLR;
  Algorithm = Integration;
  Type = mixture;
  Starts = 500 300;
  Lrtstarts = 20 5 20 5;

Output:
  Svalues; STDY; Tech1; Tech8; Tech10; Tech11; Tech14;

Plot:
  Type = Plot3;
  Series = somatic-obsess (*);
```

D.4 Semi-parametric factor model

```
Title: Semi-parametric factor model with two classes
Data:
  File is ... ;

Variable:
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Classes = c(2);

Analysis:
  Estimator = MLR;
  Algorithm = Integration;
  Type = mixture;
  Starts = 300 100;

Model:
     %OVERALL%

     factor BY somatic@1;
     factor BY fatigue* (1);
     factor BY concforg* (2);
     factor BY sleep* (3);
     factor BY irritab* (4);
     factor BY worrheal* (5);
     factor BY depress* (6);
     factor BY worry* (7);
     factor BY anxiety* (8);
     factor BY phobiasc* (9);
     factor BY panicsc* (10);
     factor BY compuls* (11);
     factor BY obsess* (12);
```

```
    [ somatic$1* ] (16);
    [ somatic$2* ] (17);
    [ somatic$3* ] (18);
    [ somatic$4* ] (19);
    ...
    [ obsess$1* ] (64);
    [ obsess$2* ] (65);
    [ obsess$3* ] (66);
    [ obsess$4* ] (67);

    [ c#1* ];

    %C#1%

    [ factor* ];
    factor* (14); !Constrains factor variances to be equal

    %C#2%

    [ factor@0 ]; !Only fixed at 0 in one class
    factor* (14);

Output:
  Svalues; STDY; Tech1; Tech8; Tech10; Tech11;

Plot:
  Type = Plot3;
  Series = somatic-obsess (*);
```

D.5 Latent class factor model

```
Title: Latent class factor model with two classes
Data:
  File is ... ;

Variable:
  Names are ... ;
```

```
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Classes = c(2);

Analysis:
  Estimator = MLR;
  Algorithm = Integration;
  Type = mixture;
  Starts = 100 50;

Model:
     %OVERALL%

     factor BY somatic@1 ;
     factor BY fatigue* (1);
     factor BY concforg* (2);
     factor BY sleep* (3);
     factor BY irritab* (4);
     factor BY worrheal* (5);
     factor BY depress* (6);
     factor BY worry* (7);
     factor BY anxiety* (8);
     factor BY phobiasc* (9);
     factor BY panicsc* (10);
     factor BY compuls* (11);
     factor BY obsess* (12);

     [ somatic$1* ] (16);
     [ somatic$2* ] (17);
     [ somatic$3* ] (18);
     [ somatic$4* ] (19);
     ...
     [ obsess$1* ] (69);
     [ obsess$2* ] (70);
```

```
    [ obsess$3* ] (71);
    [ obsess$4* ] (72);

    [ c#1* ];

    %C#1%

    [ factor* ];
    factor@0;

    %C#2%

    [ factor@0 ]; !Only fixed at zero in one class
    factor@0;

Output:
  Svalues; STDY; Tech1; Tech8; Tech10; Tech11;

Plot:
  Type = Plot3;
  Series = somatic-obsess (*);
```

D.6 Factor mixture model: loadings constrained to be equal

```
Title: Factor mixture model with two class and loadings
constrained to be equal across classes
Data:
  File is ... ;
Variable:
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
```

```
     obsess;
  Classes = c(2);

Analysis:
  Estimator = MLR;
  Algorithm = Integration;
  Type = mixture;
  Starts = 2000 500;

Model:
     %OVERALL%

     factor BY somatic@1;
     factor BY fatigue* (1);
     factor BY concforg* (2);
     factor BY sleep* (3);
     factor BY irritab* (4);
     factor BY worrheal* (5);
     factor BY depress* (6);
     factor BY worry* (7);
     factor BY anxiety* (8);
     factor BY phobiasc* (9);
     factor BY panicsc* (10);
     factor BY compuls* (11);
     factor BY obsess* (12);

     [ c#1* ];

     %C#1%

     [ factor@0 ];

     [ somatic$1* ];
     [ somatic$2* ];
     [ somatic$3* ];
     [ somatic$4* ];
     ...
     [ obsess$1* ];
     [ obsess$2* ];
```

```
    [ obsess$3* ];
    [ obsess$4* ];

    factor*;

    %C#2%

    [ factor@0 ];

    [ somatic$1* ];
    [ somatic$2* ];
    [ somatic$3* ];
    [ somatic$4* ];
    ...
    [ obsess$1* ];
    [ obsess$2* ];
    [ obsess$3* ];
    [ obsess$4* ];

    factor*;

Output:
  Svalues; STDY; Tech1; Tech8; Tech10; Tech11;

Plot:
  Type = Plot3;
  Series = somatic-obsess (*);
```

D.7 Factor mixture model: loadings allowed to vary

```
Title: Factor mixture model with two classes and loadings
allowed to vary across classes
Data:
  File is ... ;

Variable:
```

```
  Names are ... ;
  Useobservations are ... ;
  Usevariables are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Categorical are somatic fatigue concforg sleep irritab
     worrheal depress worry anxiety phobiasc panicsc compuls
     obsess;
  Classes = c(2);

Analysis:
  Estimator = MLR;
  Algorithm = Integration;
  Type = mixture;
  Starts = 500 100;

Model:
    %OVERALL%

     factor BY somatic;
     factor BY fatigue;
     factor BY concforg;
     factor BY sleep;
     factor BY irritab;
     factor BY worrheal;
     factor BY depress;
     factor BY worry;
     factor BY anxiety;
     factor BY phobiasc;
     factor BY panicsc;
     factor BY compuls;
     factor BY obsess;

     [ c#1* ];

     %C#1%

     factor BY somatic*;
     factor BY fatigue*;
```

```
factor BY concforg*;
factor BY sleep*;
factor BY irritab*;
factor BY worrheal*;
factor BY depress*;
factor BY worry*;
factor BY anxiety*;
factor BY phobiasc*;
factor BY panicsc*;
factor BY compuls*;
factor BY obsess*;

[ factor@0 ];

[ somatic$1* ];
[ somatic$2* ];
[ somatic$3* ];
[ somatic$4* ];
...
[ obsess$1* ];
[ obsess$2* ];
[ obsess$3* ];
[ obsess$4* ];

factor@1;

%C#2%

factor BY somatic*;
factor BY fatigue*;
factor BY concforg*;
factor BY sleep*;
factor BY irritab*;
factor BY worrheal*;
factor BY depress*;
factor BY worry*;
factor BY anxiety*;
factor BY phobiasc*;
factor BY panicsc*;
```

```
     factor BY compuls*;
     factor BY obsess*;

     [ factor@0 ];

     [ somatic$1* ];
     [ somatic$2* ];
     [ somatic$3* ];
     [ somatic$4* ];
     ...
     [ obsess$1* ];
     [ obsess$2* ];
     [ obsess$3* ];
     [ obsess$4* ];

     factor@1;

Output:
  Svalues; STDY; Tech1; Tech8; Tech10; Tech11;

Plot:
  Type = Plot3;
  Series = somatic-obsess (*);
```

Appendix E

Comparison results from the reserved half of the data

E.1 Factor analysis models

Table E.1: Estimated standardised loadings from the 3 factor exploratory factor analysis on the second random half of the data

Symptom	Factor 1	Factor 2	Factor 3
Somatic symptoms	0.55	0.16	-0.01
Fatigue	0.94	-0.01	-0.15
Concentration/forgetfulness	0.62	0.17	0.04
Sleep	0.50	0.17	-0.01
Irritability	0.32	0.37	0.01
Worry over physical health	0.51	0.13	0.07
Depression	0.34	0.48	0.00
Worry	-0.12	0.94	0.00
Anxiety	0.02	0.86	-0.10
Phobias	0.04	0.00	0.78
Panic	0.06	0.39	0.45
Compulsions	-0.05	0.19	0.45
Obsessions	-0.02	0.45	0.25

E.2 Latent class models

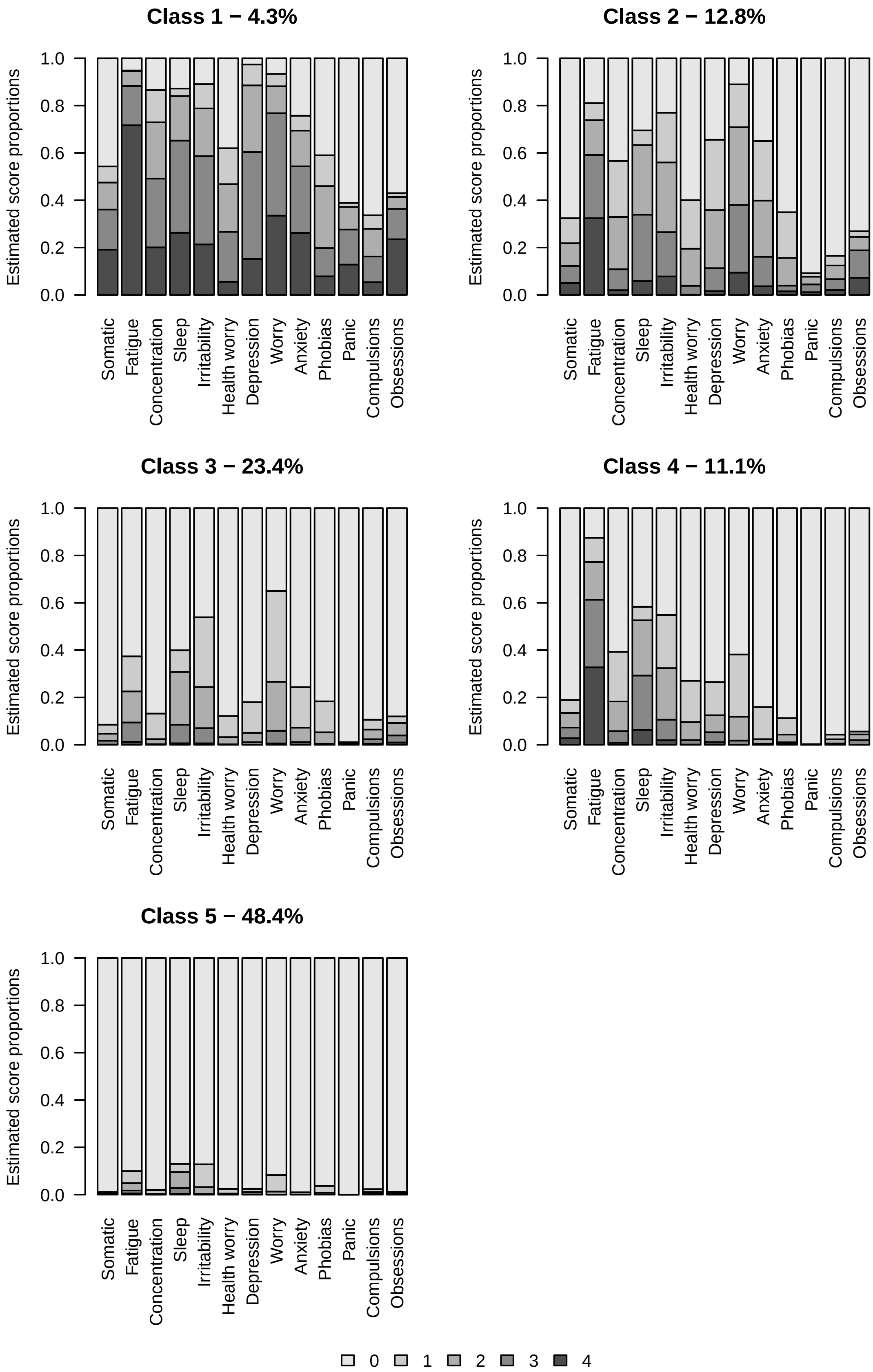

Figure E.1: Estimated probabilities by class for the 5 class latent class model in the second split half of the data

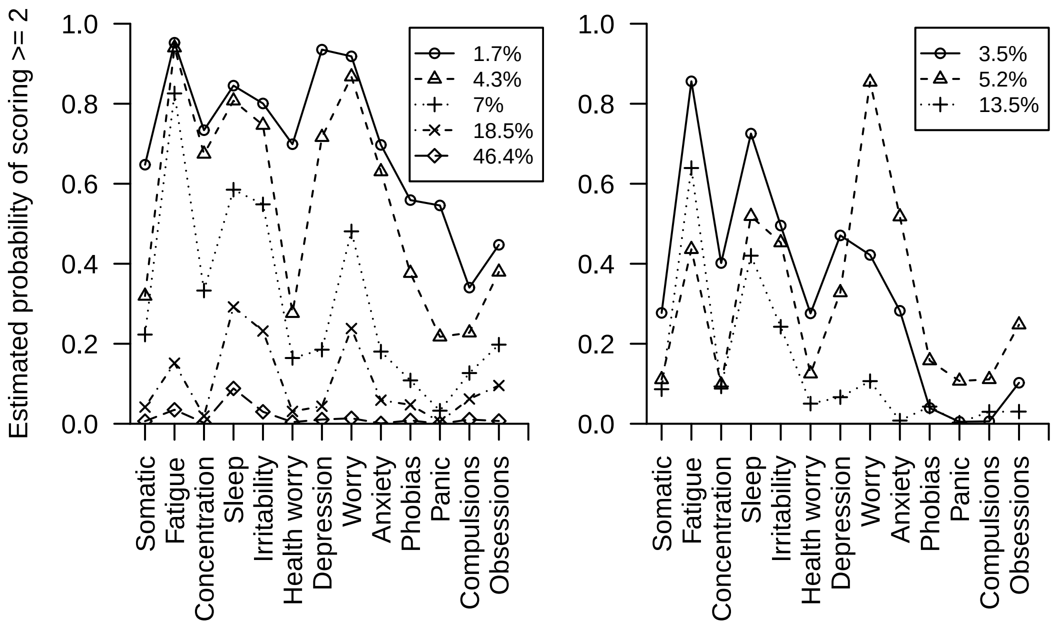

Figure E.2: Estimated probabilities by class for the 8 class latent class model in the second split half of the data

E.3 Semi-parametric factor models

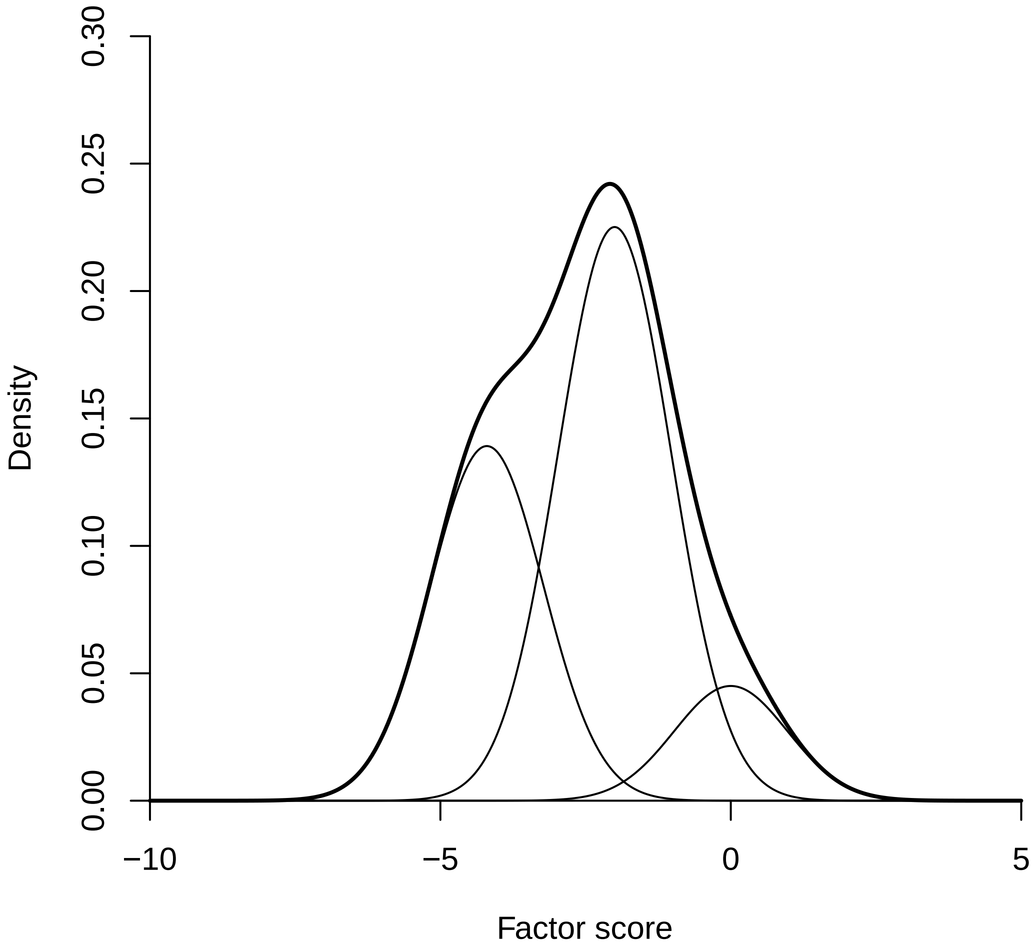

Figure E.3: The mixture model implied by the unidimensional 3 class semi-parametric factor mixture model fitted to the second split half of the data. The variances of the classes were constrained to be equal.

E.4 Latent class factor models

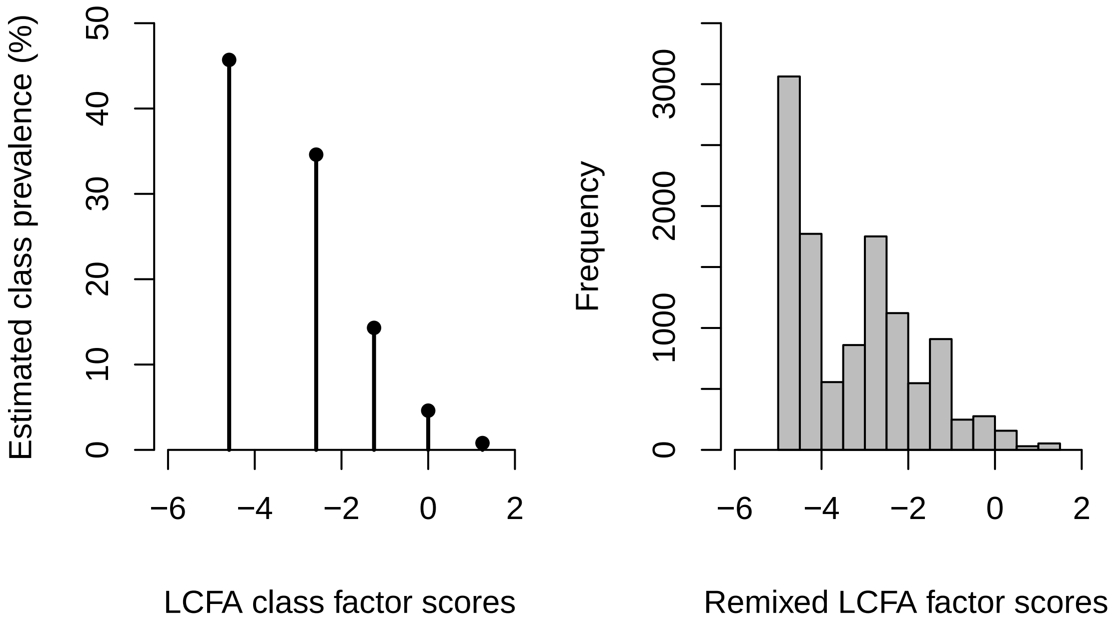

Figure E.4: Factor scores from the 1 dimensional 5 class latent class factor model estimated on the second split half of the data.

E.5 Factor mixture models: equal loadings

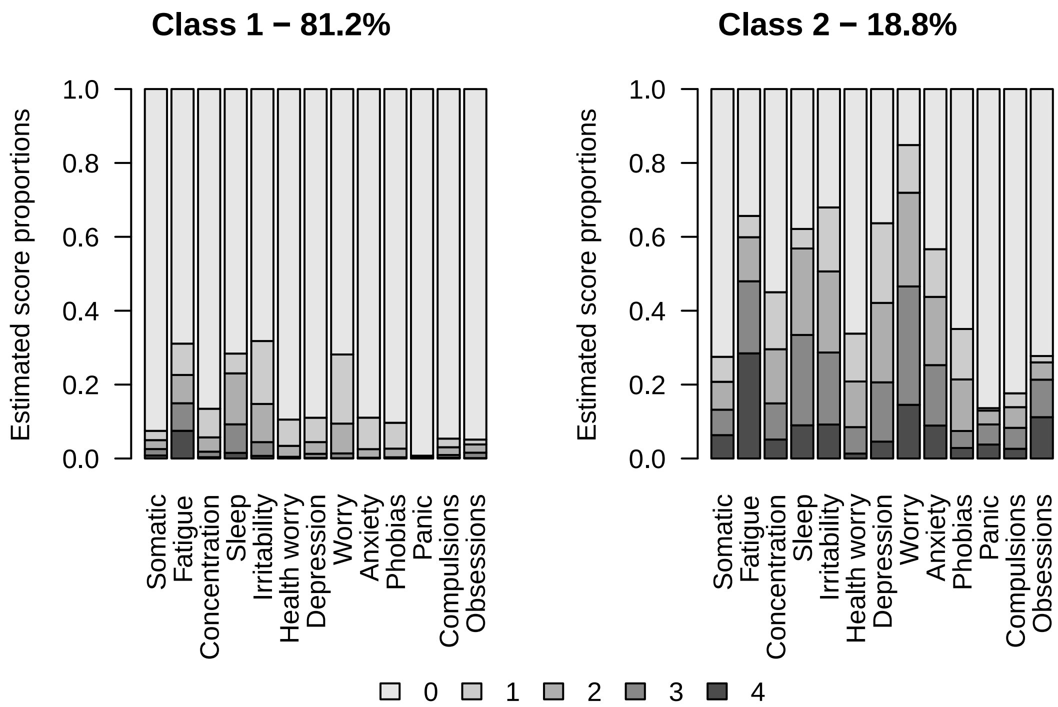

Figure E.5: Estimated score proportions from the 1 dimensional 2 class factor mixture model with intercepts allowed to vary but loadings constrained to be equal across classes, estimated on the second split half of the data.

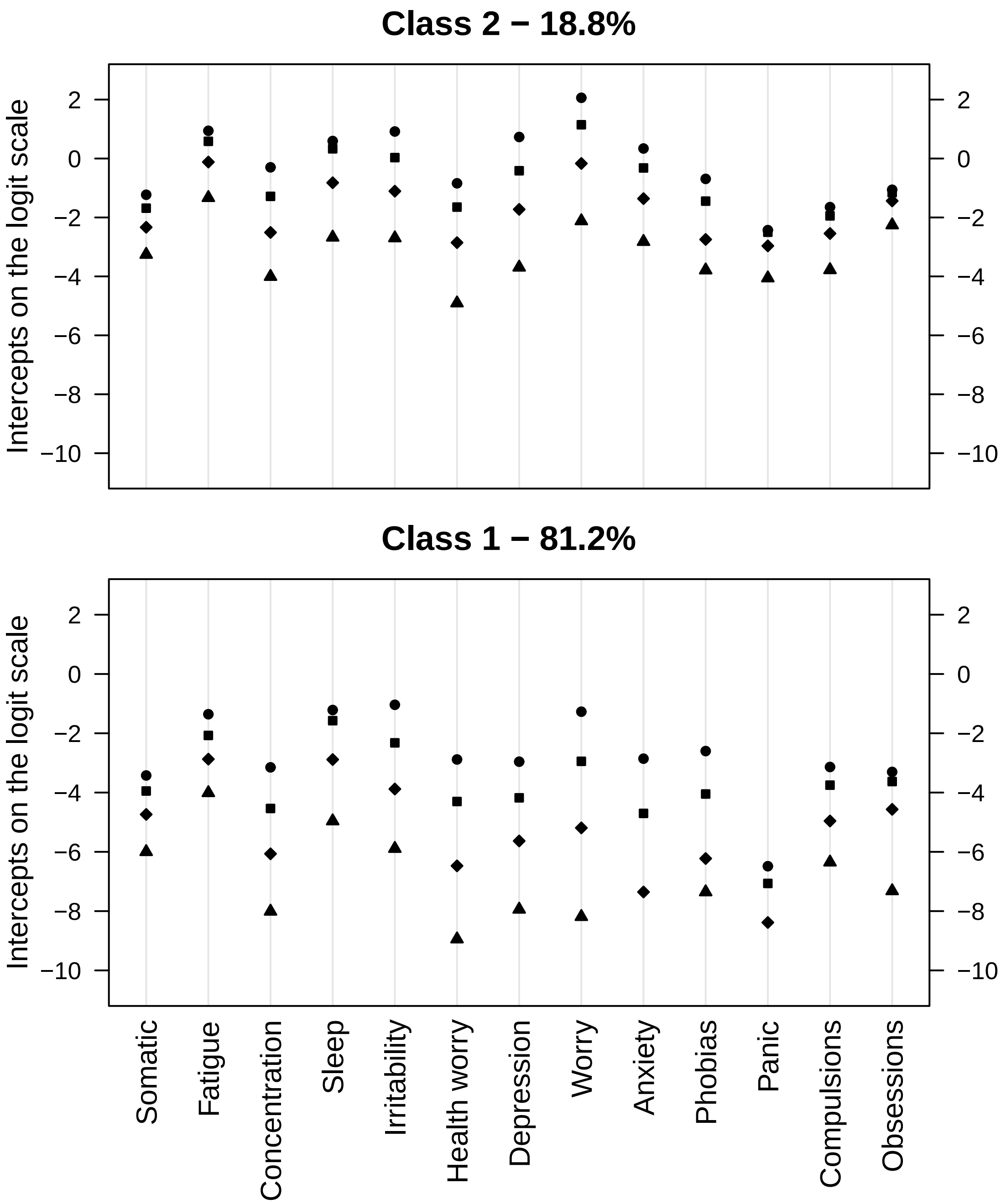

Figure E.6: Intercept parameters for each of the classes in the 2 class factor mixture model, estimated on the second split half of the data. The mean latent trait score for each class was fixed to 0.

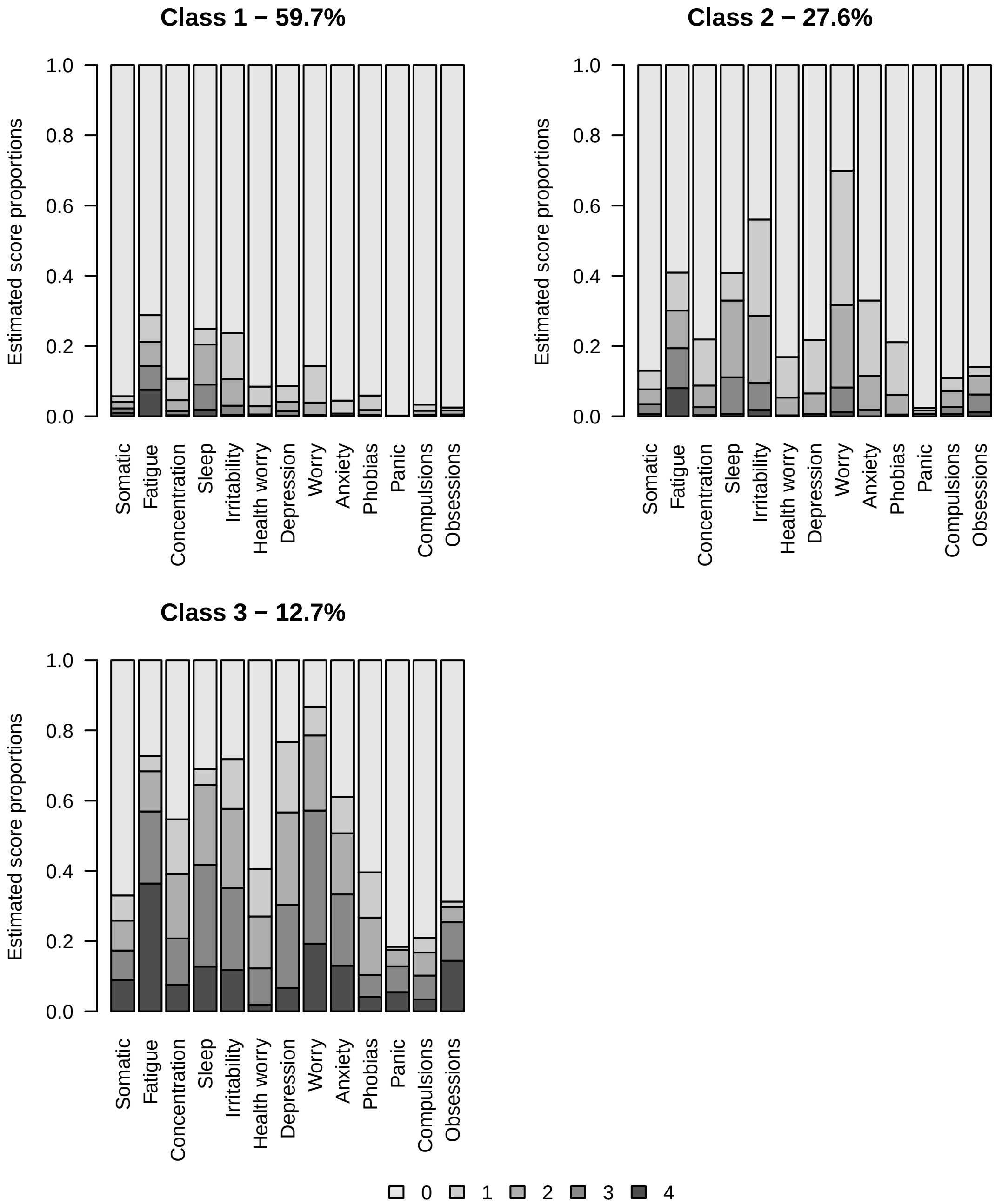

Figure E.7: Estimated score proportions from the 1 dimensional 3 class factor mixture model with intercepts allowed to vary but loadings constrained to be equal across classes, estimated on the second split half of the data.

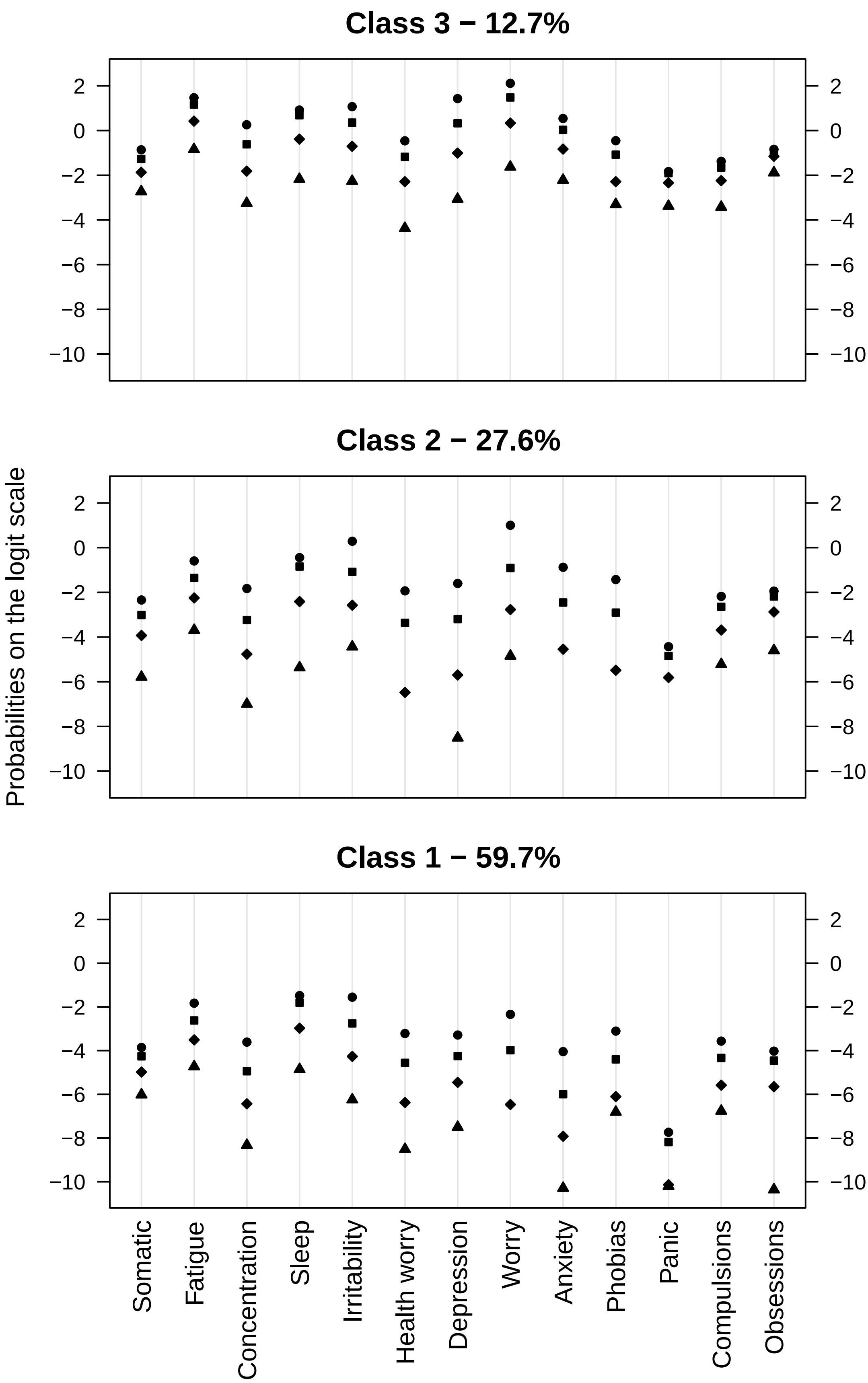

Figure E.8: Intercept parameters for each of the classes in the 3 class factor mixture model, estimated on the second split half of the data. The mean latent trait score in each class is fixed at 0.

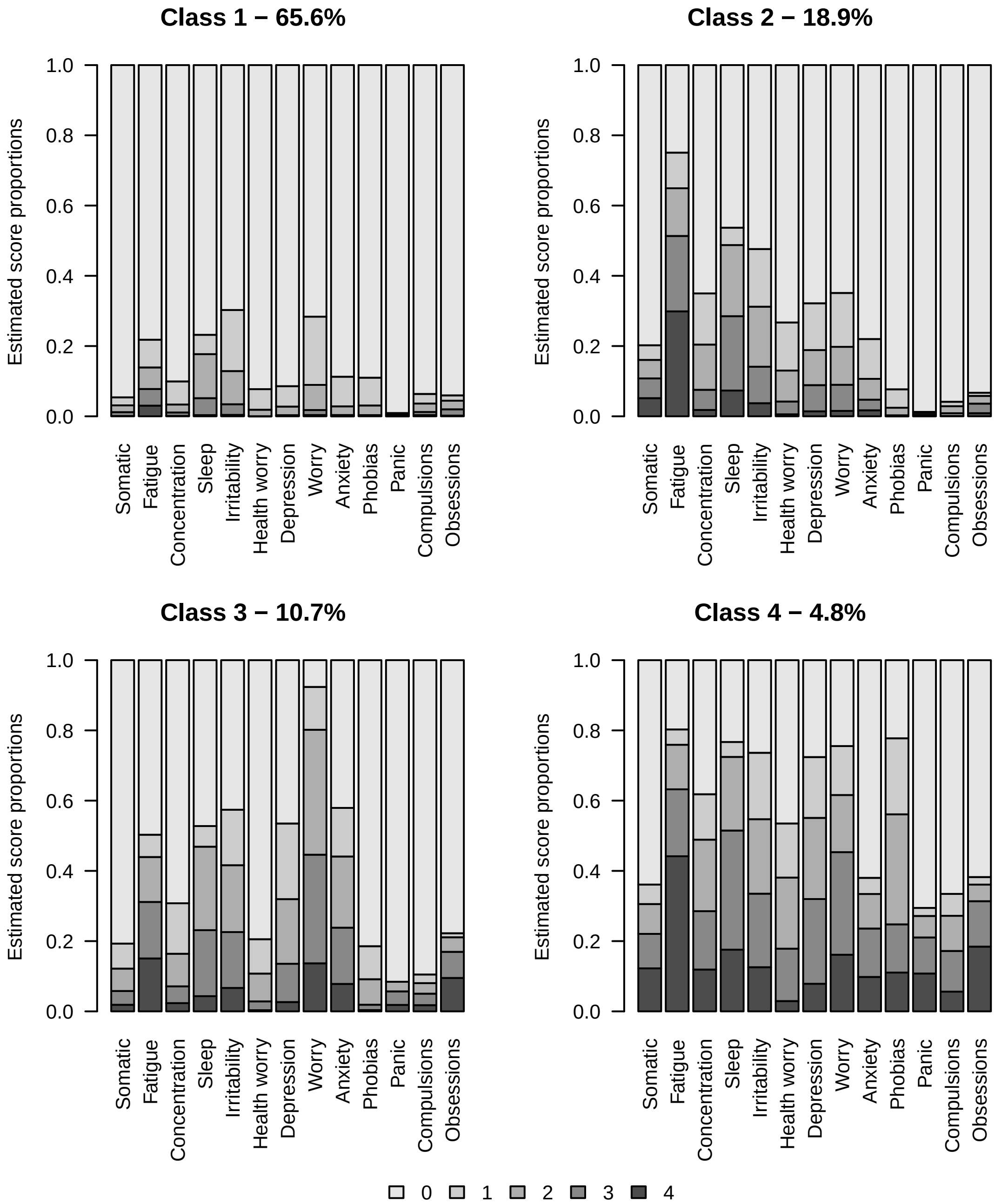

Figure E.9: Estimated score proportions from the 1 dimensional 4 class factor mixture model with intercepts allowed to vary but loadings constrained to be equal across classes, estimated on the second split half of the data.

E.6 Factor mixture models: loadings allowed to differ

Table E.2: Estimated loadings within each class for the 2 class factor mixture model in which loadings and intercepts are allowed to differ between classes — results from the second split half of the data

	Class 1 (83.4%)	Class 2 (16.6%)
Symptom	Loading (SE)	Loading (SE)
Somatic	1.6 (0.1)	1.1 (0.1)
Fatigue	**2.0** (0.1)	**1.8** (0.1)
Concentration/forgetfulness	**2.1** (0.1)	**1.9** (0.1)
Sleep	1.3 (0.0)	1.0 (0.1)
Irritability	1.4 (0.1)	0.9 (0.1)
Health worry	1.5 (0.1)	1.0 (0.1)
Depression	1.5 (0.1)	**1.4** (0.1)
Worry	1.5 (0.1)	0.8 (0.1)
Anxiety	1.8 (0.1)	0.8 (0.1)
Phobias	0.9 (0.1)	0.9 (0.1)
Panic	**2.0** (0.2)	1.2 (0.1)
Compulsions	0.8 (0.1)	0.7 (0.1)
Obsessions	1.1 (0.1)	0.5 (0.1)

Note: Factor variances were fixed to 1 in each class to identify the scale of the latent variable. The largest three loadings in each class are highlighted in bold.

Table E.3: Estimated loadings within each class for the 3 class factor mixture model in which loadings and intercepts are allowed to differ between classes — results from the second split half of the data

	Class 1 (76.2%)	Class 2 (12.1%)	Class 3 (11.7%)
Symptom	Loading (SE)	Loading (SE)	Loading (SE)
Somatic	1.5 (0.1)	0.7 (0.2)	1.1 (0.2)
Fatigue	1.7 (0.2)	0.6 (0.2)	**2.1** (0.2)
Concentration/forgetfulness	**2.0** (0.2)	0.8 (0.3)	**2.1** (0.2)
Sleep	1.1 (0.1)	0.5 (0.1)	1.0 (0.1)
Irritability	1.5 (0.1)	0.7 (0.2)	1.0 (0.1)
Health worry	1.2 (0.1)	1.0 (0.2)	1.0 (0.1)
Depression	1.4 (0.1)	0.9 (0.3)	**1.5** (0.2)
Worry	**2.0** (0.2)	1.2 (0.3)	0.9 (0.2)
Anxiety	**2.3** (0.2)	0.8 (0.3)	0.9 (0.2)
Phobias	1.1 (0.1)	**2.1** (0.5)	0.9 (0.2)
Panic	**2.0** (0.3)	**3.6** (1.4)	1.1 (0.2)
Compulsions	0.9 (0.1)	**2.0** (0.8)	0.6 (0.2)
Obsessions	1.3 (0.1)	1.4 (0.3)	0.5 (0.1)

Note: Factor variances were fixed to 1 in each class to identify the scale of the latent variable. The largest three loadings in each class are highlighted in bold.

E.7 Higher dimensional factor mixture models

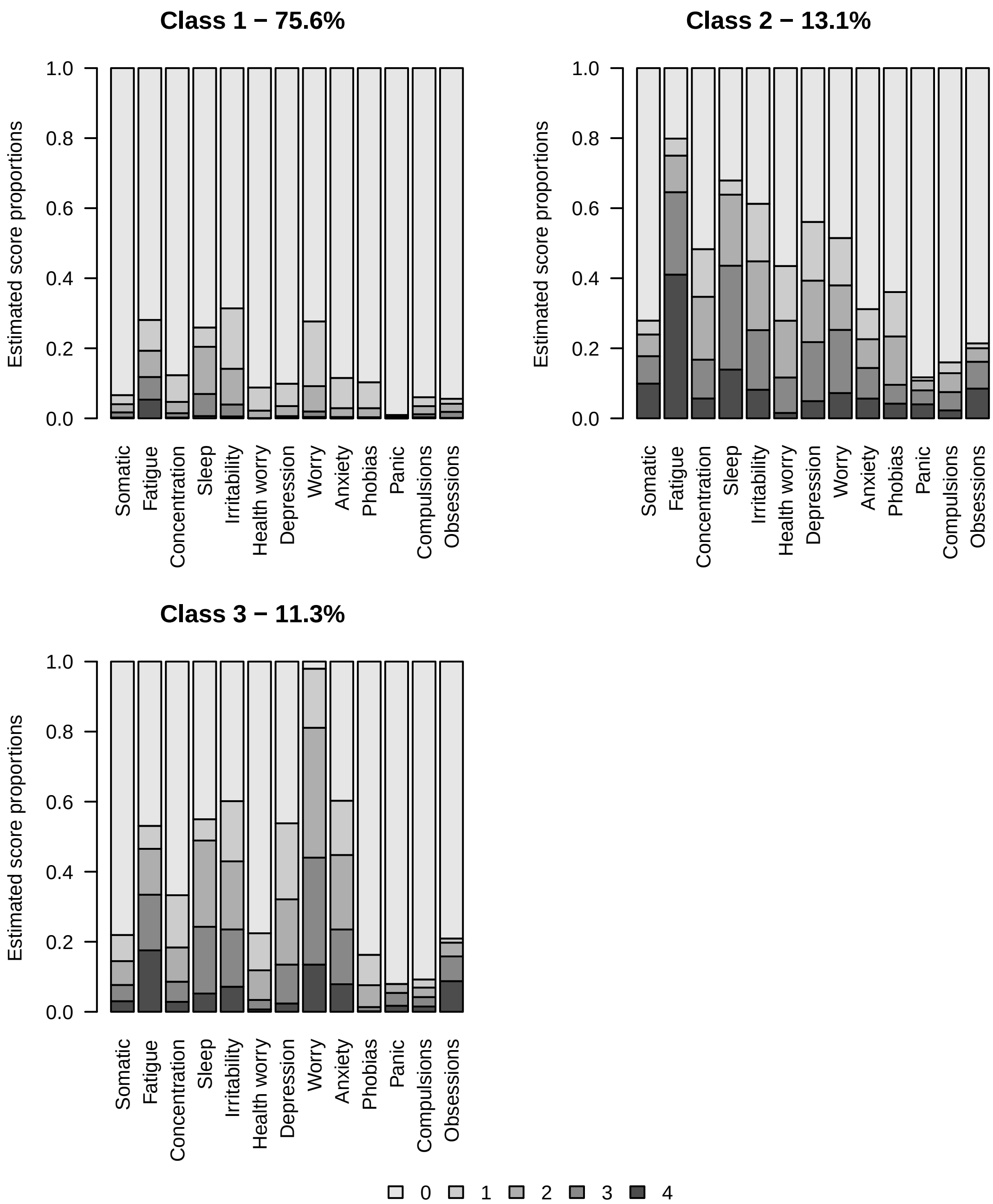

Figure E.10: Estimated score proportions in each class of the factor mixture model with 2 dimensions and 3 classes — results from the second split half of the data

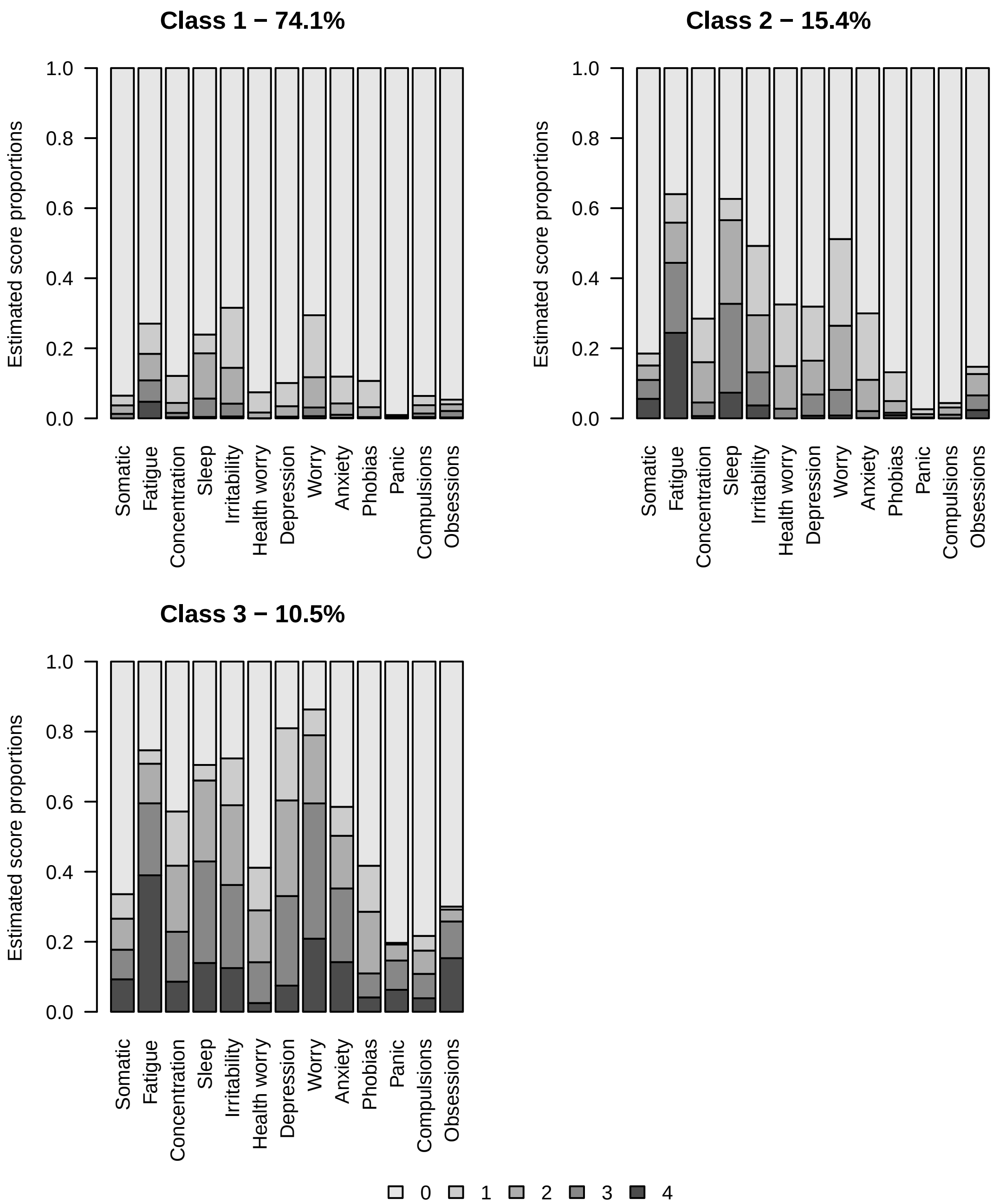

Figure E.11: Estimated score proportions in each class of the factor mixture model with 3 dimensions and 3 classes — results from the second split half of the data

E.8 Nominal quadratic model

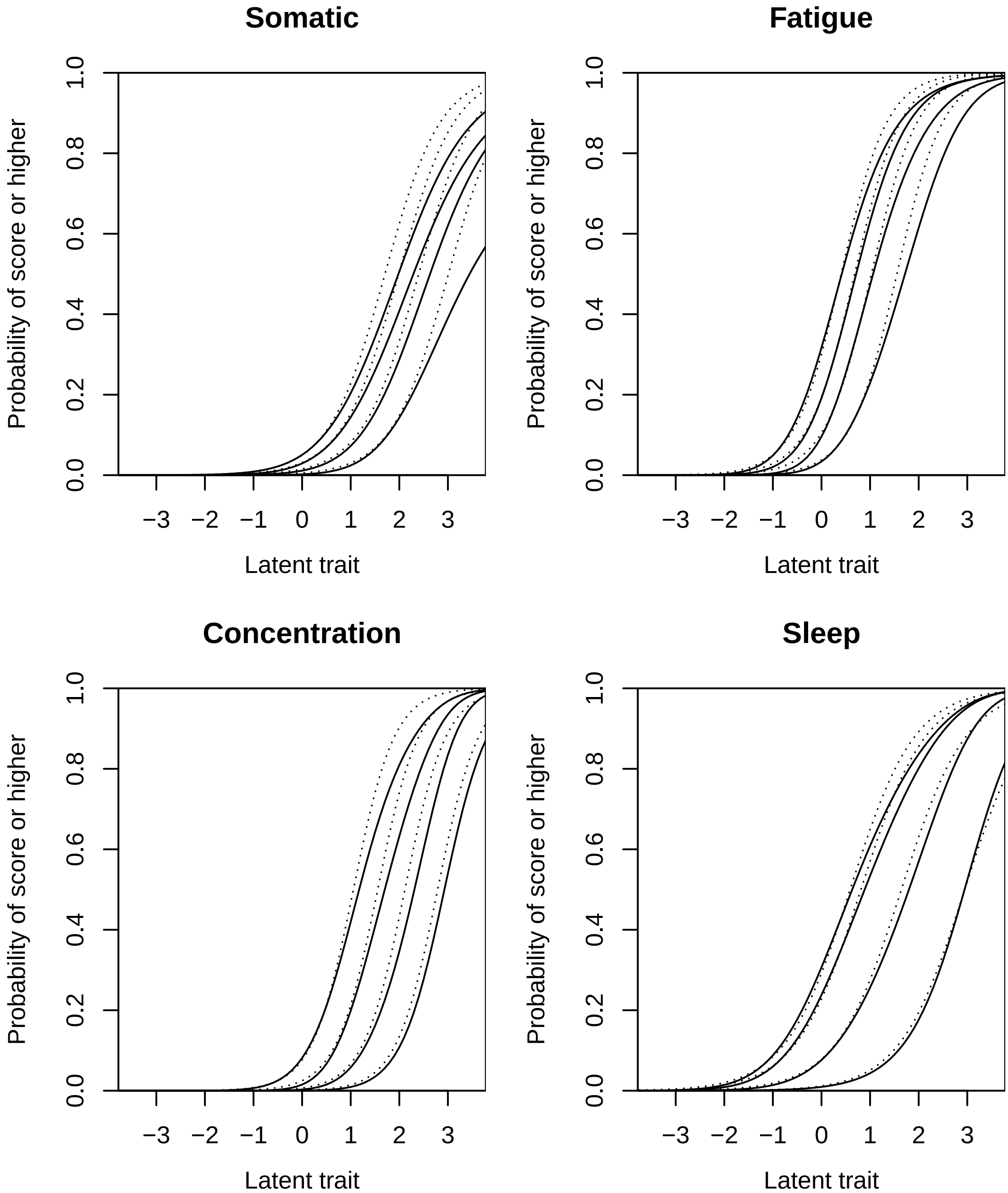

Figure E.12: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

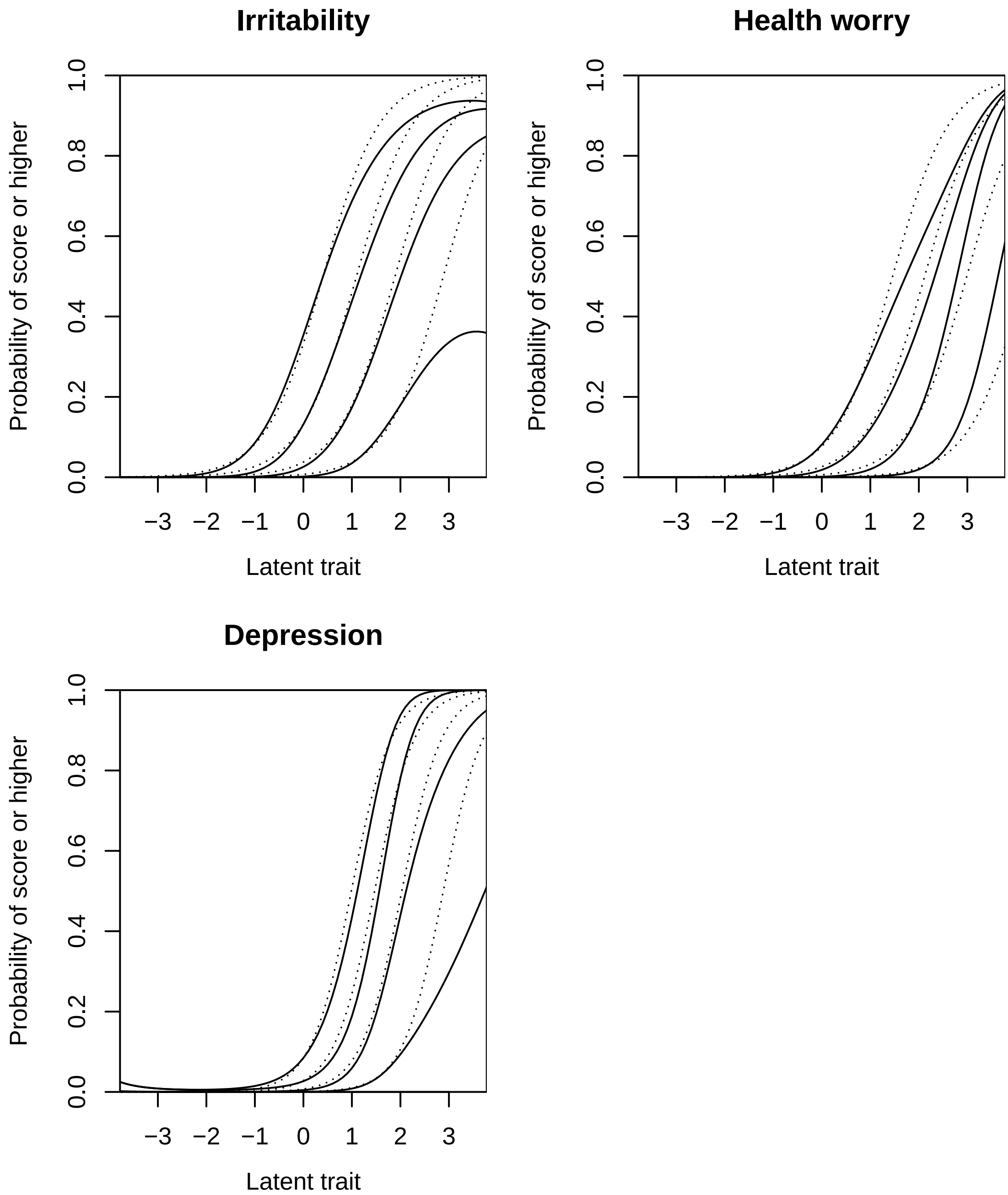

Figure E.13: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

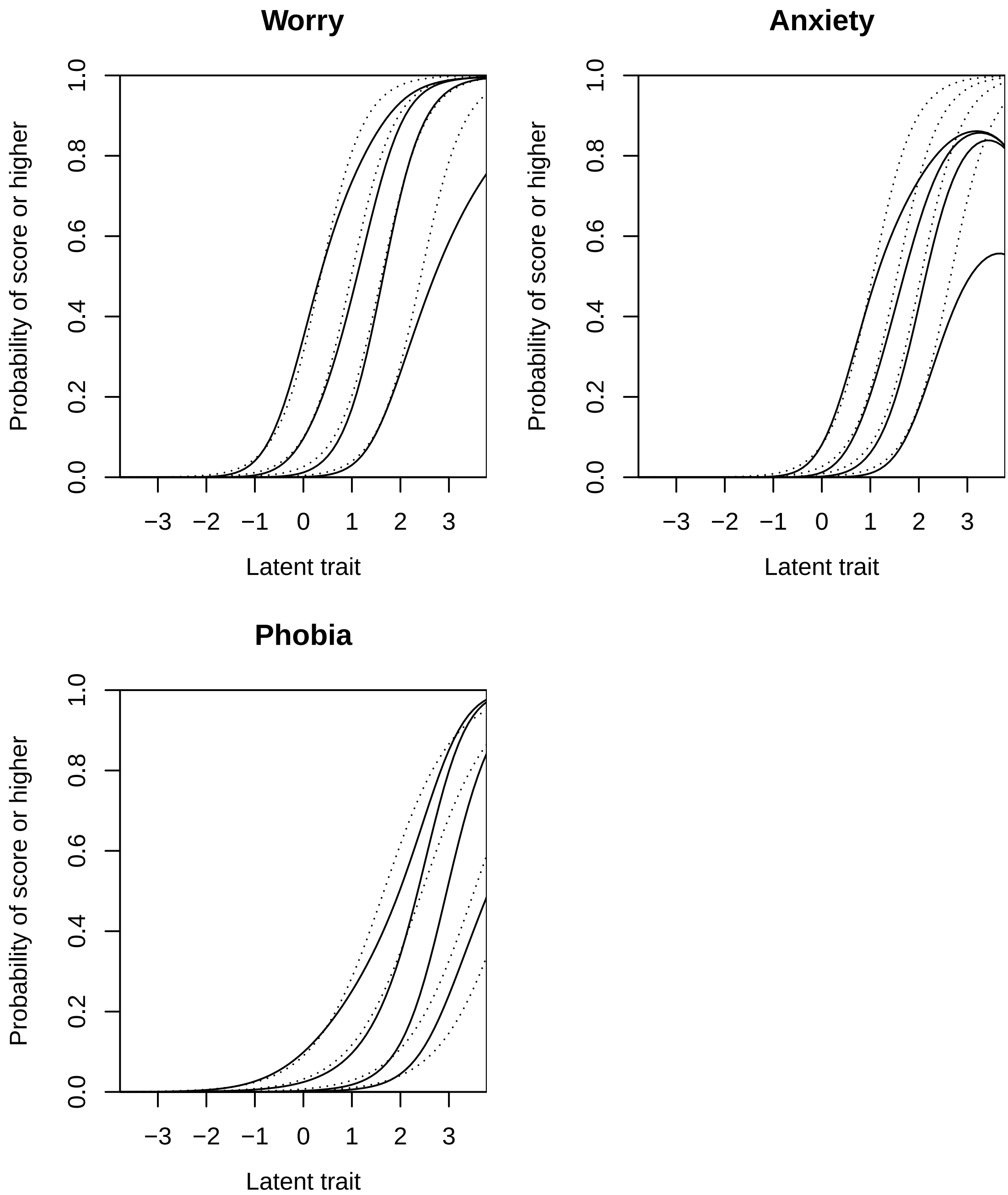

Figure E.14: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

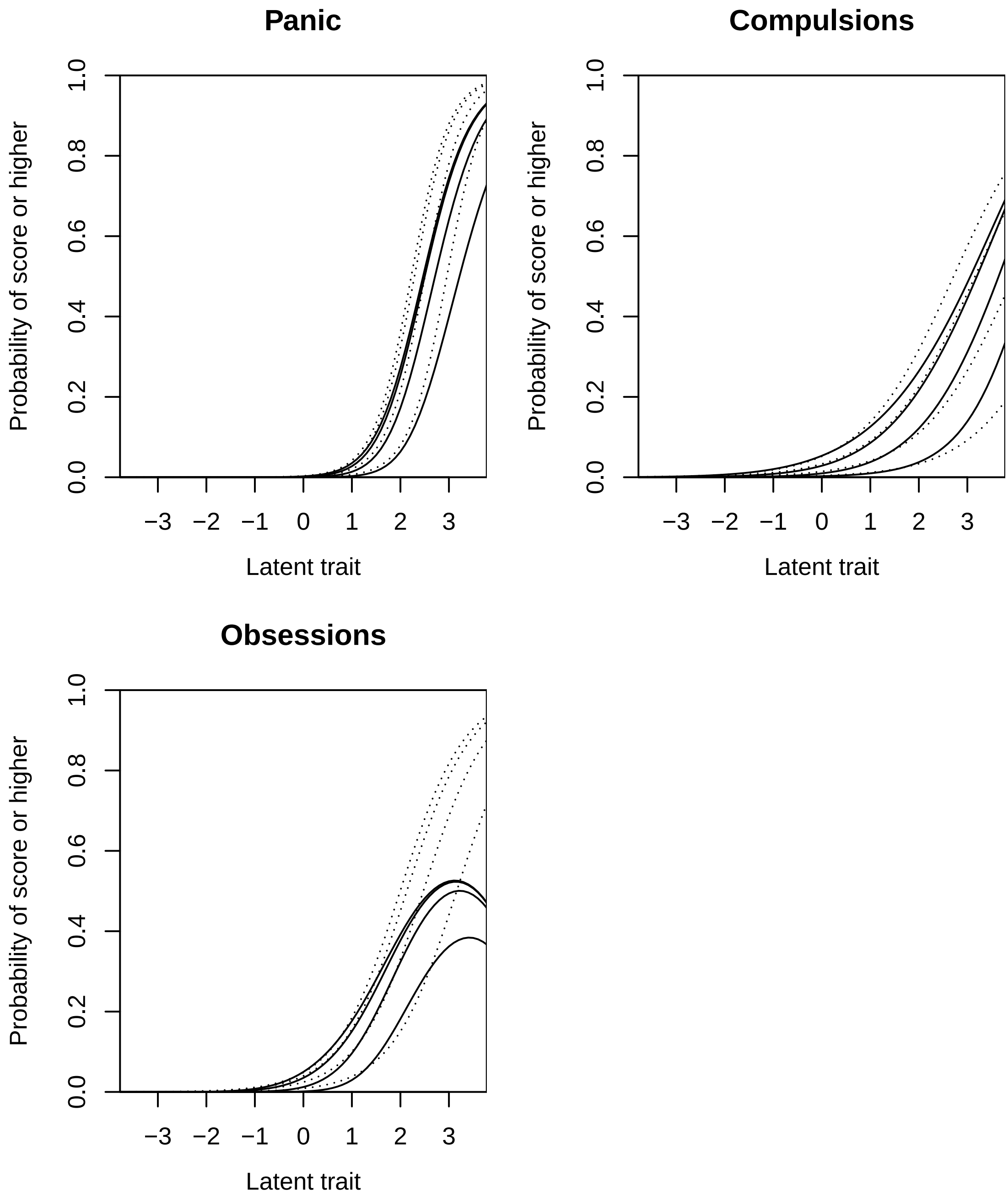

Figure E.15: Response functions from the nominal model incorporating quadratic terms (solid lines) superimposed on the response functions from the standard factor model (dotted lines)

Appendix F

Details of other models fitted to the data

F.1 Non-parametric factor analysis

F.1.1 R code

The idea to use the `sm.binomial()` function of Bowman & Azzalini (2010) came from a function for binary data written by Partchev (2012) in the package 'irtoys'. However, I have extended that idea to produce cumulative plots of ordinal data, as well as making use of restscores rather than sumscores.

This is the basic code that was used to plot the non-parametric response functions in R. (Code to improve the appearance of the plots has been omitted.)

```
## Requires the 'sm' package to be loaded.
##
## 'responses' is a matrix or data frame with one column for each
## item and one row for each individual.
## Each item has 5 ordered categories coded as 0,1,2,3 and 4.
## 'symptom' is the column number of the item to be plotted.

plot.nonparametric <- function(responses, symptom, from=-4, to=4,
      h=NA){
   sum.scores <- apply(responses, 1, sum, na.rm=TRUE)
   rest.scores <- sum.scores - responses[,symptom]
   ranks <- rank(rest.scores, ties.method="random")
```

```
    trait.quantiles <- as.matrix(qnorm(ranks/(length(ranks) + 1)))
    score4 <-as.numeric(responses[,symptom]==4)
    score3 <-as.numeric(responses[,symptom]==4|responses[,symptom]==3)
    score2 <-as.numeric(responses[,symptom]==4|responses[,symptom]==3|
       responses[,symptom]==2)
    score1 <-as.numeric(responses[,symptom]==4|responses[,symptom]==3|
       responses[,symptom]==2|responses[,symptom]==1)
    item.matrix <- cbind(score1,score2,score3,score4)
    plot(c(from, to), c(0, 1), type="n")
    for (i in 1:ncol(item.matrix)){
       if(is.na(h)){
          h <- h.select(x=trait.quantiles, y=item.matrix[, i])
       }
       br <- sm.binomial(x=trait.quantiles, y=item.matrix[, i],
          h=h, display="none")
       lines(br$eval.points, br$estimate)
    }
}
```

F.2 Nominal model

F.2.1 Model description

The nominal model is described by

$$P(Y_{ij} = k) = \frac{e^{\alpha_{jk}\theta_i + \gamma_{jk}}}{\sum_{h=1}^{m} e^{\alpha_{jh}\theta_i + \gamma_{jh}}}. \tag{F.1}$$

This gives the probability of an individual i responding in the kth category of item j (which has m response categories). A key aspect of this model is that the slope/discrimination parameters α_{jk} are allowed to vary for different response categories of the same item. The model can alternatively be written as

$$log\left(\frac{P(Y_{ij} = k)}{P(Y_{ij} = 0)}\right) = \alpha_{jk}\theta_i + \gamma_{jk}, \quad k \neq 0, \tag{F.2}$$

where $Y_{ij} = 0$ specifies the baseline category against which all the other response categories are being compared. This emphasises that the nominal model is related to the multinomial logistic regression model for unordered categorical data. In order to identify the model, the slope α_{j0} and intercept γ_{j0} parameters for the baseline category must be fixed at 0, as well as the usual constraints on the location and scale of the latent trait θ. (It is customary to fix the mean of θ to 0 and its standard deviation to 1.)

For the model specified in OpenBUGS below, the latent trait scores were given a standard normal prior, while the slope and intercept parameters were given weakly informative priors following the recommendation of Gelman *et al.* (2008) for logistic regression parameters (although using a normal rather than Student-t prior).

$$\theta_i \sim N(0, 1)$$

$$\alpha_{jk} \sim N(0, 25)$$

$$\gamma_{jk} \sim N(0, 25)$$

F.2.2 OpenBUGS code

```
model{
   for (i in 1:n){
      for (j in 1:p){
         Y[i,j] ~ dcat(prob[i,j,1:ncat[j]])
      }
      theta[i] ~ dnorm(0, 1)
      for (j in 1:p){
         for (k in 1:(ncat[j])){
            eta[i,j,k] <- alpha[j,k]*theta[i] + inter[j,k]
            exp.eta[i,j,k] <- exp(eta[i,j,k])
         }
         for (k in 1:(ncat[j])){
            prob[i,j,k] <- exp.eta[i,j,k]/sum(exp.eta[i,j,1:ncat[j]])
         }
      }
   }
   for (j in 1:p){
      for (k in 2:(ncat[j])){
         alpha[j,k] ~ dnorm(1.0, pr.alpha)
         alpha.std[j,k] <- alpha[j,k]*theta.sd
         inter[j,k] ~ dnorm(0.0, pr.inter)
         inter.std[j,k] <- inter[j,k] - theta.mean*alpha[j,k]
      }
   }
   theta.mean <- mean(theta[])
   theta.sd <- sd(theta[])
   pr.alpha <- pow(5, -2)
   pr.inter <- pow(5, -2)
}
```

F.2.3 OpenBUGS notes

1. The code above should work equally well in WinBUGS. I chose to use OpenBUGS because it is easier to change the default sampling method (see next point) and because the 'Externalize' option allows a model in progress to be

saved so that it can be edited or restarted later (which is extremely useful for slow running models).

2. OpenBUGS was forced to use slice sampling rather than the default sampling method based on a recommendation of Jackman (2010).

3. This model runs *very slowly* for the sample size of 11,230 used here. The running time on a desktop PC with an Intel Core i7 2.93 GHz processor was approximately 14 hours for each 1,000 iterations.

4. While the scale and location of the latent trait are identified in the above code by the use of a normal prior for the `theta[i]` with mean 0 and variance 1, the actual mean and variance may drift slightly up and down over successive iterations. This may lead to high levels of autocorrelation in the MCMC chains, and more iterations will then be required. For this reason, `theta.mean` and `theta.sd` monitor the mean and variance of the latent trait in any particular iteration, and these values are used to standardise `alpha.std[j,k]` and `inter.std[j,k]`.

5. `alpha[j,k]` and `intercept[j,k]` coefficients for the first category level ($j = 1$) were fixed to 0 to complete the specification of the model. This was achieved by submitting a matrix for these parameters to OpenBUGS as data with values in the first column specified as 0, but all other values specified as `NA`.

6. Other values that were submitted as data were `p` (the number of items), `n` (the sample size) and `ncat` (a vector of length `p` giving the number of categories for each item).

7. As specified, the model is still not strictly identified — it is reflectionally non-invariant, so it has a bimodal posterior density where the two mirror image modes correspond to sets of parameters with reversed signs. Since the two modes are well-separated, each MCMC chain explores only a single mode of the posterior density and does not jump between them. However, it is necessary to check that the starting values chosen for each chain result in both chains exploring the same mode.

8. Two MCMC chains were run for 20,000 iterations, and the initial 3,000 iterations were discarded from each chain. The 'coda' package in R was used to check convergence. Convergence was deemed acceptable for the estimates of posterior means of model parameters, but according to the Raftery-Lewis diagnostic many more iterations would be required for inferences based on extreme quantiles of the posterior distributions (such as 95% credible intervals).

F.2.4 Nominal model results

Table F.1: Estimates of slope coefficients and intercepts from the nominal model

	Slopes/loadings				Intercepts			
Symptom	1	2	3	4	1	2	3	4
Somatic symptoms	1.3	1.5	2.0	2.5	−3.8	−4.0	−4.7	−5.9
Fatigue	1.1	1.6	2.4	3.4	−2.0	−2.2	−2.6	−3.4
Concentration	1.6	2.6	3.9	4.7	−2.8	−4.1	−6.4	−8.6
Sleep	0.7	1.2	2.0	2.8	−2.5	−1.6	−2.6	−4.9
Irritability	1.2	1.7	2.5	3.5	−1.3	−1.9	−3.3	−5.8
Health worry	1.1	1.8	3.0	3.6	−2.6	−3.9	−6.7	−9.5
Depression	2.0	2.6	3.6	4.0	−2.8	−3.8	−5.4	−7.4
Worry	1.3	2.4	3.8	4.8	−1.2	−2.2	−4.3	−6.6
Anxiety	1.5	2.5	3.5	5.1	−2.5	−3.9	−6.1	−9.1
Phobias	0.9	1.6	2.2	2.8	−2.6	−4.0	−6.0	−7.5
Panic	1.8	2.3	2.9	4.3	−6.6	−6.6	−7.1	−10.3
Compulsions	0.8	1.2	1.5	1.9	−3.9	−4.1	−4.8	−6.0
Obsessions	0.7	1.1	1.7	2.4	−4.2	−3.9	−4.3	−5.6

Note: The mean of the latent trait was fixed at 0 and its variance was fixed at 1, so coefficients are in the standard IRT metric. Slope coefficients and intercepts for the 0 score baseline category were fixed at 0 to identify the model.

F.3 Nominal model with quadratic terms

F.3.1 Model description

This model is simply an extension of the nominal model in which a quadratic term $\alpha 2_{jk}$ is added to Equation F.2 to give

$$log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=0)}\right) = \alpha 1_{jk}\theta_i + \alpha 2_{jk}\theta_i^2 + \gamma_{jk}, \quad k \neq 0. \tag{F.3}$$

This equation describes the probability of an individual i responding in the kth category of item j. Again, the $\alpha 1_{jk}$ and $\alpha 2_{jk}$ parameters are each allowed to vary for the different response categories within an item, and the $\alpha 1_{j0}$, $\alpha 2_{j0}$ and γ_{j0} parameters for the baseline category are fixed to 0 to identify the model.

For the OpenBUGS code below, the following priors were used:

$$\boldsymbol{B} = \begin{bmatrix} \alpha 1_{jk} \\ \alpha 2_{jk} \\ \gamma_{jk} \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 25 \end{bmatrix}$$

$$\theta_i \sim N(\mu_\theta, \sigma_\theta)$$

$$\mu_\theta \sim N(0, 1\,000\,000)$$

$$\sigma_\theta \sim U(0, 100)$$

See the OpenBUGS notes below in Section F.3.3 for further details about the choice of priors and how the location and scale of the latent trait were identified.

F.3.2 OpenBUGS code

```
model{
   for(i in 1:n){
      for(j in 1:p){
         Y[i,j] ~ dcat(prob[i,j,1:ncat[j]])
      }
      theta[i] ~ dnorm(mu.theta, pr.theta)
```

```
        eta[i,1,1] <- 0
        eta[i,1,2] <- b1[1]*theta[i] + b1[2]*pow(theta[i],2) + b1[3]
        eta[i,1,3] <- b2[1]*theta[i] + b2[2]*pow(theta[i],2) + b2[3]
        eta[i,1,4] <- b3[1]*theta[i] + b3[2]*pow(theta[i],2) + b3[3]
        eta[i,1,5] <- b4[1]*theta[i] + b4[2]*pow(theta[i],2) + b4[3]

        eta[i,2,1] <- 0
        eta[i,2,2] <- b5[1]*theta[i] + b5[2]*pow(theta[i],2) + b5[3]
        eta[i,2,3] <- b6[1]*theta[i] + b6[2]*pow(theta[i],2) + b6[3]
        eta[i,2,4] <- b7[1]*theta[i] + b7[2]*pow(theta[i],2) + b7[3]
        eta[i,2,5] <- b8[1]*theta[i] + b8[2]*pow(theta[i],2) + b8[3]
        ...
        eta[i,13,1] <- 0
        eta[i,13,2] <- b49[1]*theta[i] + b49[2]*pow(theta[i],2) +b49[3]
        eta[i,13,3] <- b50[1]*theta[i] + b50[2]*pow(theta[i],2) +b50[3]
        eta[i,13,4] <- b51[1]*theta[i] + b51[2]*pow(theta[i],2) +b51[3]
        eta[i,13,5] <- b52[1]*theta[i] + b52[2]*pow(theta[i],2) +b52[3]

        for(j in 1:p){
            for(k in 1:(ncat[j])){
                exp.eta[i,j,k] <- exp(eta[i,j,k])
            }
            for (k in 1:(ncat[j])){
                prob[i,j,k] <- exp.eta[i,j,k]/sum(exp.eta[i,j,1:ncat[j]])
            }
        }
    }
    b1[1] <- 1.33
    b1[2] ~ dnorm(0, 4)
    b1[3] <- -3.8
    b2[1:3] ~ dmnorm(b0[1:3], B0[1:3, 1:3])
    b3[1:3] ~ dmnorm(b0[1:3], B0[1:3, 1:3])
    b4[1:3] ~ dmnorm(b0[1:3], B0[1:3, 1:3])
    ...
    b51[1:3] ~ dmnorm(b0[1:3], B0[1:3, 1:3])
    b52[1:3] ~ dmnorm(b0[1:3], B0[1:3, 1:3])
    mu.theta ~ dnorm(0, 0.000001)
    pr.theta ~ dunif(0, 100)
}
```

F.3.3 OpenBUGS notes

1. This model (as specified above) runs even more slowly than the plain nominal model. The running time on a desktop PC with an Intel Core i7 2.93 GHz processor was approximately 24 hours for each 1,000 iterations.

2. Now that the model incorporates quadratic terms, it is no longer possible to standardise the item parameters on each iteration. For this reason, rather than using a standard normal prior for the `theta[i]`s to identify the model (which only roughly identifies the model parameters), the scale and location of the latent trait were identified more precisely by fixing the values of an $\alpha 1_{jk}$ parameter `b1[1]` and an intercept γ_{jk} parameter `b1[3]`. These values were chosen so that the mean and standard deviation of the latent trait would be approximately 0 and 1 respectively.

3. The MCMC chains for many of the parameters in this model are slow mixing. In order to improve the mixing and speed up convergence, the OpenBUGS code makes use of 'blocking' of the parameters (as recommended by Jackman (2010)). Since the three parameters for each ordered response category are likely to be highly correlated (the $\alpha 1_{jk}$, $\alpha 2_{jk}$ and γ_{jk} parameters for each ordered response category), these groups of three parameters are treated in blocks in the MCMC updating scheme. This is achieved by specifying a multivariate normal prior for each set of three parameters. The OpenBUGS model specifies the sets of three parameters as, for example, `b1[1]`, `b1[2]` and `b1[3]` to help with the specification of the multivariate priors.

4. Once again, OpenBUGS was forced to use slice sampling. This actually updates the parameters univariately, so is not as efficient as using a true block updater. However, the block updaters in OpenBUGS were unfeasibly slow for this model. Nonetheless, the specification of multivariate priors still helps to reduce the correlations between the sampled values of the blocked parameters.

5. Two MCMC chains were run for 20,000 iterations, and the initial 3,000 iterations were discarded from each chain. The 'coda' package in R was used to check convergence. However, the Raftery-Lewis diagnostic indicated that hundreds of thousands of iterations would be required to provide stable estimates of the extreme quantiles of the posterior distribution.

F.3.4 Model results

Table F.2: Estimates of coefficients from the nominal model with quadratic terms

	$\alpha 1$				$\alpha 2$				γ			
Symptom	1	2	3	4	1	2	3	4	1	2	3	4
Somatic symptoms	1.3	2.0	2.5	2.0	−0.1	−0.3	−0.3	0.1	−3.8	−4.0	−4.9	−5.6
Fatigue	1.3	1.8	2.7	3.1	−0.5	−0.3	−0.4	−0.2	−1.8	−2.1	−2.5	−3.1
Concentration	2.0	3.3	4.1	4.4	−0.4	−0.5	−0.3	−0.2	−2.7	−4.3	−6.3	−8.2
Sleep	0.8	1.3	1.9	1.9	−0.2	−0.2	0.0	0.2	−2.4	−1.5	−2.5	−4.6
Irritability	1.4	1.9	2.8	3.6	−0.4	−0.4	−0.4	−0.3	−1.1	−1.8	−3.2	−5.6
Health worry	1.6	2.1	3.2	2.4	−0.4	−0.2	−0.2	0.2	−2.5	−3.9	−6.8	−8.6
Depression	2.3	2.0	2.4	2.0	−0.3	0.3	0.5	0.7	−2.8	−3.6	−5.0	−6.5
Worry	1.7	3.0	3.4	4.6	−0.7	−0.7	−0.3	−0.4	−1.0	−2.2	−3.8	−6.3
Anxiety	2.7	3.9	5.0	5.9	−1.0	−0.9	−0.9	−0.7	−2.5	−4.2	−6.5	−9.3
Phobias	1.1	1.7	1.8	1.8	−0.1	−0.1	0.1	0.3	−2.6	−3.9	−5.8	−7.0
Panic	3.4	3.5	3.7	4.8	−0.8	−0.5	−0.4	−0.3	−7.2	−7.1	−7.5	−10.7
Compulsions	0.9	1.3	1.4	1.9	−0.2	−0.1	0.0	0.0	−3.8	−4.0	−4.8	−6.0
Obsessions	1.6	1.9	2.7	3.0	−0.8	−0.5	−0.6	−0.4	−4.0	−3.9	−4.5	−5.8

Note: The location and scale of the latent trait were fixed so that they were approximately 0 and 1 respectively, therefore coefficients are in the standard IRT metric.
All coefficients for the 0 score baseline category were fixed at 0 to identify the model.

F.4 Simulating data from the nominal quadratic model

These two functions together allow data to be simulated based on the nominal quadratic model described in Section F.3. The second function depends on the probabilities calculated by the first. The idea to use this two step approach came from Habing (no date).

```
### Calculating the probabilities of each symptom score based on the
### supplied model parameters for a given set of theta scores.

nom.quadratic.probs <- function(ints, a1s, a2s,
                                 theta.req=seq(-4.5, 4.5, by=0.1)) {
   eta <- matrix(rep(rep(NA, 5), length(theta.req)),
              nrow=length(theta.req), ncol=5, byrow=T)
   exp.eta <- matrix(rep(rep(NA, 5), length(theta.req)),
              nrow=length(theta.req), ncol=5, byrow=T)
   prob <- matrix(rep(rep(NA, 5), length(theta.req)),
              nrow=length(theta.req), ncol=5, byrow=T)
   p <- matrix(rep(rep(NA, 5), length(theta.req)),
              nrow=length(theta.req), ncol=5, byrow=T)
   for (i in 1:length(theta.req)){
      eta[i,1] <- 0*theta.req[i] + 0*theta.req[i]^2 + 0
      eta[i,2] <- a1s[1]*theta.req[i] +a2s[1]*theta.req[i]^2 +ints[1]
      eta[i,3] <- a1s[2]*theta.req[i] +a2s[2]*theta.req[i]^2 +ints[2]
      eta[i,4] <- a1s[3]*theta.req[i] +a2s[3]*theta.req[i]^2 +ints[3]
      eta[i,5] <- a1s[4]*theta.req[i] +a2s[4]*theta.req[i]^2 +ints[4]
      for (k in 1:5){
         exp.eta[i,k] <- exp(eta[i,k])
      }
      for (k in 1:5){
         p[i,k] <- exp.eta[i,k] / (sum(exp.eta[i,1:5]))
      }
      prob[i,1] <- p[i,2] + p[i,3] + p[i,4] + p[i,5]
      prob[i,2] <- p[i,3] + p[i,4] + p[i,5]
      prob[i,3] <- p[i,4] + p[i,5]
      prob[i,4] <- p[i,5]
   }
   return(prob)
```

```
}

### Simulating ordinal response data based on the probabilities
### calculated in the first function.

gen.nom.quad <- function(thetas=rnorm(100,mean=0,sd=1), alpha1s,
              alpha2, intercepts, theta.mean=0, theta.sd=1){
   nitems <- nrow(alpha1s)
   data <- matrix(0, nrow=length(thetas), ncol=nitems)
   for(j in 1: nrow(alpha1s)){
      cum.probs <- nom.quadratic.probs(a1s=alpha1s[j,],
              a2s=alpha2s[j,], ints=intercepts[j,], theta.req=thetas)
      sim <- runif(length(thetas),0,1)
      for (i in 1:length(thetas)){
         if(sim[i] <= cum.probs[i,1]){
            data[i, j]<-1
         }
         if(sim[i] <= cum.probs[i,2]){
            data[i, j]<-2
         }
         if(sim[i] <= cum.probs[i,3]){
            data[i, j]<-3
         }
         if(sim[i] <= cum.probs[i,4]){
            data[i, j]<-4
         }
      }
   }
   return(data)
}
```

The following commands actually simulate the data using the two functions described above. The parameter values specified here come from the results of the nominal quadratic model estimated in OpenBUGS shown in Table F.2 on page 318.

```
scores <- rnorm(10000,mean=0,sd=1)
quadratic_alpha1s <- rbind(c(1.3,2.0,2.5,2.0), c(1.3,1.8,2.7,3.1),
              c(2.0,3.3,4.1,4.4), c(0.8,1.3,1.9,1.9),
              c(1.4,1.9,2.8,3.6), c(1.6,2.1,3.2,2.4),
              c(2.3,2.0,2.4,2.0), c(1.7,3.0,3.4,4.6),
              c(2.7,3.9,5.0,5.9), c(1.1,1.7,1.8,1.8),
              c(3.4,3.5,3.7,4.8), c(0.9,1.3,1.4,1.9),
              c(1.6,1.9,2.7,3.0))
quadratic_alpha2s <- rbind(c(-0.1,-0.3,-0.3,0.1),
              c(-0.5,-0.3,-0.4,-0.2), c(-0.4,-0.5,-0.3,-0.2),
              c(-0.2,-0.2,0.0,0.2), c(-0.4,-0.4,-0.4,-0.3),
              c(-0.4,-0.2,-0.2,0.2), c(-0.3,0.3,0.5,0.7),
              c(-0.7,-0.7,-0.3,-0.4), c(-1,-0.9,-0.9,-0.7),
              c(-0.1,-0.1,0.1,0.3), c(-0.8,-0.5,-0.4,-0.3),
              c(-0.2,-0.1,0.0,0.0),c(-0.8,-0.5,-0.6,-0.4))
quadratic_intercepts <- rbind(c(-3.8,-4.0,-4.9,-5.6),
              c(-1.8,-2.1,-2.5,-3.1), c(-2.7,-4.3,-6.3,-8.2),
              c(-2.4,-1.5,-2.5,-4.6), c(-1.1,-1.8,-3.2,-5.6),
              c(-2.5,-3.9,-6.8,-8.6), c(-2.8,-3.6,-5.0,-6.5),
              c(-1.0,-2.2,-3.8,-6.3), c(-2.5,-4.2,-6.5,-9.3),
              c(-2.6,-3.9,-5.8,-7.0), c(-7.2,-7.1,-7.5,-10.7),
              c(-3.8,-4.0,-4.8,-6.0), c(-4.0,-3.9,-4.5,-5.8))
simdata_quad <- gen.nom.quad(thetas=scores,
              alpha1s=quadratic_alpha1s,alpha2s=quadratic_alpha2s,
              intercepts=quadratic_intercepts)
```

F.4.1 Simulating data from the standard factor model

The functions simulating data under the standard factor model follow the same format as those in the previous section for simulating data under the nominal quadratic model.

```
grm.probs <- function(load, int, factor.req=seq(-4.5, 4.5, by=0.1)){
   eta <- matrix(rep(rep(NA, 4), length(factor.req)),
              nrow=length(factor.req), ncol=4, byrow=T)
   exp.eta <- matrix(rep(rep(NA, 4), length(factor.req)),
              nrow=length(factor.req), ncol=4, byrow=T)
   prob <- matrix(rep(rep(NA, 4), length(factor.req)),
              nrow=length(factor.req), ncol=4, byrow=T)
   for (i in 1:length(factor.req)){
      for (k in 1:4){
         eta[i, k] <- int[k] + (load * factor.req[i])
      }
      exp.eta <- exp(eta)
      for (k in 1:4){
         prob[i, k] <- 1 -(1 / (1 + exp.eta[i, k]))
      }
   }
   return(prob)
}


gen.grm <- function(thetas=rnorm(100,mean=0,sd=1),
              loadings=rep(1,10),
              intercepts=matrix(rep(c(-1,-2,-3,-4), 10),
              nrow=10, ncol=4, byrow=T), theta.mean=0, theta.sd=1){
   nitems<-length(loadings)
   data<-matrix(0,nrow=length(thetas),ncol=nitems)
   for(j in 1: length(loadings)){
      cum.probs <- grm.probs(load=loadings[j], int=intercepts[j,],
              factor.req=thetas)
      sim <- runif(length(thetas),0,1)
      for (i in 1:length(thetas)){
         if(sim[i] <= cum.probs[i,1]){
```

```
            data[i, j]<-1
         }
         if(sim[i] <= cum.probs[i,2]){
            data[i, j]<-2
         }
         if(sim[i] <= cum.probs[i,3]){
            data[i, j]<-3
         }
         if(sim[i] <= cum.probs[i,4]){
             data[i, j]<-4
         }
      }
   }
   return(data)
}

## Actually simulating the ordinal response data

scores <- rnorm(10000,mean=0,sd=1)
grm_loadings <- c(1.7, 2, 2.3, 1.5, 1.6, 1.5, 2.2, 2.1, 2.2,
              1.3, 2.6, 1.2, 1.5)
grm_intercepts <- rbind(c(-2.9,-3.4,-4.2,-5.3),
              c(-0.9,-1.5,-2.2,-3.3), c(-2.6,-3.6,-4.9,-6.5),
              c(-0.9,-1.2,-2.6,-4.6), c(-0.7,-1.9,-3.3,-5.1),
              c(-2.4,-3.5,-5.1,-7.2), c(-2.3,-3.4,-4.7,-6.8),
              c(-0.8,-2.2,-3.6,-5.4), c(-2.3,-3.5,-4.8,-6.0),
              c(-2.4,-3.5,-4.9,-5.9), c(-5.7,-5.9,-6.4,-7.7),
              c(-3.0,-3.4,-4.2,-5.4), c(-2.9,-3.2,-3.7,-4.7))
simdata_grm <- gen.grm(thetas=scores, loadings=grm_loadings,
              intercepts=grm_intercepts)
```

F.4.2 Comparing the simulated data to the real data

A notable difference between the real data and the data simulated from the nominal quadratic model is that in the real data the Bayesian information criterion favours the three class factor mixture model over models with one or two classes, whereas in the simulated data it favours the two class model. This discrepancy may reflect the failure of the nominal quadratic model to capture some key aspects of the CIS-

R data. For this reason, it may be helpful to compare directly the relationships between symptoms and the latent trait seen in the real and simulated data. If the relationships seen in the two data sets are extremely similar, this would indicate that the nominal quadratic model does a very good job of describing and replicating the CIS-R data. However, if there are substantial differences, this would indicate some form of shortcoming in the nominal quadratic model.

Figures F.1 to F.4 show a comparison of the non-parametric plots for each symptom for the data simulated from the nominal quadratic model (solid lines) and the real data (dotted lines). For these plots, the curves for the real data are based on the first split half of the data only (unlike the non-parametric plots in Chapter 7, which used the full CIS-R data). This is because the parameters used to generate the simulated data were based on the results of the nominal quadratic model fitted to the first half of the data only; therefore, using the full CIS-R data might distort the comparison (since fitting the nominal quadratic model to the full data may yield slightly different parameter estimates). Furthermore, the non-parametric regression function may be sensitive to large differences in sample size at the extremes of the plotting region.

Figures F.1 to F.4 suggest that the nominal quadratic model does a reasonable job of describing the CIS-R data for some symptoms. However, there are a few symptoms for which the approximation appears poor (particularly in the upper regions of the latent trait above scores of approximately 2). For both somatic symptoms and anxiety, the nominal quadratic model accommodates *some* flattening off in the response curves at high latent trait scores, relative to the response curves allowed by the standard factor model (as was seen in Figure 7.14 on page 198 and Figure 7.16 on page 200). However, even the flattening off allowed by this more flexible factor model does not appear to match the actual degree of flattening off seen in the real data. Meanwhile, for the obsessions symptom the reverse appears to be true: the nominal quadratic model appears to do *too much* flattening off at high trait scores.

This seems likely to be related to the relatively limited range of shapes that can be accommodated by the quadratic term in the nominal quadratic model: on the logit scale, the response functions must be parabolic curves. The situation is made more complex by the fact that the transformation back to the probability scale (the inverse logit transformation) will have different effects on the shape of the quadratic curve depending on the vertical position of the curve at a given trait score. After transformation, changes in the gradient of a curve associated with a quadratic term will have most impact on the predicted probabilities when they occur at logit values corresponding to probabilities of approximately 0.5 (logit values around 0). At

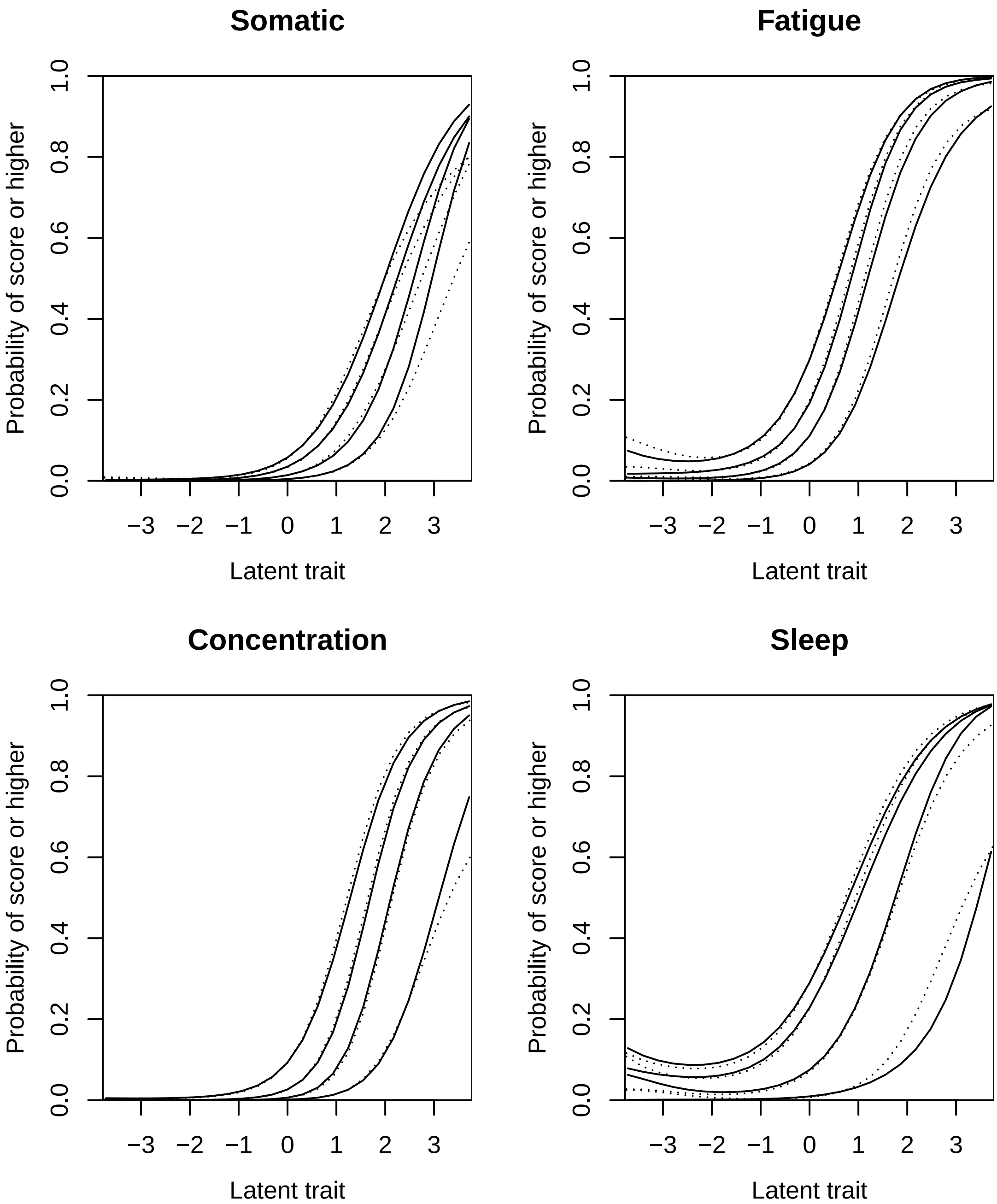

Figure F.1: Non-parametric plots of data simulated from the nominal quadratic model (solid lines) superimposed on non-parametric plots based on the first half of the real data (dotted lines)

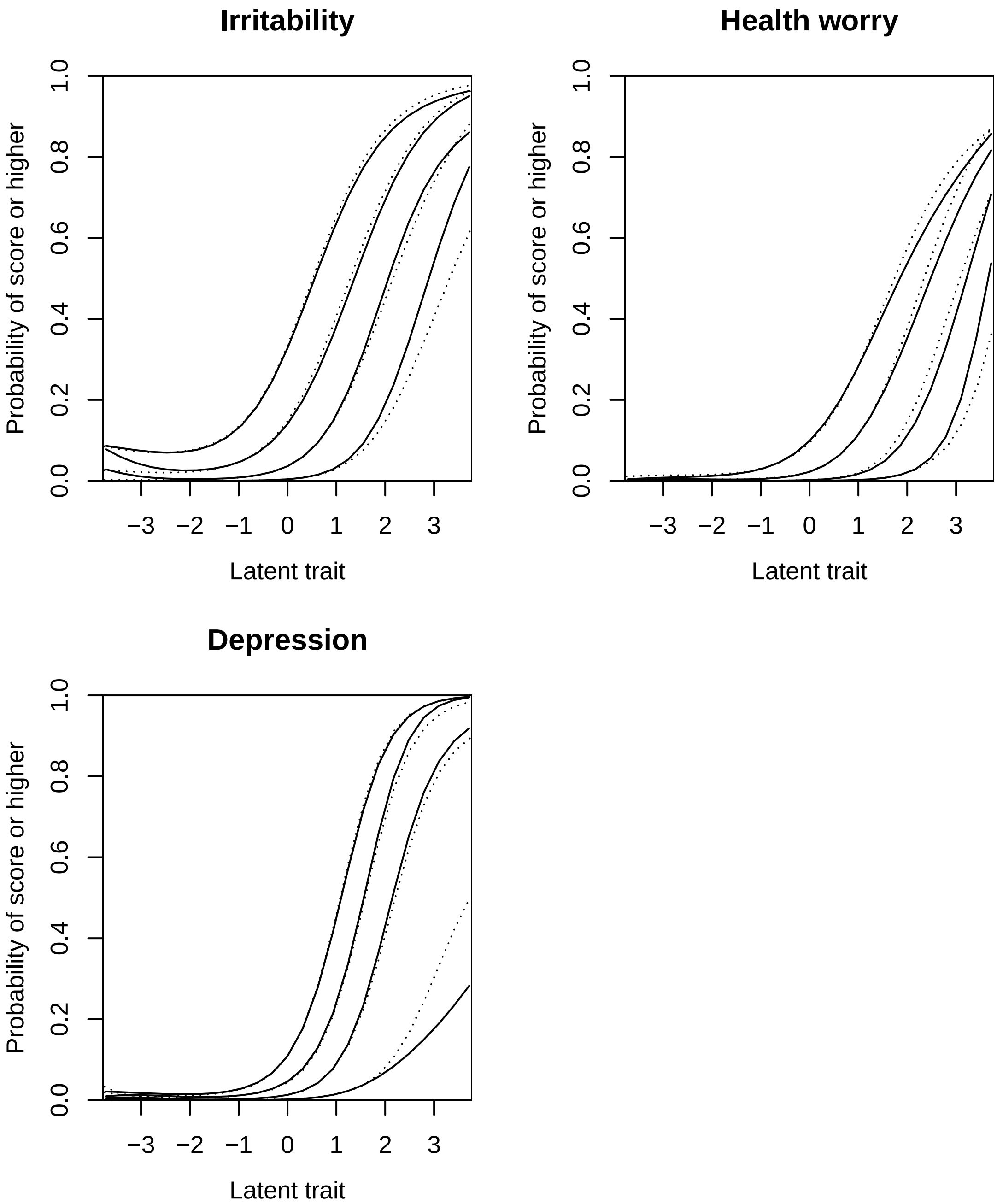

Figure F.2: Non-parametric plots of data simulated from the nominal quadratic model (solid lines) superimposed on non-parametric plots based on the first half of the real data (dotted lines)

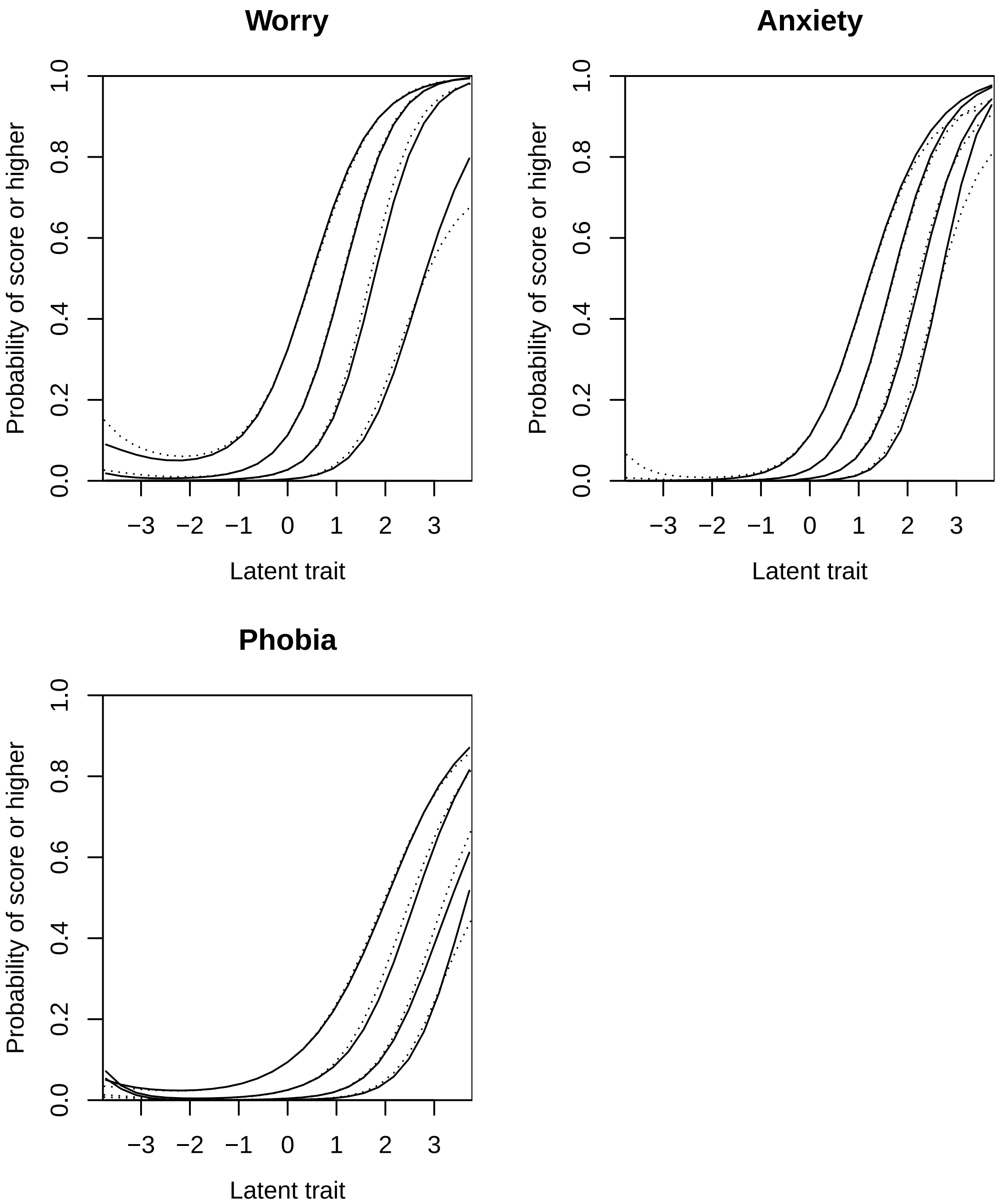

Figure F.3: Non-parametric plots of data simulated from the nominal quadratic model (solid lines) superimposed on non-parametric plots based on the first half of the real data (dotted lines)

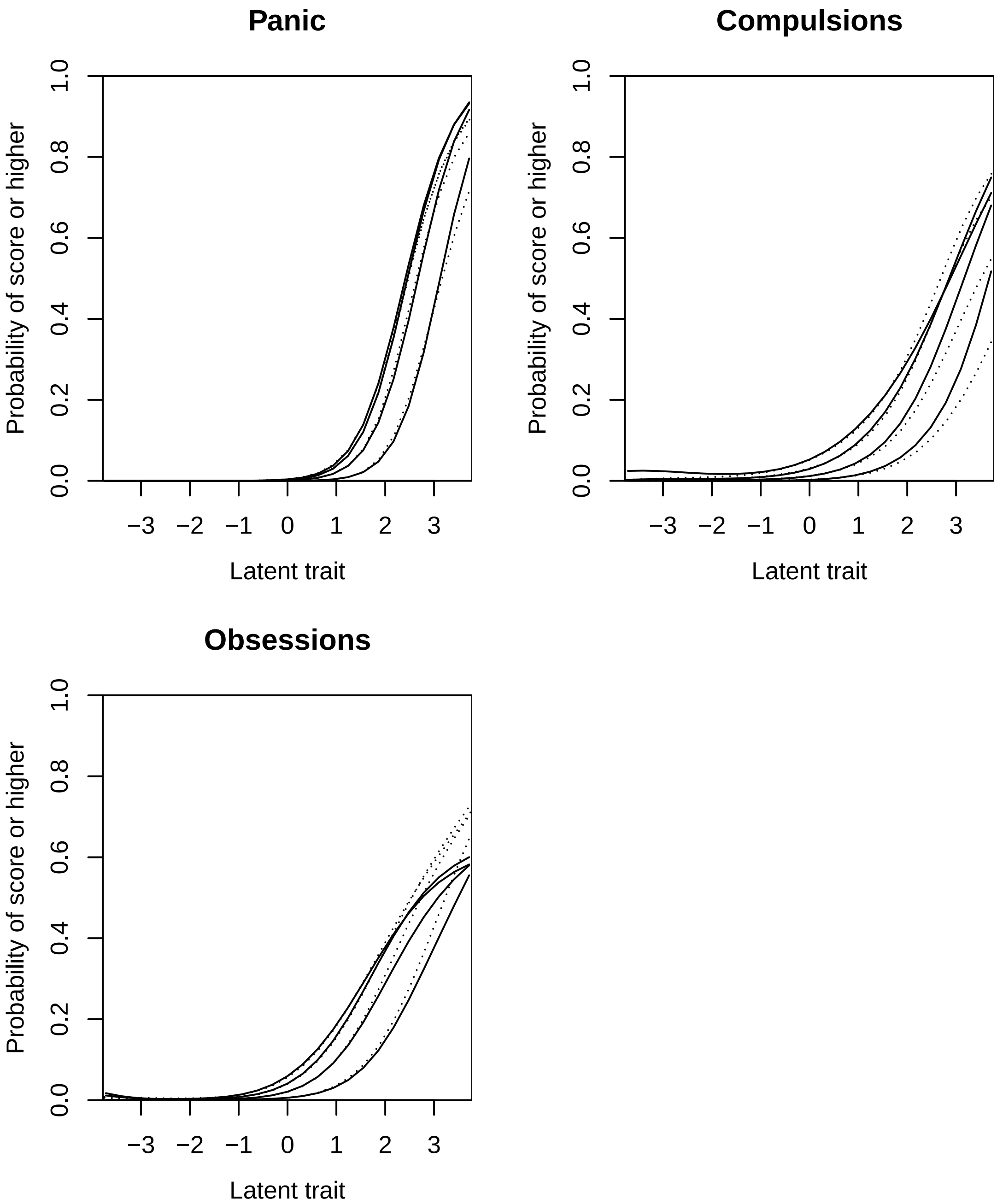

Figure F.4: Non-parametric plots of data simulated from the nominal quadratic model (solid lines) superimposed on non-parametric plots based on the first half of the real data (dotted lines)

probabilities closer to 0 or 1 (corresponding to large positive or negative logit values), the changes in gradient described by the quadratic term will have a much lesser impact on the predicted probabilities. Therefore, it may be difficult to find a single quadratic curve that can accurately describe the shape of a gradually flattening curve across all values of the latent trait.

It seems plausible that it is the limited degree of flattening off captured by the nominal quadratic model for items such as somatic symptoms and anxiety that underpins the lack of support for the three class factor mixture model in the simulated data. This highlights the need for more flexible factor or item response theory models to describe psychopathology data.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. John Wiley & Sons, Hoboken, NJ, 2nd edn.

Aldhous, P. (2012). Trials highlight worrying flaws in psychiatry 'bible'. *New Scientist*, **2865**, 17 May.

Almansa, J., Vermunt, J.K., Forero, C.G., Vilagut, G., De Graaf, R., De Girolamo, G. & Alonso, J. (2011). Measurement and description of underlying dimensions of comorbid mental disorders using factor mixture models: results of the ESEMeD project. *International Journal of Methods in Psychiatric Research*, **20**, 116–133.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Washington, DC, 4th edn.

American Psychiatric Association (2010a). *Cross-Cutting Dimensional Assessment in DSM-5*. http://www.dsm5.org/ProposedRevisions/Pages/Cross-CuttingDimensionalAssessmentinDSM-5.aspx, accessed: 12 October 2011.

American Psychiatric Association (2010b). *Mixed Anxiety Depression*. http://www.dsm5.org/ProposedRevisions/Pages/proposedrevision.aspx?rid=407, accessed: 12 October 2011.

American Psychiatric Association (2010c). *Mood Disorders*. http://www.dsm5.org/PROPOSEDREVISIONS/Pages/MoodDisorders.aspx, accessed: 12 October 2011.

American Psychiatric Association (2012). *Proposed Revision | APA DSM-5*. http://www.dsm5.org/ProposedRevisions/Pages/proposedrevision.aspx?rid=407, accessed: 25 August 2012.

Asmundson, G.J., Taylor, S., Nicholas Carleton, R., Weeks, J.W. & Hadjstavropoulos, H.D. (2012). Should health anxiety be carved at the joint? a look at the health anxiety construct using factor mixture modeling in a non-clinical sample. *Journal of Anxiety Disorders*, **26**, 246–251.

Baillie, A.J. & Teesson, M. (2010). Continuous, categorical and mixture models of DSM-IV alcohol and cannabis use disorders in the australian community. *Addiction*, **105**, 1246–1253.

Baker, R.P., Bradburn, N.M. & Johnson, R.A. (1995). Computer-assisted personal interviewing: An experimental evaluation of data quality and costs. *Journal of Official Statistics*, **11**, 413–431.

Bartholomew, D., Knott, M. & Moustaki, I. (2011). *Latent variable models and factor analysis: a unified approach*. John Wiley & Sons, Chichester, 3rd edn.

Bartholomew, D.J., Steele, F., Galbraith, J. & Moustaki, I. (2008). *Analysis of Multivariate Social Science Data*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edn.

Bauer, D. & Curran, P. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, **8**, 338–363.

Bauer, D.J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, **10**, 305–316.

Bauer, D.J. & Curran, P.J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, **9**, 3–29.

Bauermeister, J.J., Canino, G., Polanczyk, G. & Rohde, L.A. (2010). ADHD across cultures: Is there evidence for a bidimensional organization of symptoms? *Journal of Clinical Child & Adolescent Psychology*, **39**, 362–372.

Bernstein, A., Stickle, T.R., Zvolensky, M.J., Taylor, S., Abramowitz, J. & Stewart, S. (2010). Dimensional, categorical, or dimensional-categories: Testing the latent structure of anxiety sensitivity among adults using factor-mixture modeling. *Behavior Therapy*, **41**, 515–529.

Bernstein, A., Cárdenas, S., Coy, P. & Zvolensky, M. (2011). Test of a factor mixture-based taxonic-dimensional model of anxiety sensitivity and panic attack vulnerability among university and clinical samples in mexico city. *Journal of Psychopathology and Behavioral Assessment*, **33**, 491–500.

Blashfield, R. & Morey, L. (1979). Classification of Depression Through Cluster-Analysis. *Compr.Psychiatry*, **20**, 516–527.

Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

Bowman, A.W. & Azzalini, A. (2010). *R package `sm`: nonparametric smoothing methods (version 2.2-4)*. University of Glasgow, UK and Università di Padova, Italia.

Brown, T., Campbell, L., Lehman, C., Grisham, J. & Mancill, R. (2001). Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *J.Abnorm.Psychol.*, **110**, 585–599.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. Guilford Press, New York, NY.

Brugha, T. (2002). The end of the beginning: a requiem for the categorization of mental disorder? *Psychol.Med.*, **32**, 1149–1154.

Byrne, D. (1978). Note on Cluster-Analysis and Depression - Disparities in Results Produced by Application of Different Clustering Methods. *Aust.N.Z.J.Psychiatry*, **12**, 99–102.

Carragher, N., Adamson, G., Bunting, B. & McCann, S. (2009). Subtypes of depression in a nationally representative sample. *J.Affect.Disord.*, **113**, 88–99.

Chen, L., Eaton, W., Gallo, J. & Nestadt, G. (2000). Understanding the heterogeneity of depression through the triad of symptoms, course and risk factors: a longitudinal, population-based study. *J.Affect.Disord.*, **59**, 1–11.

Clark, L. (1989). The anxiety and depressive disorders: descriptive psychopathology and differential diagnosis. In P. Kendall & D. Watson, eds., *Anxiety and Depression: Distinctive and Overlapping Features*, 83–129, Academic Press, San Diego.

Clark, S.L. & Muthen, B.O. (2009). *Relating latent class analysis results to variables not included in the analysis*. http://www.statmodel.com/download/relatinglca.pdf, accessed: 22 Decmber 2011.

Clark, S.L., Muthen, B.O., Kaprio, J., D'Onofrio, B.M., Viken, R., Rose, R.J. & Smalley, S.L. (2009). *Models and strategies for factor mixture analysis: Two examples concerning the structure underlying psychological disorders*. http://www.statmodel.com/papers.shtml, accessed: 12 October 2011.

Collins, L.M. & Lanza, S.T. (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley, Chichester.

Cuijpers, P., De Graaf, R. & van Dorsselaer, S. (2004). Minor depression: risk profiles, functional disability, health care use and risk of developing major depression. *J.Affect.Disord.*, **79**, 71–79.

Daniels, M.G. (2009). *hdf5: Interface to the NCSA HDF5 library*. http://cran.r-project.org/web/packages/hdf5/index.html, R package version 1.6.9.

Das-Munshi, J., Goldberg, D., Bebbington, P., Bhugra, D., Brugha, T., Dewey, M., Jenkins, R., Stewart, R. & Prince, M. (2008). Public health significance of mixed anxiety and depression: beyond current classification. *Br.J.Psychiatry*, **192**, 171–177.

de Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. The Guildford Press, New York, NY.

Dickman, K. (1960). *Factorial validity of a rating instrument*. Ph.D. thesis, University of Illinois.

Drasgow, F. & Lissak, R.I. (1983). Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, **68**, 363–373.

Eaton, W., Dryman, A., Sorenson, A. & Mccutcheon, A. (1989a). DSM-III major depressive disorder in the community — a latent class analysis of data from the NIMH Epidemiologic Catchment-Area Program. *Br.J.Psychiatry*, **155**, 48–54.

Eaton, W., Mccutcheon, A., Dryman, A. & Sorenson, A. (1989b). Latent class analysis of anxiety and depression. *Sociological Methods & Research*, **18**, 104–125.

Elhai, J.D., Naifeh, J.A., Forbes, D., Ractliffe, K.C. & Tamburrino, M. (2011). Heterogeneity in clinical presentations of posttraumatic stress disorder among medical patients: Testing factor structure variation using factor mixture modeling. *Journal of Traumatic Stress*, **24**, 435–443.

Elhai, J.D., Contractor, A.A., Tamburrino, M., Fine, T.H., Prescott, M.R., Shirley, E., Chan, P.K., Slembarski, R., Liberzon, I., Galea, S. & Calabrese, J.R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, advance online publication.

Ferdinand, R., de Nijs, P., van Lier, P. & Verhulst, F. (2005). Latent class analysis of anxiety and depressive symptoms in referred adolescents. *J.Affect.Disord.*, **88**, 299–306.

Finch, H. & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, **21**, 119–140.

Fox, J. (2010). *polycor: Polychoric and Polyserial Correlations*. http://CRAN.R-project.org/package=polycor, r package version 0.7-8.

Frazier, T.W., Youngstrom, E.A., Speer, L., Embacher, R., Law, P., Constantino, J., Findling, R.L., Hardan, A.Y. & Eng, C. (2012). Validation of proposed DSM-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, **51**, 28–40.e3.

Gallo, J.J., Anthony, J.C. & Muthen, B.O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology*, **49**, 251–264.

Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

Gillespie, N.A., Kendler, K.S. & Neale, M.C. (2011). Psychometric modeling of cannabis initiation and use and the symptoms of cannabis abuse, dependence and withdrawal in a sample of male and female twins. *Drug and Alcohol Dependence*, **118**, 166–172.

Gillespie, N.A., Neale, M.C., Legrand, L.N., Iacono, W.G. & McGue, M. (2012). Are the symptoms of cannabis use disorder best accounted for by dimensional, categorical, or factor mixture models? a comparison of male and female young adults. *Psychology of Addictive Behaviors*, **26**, 68–77.

Goldberg, D. (2008). Towards DSM-V: the relationship between generalized anxiety disorder and major depressive episode. *Psychol.Med.*, **38**, 1671–1675.

Goldberg, D. & Huxley, P. (1992). *Common Mental Disorders: a Bio-social Model*. Tavistock/Routledge, London.

Goldberg, D., Kendler, K.S., Sirovatka, P.J. & Regier, D.A. (2010). *Diagnostic Issues in Depression and Generalized Anxiety Disorder: Refining the Research Agenda for DSM-V*. American Psychiatric Association, Arlington, VA.

Habing, B. (no date). *Item Response Theory - Examples in R*. http://www.unt.edu/rss/irt_r_scripts.htm#h3, accessed: 03 August 2012.

Hagenaars, J.A. & McCutcheon, A.L., eds. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Hallquist, M. (2012). *MplusAutomation: Automating Mplus Model Estimation and Interpretation*. http://cran.r-project.org/web/packages/MplusAutomation/index.html, R package version 0.5-2.

Heeringa, S.G., West, B.T. & Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.

Hildebrandt, T., Alfano, L. & Langenbucher, J.W. (2010). Body image disturbance in 1000 male appearance and performance enhancing drug users. *Journal of Psychiatric Research*, **44**, 841–846.

Hong, S. & Min, S. (2007). Mixed Rasch modeling of the self-rating depression scale - Incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, **67**, 280–299.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, **30**, 179–185.

Hu, L.t. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**, 1–55.

Hybels, C., Blazer, D., Pieper, C., Landerman, L. & Steffens, D. (2009). Profiles of depressive symptoms in older adults diagnosed with major depression: Latent cluster analysis. *Am.J.Geriatr.Psychiatry*, **17**, 387–396.

Jackman, S. (2010). 19 worked examples in WinBUGS. http://jackman.stanford.edu/mcmc/mainframeWinBugs.php, accessed: 16/05/2012.

Jeffries, N.O. (2003). A note on 'Testing the number of components in a normal mixture'. *Biometrika*, **90**, 991 –994.

Jenkins, R., Meltzer, H., Bebbington, P., Brugha, T., Farrell, M., McManus, S. & Singleton, N. (2009). The british mental health survey programme: achievements and latest findings. *Social Psychiatry and Psychiatric Epidemiology*, **44**, 899–904.

Joreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, **34**, 183–202.

Junker, B.W. & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, **24**, 65–81.

Kamata, A. & Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, **15**, 136–153.

Keller, F. & Kempf, W. (1993). Stability and Validity of Latent Classes in the Beck Depression Inventory (BDI). *Psychiatr.Prax.*, **20**, 68–71.

Kendler, K., Eaves, L., Walters, E., Neale, M., Heath, A. & Kessler, R. (1996). The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Arch.Gen.Psychiatry*, **53**, 391–399.

Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1992). Major depression and generalized anxiety disorder: Same genes, (partly) different environments? *Archives of General Psychiatry*, **49**, 716–722.

Kessler, R. & Ustun, T. (2004). The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, **13**, 93–121.

Kessler, R., Nelson, C., McGonagle, K., Liu, J., Swartz, M. & Blazer, D. (1996). Comorbidity of DSM-III-R major depressive disorder in the general population: Results from the US National Comorbidity Survey. *Br.J.Psychiatry*, **168**, 17–30.

Kim, S.H., Beretvas, S.N. & Sherry, A.R. (2010). A validation of the factor structure of OQ-45 scores using factor mixture modeling. *Measurement and Evaluation in Counseling and Development*, **42**, 275–295.

Kim, Y. & Muthen, B.O. (2009). Two-part factor mixture modeling: Application to an aggressive behavior measurement instrument. *Structural Equation Modeling-A Multidisciplinary Journal*, **16**, 602–624.

King, M., Bottomley, C., Bellón-Saameño, J.A., Torres-Gonzalez, F., Švab, I., Rifel, J., Maaroos, H.I., Aluoja, A., Geerlings, M.I., Xavier, M., Carraça, I., Vicente, B., Saldivia, S. & Nazareth, I. (2011). An international risk prediction algorithm for the onset of generalized anxiety and panic syndromes in general practice attendees: predictA. *Psychological Medicine*, **41**, 1625–1639.

Korn, E.L. & Graubard, B.I. (1999). *Analysis of Health Surveys*. Wiley, New York.

Kramer, A. & Zimmerman, J. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit.Care Med.*, **35**, 2052–2056.

Kuo, P.H., Aggen, S.H., Prescott, C.A., Kendler, K.S. & Neale, M.C. (2008). Using a factor mixture modeling approach in alcohol dependence in a general population sample. *Drug and Alcohol Dependence*, **98**, 105–114.

Leoutsakos, J., Zandi, P., Bandeen-Roche, K. & Lyketsos, C. (2010). Searching for valid psychiatric phenotypes: discrete latent variable models. *International Journal of Methods in Psychiatric Research*, **19**, 63–73.

Levine, J., Cole, D.P., Chengappa, K.N.R. & Gershon, S. (2001). Anxiety disorders and major depression, together or apart. *Depression and Anxiety*, **14**, 94–104.

Lewis, G., Pelosi, A., Araya, R. & Dunn, G. (1992). Measuring psychiatric disorder in the community — a standardized assessment for use by lay interviewers. *Psychol.Med.*, **22**, 465–486.

Lo, Y., Mendell, N.R. & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

Lubke, G. & Muthen, B.O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, **10**, 21–39.

Lubke, G. & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, **41**, 499–532.

Lubke, G. & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, **43**, 592–620.

Lubke, G. & Tueller, S. (2010). Latent class detection and class assignment: A comparison of the MAXEIG taxometric procedure and factor mixture modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, **17**, 605–628.

Lubke, G., Muthen, B.O., Moilanen, I., McGough, J., Loo, S., Swanson, J., Yang, M., Taanila, A., Hurtig, T., Jarvelin, M. & Smalley, S. (2007). Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. *J.Am.Acad.Child Adolesc.Psychiatry*, **46**, 1584–1593.

Lubke, G.H., Hudziak, J.J., Derks, E.M., van Bijsterveldt, T.C. & Boomsma, D.I. (2009). Maternal ratings of attention problems in ADHD: evidence for the existence of a continuum. *Journal of the American Academy of Child & Adolescent Psychiatry*, **48**, 1085–1093.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, **28**, 3049–3067.

Maj, M. (2005). 'Psychiatric comorbidity': an artefact of current diagnostic systems? *Br.J.Psychiatry*, **186**, 182–184.

Markon, K.E. (2010). Modeling psychopathology structure: A symptom-level analysis of axis i and II disorders. *Psychological Medicine*, **40**, 273–288.

Masyn, K.E., Henderson, E., Craig & Greenbaum, P.E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional-categorical spectrum. *Social Development*, **19**, 470–493.

McBride, O., Teesson, M., Baillie, A. & Slade, T. (2011). Assessing the dimensionality of lifetime DSM-IV alcohol use disorders and a Quantity–Frequency alcohol use criterion in the australian population: a factor mixture modelling approach. *Alcohol and Alcoholism*, **46**, 333–341.

McDonald, R.P. (1965). Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology*, **18**, 11–23.

McKay Curtis, S. (2010). BUGS code for item response theory. *Journal of Statistical Software, Code Snippets*, **36**, 1–34.

McLachlan, G.J. & Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York, NY.

Meijer, R.R. & Baneke, J.J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods*, **9**, 354–368.

Middeldorp, C.M., Cath, D.C., Van Dyck, R. & Boomsma, D.I. (2005). The comorbidity of anxiety and depression in the perspective of genetic epidemiology. a review of twin and family studies. *Psychological Medicine*, **35**, 611–624.

Mineka, S., Watson, D. & Clark, L. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annu.Rev.Psychol.*, **49**, 377–412.

Molenaar, P.C. & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C.C. Clogg, eds., *Latent variables analysis: applications for developmental research*, 226–242, Sage Publications, Thousand Oaks.

Mora, P.A., Beamon, T., Preuitt, L., DiBonaventura, M., Leventhal, E.A. & Leventhal, H. (2012). Heterogeneity in depression symptoms and health status among older adults. *Journal of Aging and Health*, **24**, 879–896.

Muthen, B.O. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction*, **101**, 6–16.

Muthen, B.O. (2007). Latent variable hybrids: Overview of old and new models. In G.R. Hancock & K.M. Samuelsen, eds., *Advances in Latent Variable Mixture Models*, 1–24, Information Age Publishing Inc., Charlotte, NC.

Muthen, B.O. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addict.Behav.*, **31**, 1050–1066.

Muthen, L.K. & Muthen, B.O. (2010a). *Mplus User's Guide. Sixth Edition*. Muthen & Muthen, Los Angeles, CA.

Muthen, L.K. & Muthen, B.O. (2010b). *Mplus Version 6.1*. Muthen & Muthen, Los Angeles, CA.

Naifeh, J.A., Richardson, J.D., Del Ben, K.S. & Elhai, J.D. (2010). Heterogeneity in the latent structure of PTSD symptoms among canadian veterans. *Psychological Assessment*, **22**, 666–674.

National Centre for Social Research (2009a). *Adult Psychiatric Morbidity in England, 2007: Results of a Household Survey*. http://www.ic.nhs.uk/statistics-and-data-collections/mental-health/mental-health-surveys/adult-psychiatric-morbidity-in-england-2007:-results-of-a-household-survey.

National Centre for Social Research (2009b). *Adult Psychiatric Morbidity in England, 2007: Results of a Household Survey, Appendices and Glossary*. http://www.ic.nhs.uk/statistics-and-data-collections/mental-health/mental-health-surveys/adult-psychiatric-morbidity-in-england-2007:-results-of-a-household-survey.

National Centre for Social Research (2010). *Adult Psychiatric Morbidity Survey 2007: Data Documentation*. http://www.esds.ac.uk/doc/6379%5Cmrdoc%5Cpdf%5C6379datadocs.pdf.

National Centre for Social Research & University of Leicester (2010). *Adult Psychiatric Morbidity Survey, 2007*. Colchester, Essex, UK Data Archive [distributor], http://dx.doi.org/10.5255/UKDA-SN-6379-1.

Nylund, K., Asparouhov, T. & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, **14**, 535–569.

Office for National Statistics (2001). *Psychiatric Morbidity among Adults Living in Private Households, 2000*. The Stationery Office, London.

Office for National Statistics (2003). *Psychiatric Morbidity among Adults Living in Private Households, 2000*. Colchester, Essex, UK Data Archive [distributor], http://dx.doi.org/10.5255/UKDA-SN-4653-1.

Office of Population Censuses and Surveys (1995). *OPCS Surveys of Psychiatric Morbidity in Great Britain, Report 1: The Prevalence of Psychiatric Morbidity among Adults Living in Private Households*. HMSO, London.

Office of Population Censuses and Surveys, Social Survey Division (1996). *OPCS Surveys of Psychiatric Morbidity: Private Household Survey, 1993*. Colchester, Essex, UK Data Archive [distributor], http://dx.doi.org/10.5255/UKDA-SN-3560-1.

Parker, G., Hadzipavlovic, D., Brodaty, H., Austin, M., Mitchell, P., Wilhelm, K. & Hickie, I. (1995). Sub-typing depression. 2. Clinical distinction of psychotic depression and nonpsychotic melancholia. *Psychol.Med.*, **25**, 825–832.

Parker, G., Wilhelm, K., Mitchell, P., Roy, K. & Hadzi-Pavlovic, D. (1999). Subtyping depression — Testing algorithms and identification of a tiered model. *J.Nerv.Ment.Dis.*, **187**, 610–617.

Partchev, I. (2012). *irtoys: Simple interface to the estimation and plotting of IRT models*. http://CRAN.R-project.org/package=irtoys, R package version 0.1.5.

Picardi, A., Viroli, C., Tarsitani, L., Miglio, R., de Girolamo, G., Dell'Acqua, G. & Biondi, M. (2012). Heterogeneity and symptom structure of schizophrenia. *Psychiatry Research*.

Piccinelli, M., Rucci, P., Ustun, B. & Simon, G. (1999). Typologies of anxiety, depression and somatization symptoms among primary care attenders with no formal mental disorder. *Psychol.Med.*, **29**, 677–688.

Pickles, A. & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Dev.Psychopathol.*, **15**, 529–551.

Piper, M.E., Bolt, D.M., Kim, S.Y., Japuntich, S.J., Smith, S.S., Niederdeppe, J., Cannon, D.S. & Baker, T.B. (2008). Refining the tobacco dependence phenotype using the wisconsin inventory of smoking dependence motives. *Journal of Abnormal Psychology*, **117**, 747–761.

Preston, K., Reise, S., Cai, L. & Hays, R.D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, **71**, 523 –550.

Prisciandaro, J. & Roberts, J. (2005). A taxometric investigation of unipolar depression in the national comorbidity survey. *J.Abnorm.Psychol.*, **114**, 718–728.

Prisciandaro, J. & Roberts, J. (2009). A comparison of the predictive abilities of dimensional and categorical models of unipolar depression in the National Comorbidity Survey. *Psychol.Med.*, **39**, 1087–1096.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Pickles, A. & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, **3**, 215–232.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Ramsay, J. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. http://www.psych.mcgill.ca/misc/fda/downloads/testgraf/TestGraf98.pdf, accessed: 28 May 2012.

Ranby, K.W., Boynton, M.H., Kollins, S.H., McClernon, F.J., Yang, C. & Fuemmeler, B.F. (2012). Understanding the phenotypic structure of adult retrospective ADHD symptoms during childhood in the united states. *Journal of Clinical Child & Adolescent Psychology*, **41**, 261–274.

Reise, S.P. & Waller, N.G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, **8**, 164–184.

Roberson-Nay, R. & Kendler, K.S. (2011). Panic disorder and its subtypes: a comprehensive analysis of panic symptom heterogeneity using epidemiological and treatment seeking samples. *Psychological Medicine*, **41**, 2411–2421.

Roy, M.A., Neale, M.C., Pedersen, N.L., Mathé, A.A. & Kendler, K.S. (1995). A twin study of generalized anxiety disorder and major depression. *Psychological Medicine*, **25**, 1037–1049.

Ruscio, J. & Ruscio, A. (2000). Informing the continuity controversy: A taxometric analysis of depression. *J.Abnorm.Psychol.*, **109**, 473–487.

Ruscio, J., Zimmerman, M., McGlinchey, J., Chelminski, I. & Young, D. (2007). Diagnosing major depressive disorder XI — A taxometric investigation of the structure underlying DSM-IV symptoms. *J.Nerv.Ment.Dis.*, **195**, 10–19.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, **34**.

Schilling, S. & Bock, R.D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, **70**, 533–555.

Schultz-Larsen, K., Kreiner, S. & Lomholt, R.K. (2007). Mini-mental status examination: Mixed rasch model item analysis derived two different cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology*, **60**, 268–279.

Shand, F.L., Slade, T., Degenhardt, L., Baillie, A. & Nelson, E.C. (2011). Opioid dependence latent structure: two classes with differing severity? *Addiction*, **106**, 590–598.

Sharp, C., Goodyer, I. & Croudace, T. (2006). The short mood and feelings questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology*, **34**, 365–377.

Shevlin, M., Adamson, G., Vollebergh, W., de Graaf, R. & van Os, J. (2007). An application of item response mixture modelling to psychosis indicators in two large community samples. *Social Psychiatry and Psychiatric Epidemiology*, **42**, 771–779.

Shorter, E. & Tyrer, P. (2003). Separation of anxiety and depressive disorders: blind alley in psychopharmacology and classification of disease. *British Medical Journal*, **327**, 158–160.

Silvia, P.J., Kaufman, J.C. & Pretz, J.E. (2009). Is creativity domain-specific? Latent class models of creative accomplishments and creative self-descriptions. *Psychology of Aesthetics, Creativity, and the Arts*, **3**, 139–148.

StataCorp (2007). *Stata Statistical Software: Release 10.1*. StataCorp LP, College Station, TX.

Sullivan, P. & Kendler, K. (1998). Typology of common psychiatric syndromes - An empirical study. *Br.J.Psychiatry*, **173**, 312–319.

Sullivan, P., Kessler, R. & Kendler, K. (1998). Latent class analysis of lifetime depressive symptoms in the National Comorbidity Survey. *Am.J.Psychiatry*, **155**, 1398–1406.

Sullivan, P., Prescott, C. & Kendler, K. (2002). The subtypes of major depression in a twin registry. *J.Affect.Disord.*, **68**, 273–284.

Thissen, D. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, **49**, 501–519.

Thissen, D., Cai, L. & Bock, R.D. (2010). The Nominal Categories Item Response Model. In M.L. Nering & R. Ostini, eds., *Handbook of polytomous item response theory models*, 43–76, Routledge, New York, NY.

Tueller, S. & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling: A Multidisciplinary Journal*, **17**, 165–192.

Tyrer, P. (1985). Neurosis divisible? *The Lancet*, **325**, 685–688.

Tyrer, P. (2001). The case for cothymia: mixed anxiety and depression as a single diagnosis. *Br.J.Psychiatry*, **179**, 191–193.

Tyrer, P., Seivewright, H. & Johnson, T. (2003). The core elements of neurosis: Mixed anxiety-depression (cothymia) and personality disorder. *Journal of Personality Disorders*, **17**, 129–138.

UCLA: Statistical Consulting Group (no date). *Stata FAQ: How can I convert a Stata file to an Mplus file?*. http://www.ats.ucla.edu/stat/stata/faq/stata2mplus.htm, accessed: 26 November 2012.

Uebersax, J. (no date). *Latent Class Analysis Frequently Asked Questions (FAQ)*. http://www.john-uebersax.com/stat/faq.htm#ident, accessed: 19 December 2012.

Unick, G., Snowden, L. & Hastings, J. (2009). Heterogeneity in comorbidity between major depressive disorder and generalized anxiety disorder and its clinical consequences. *J.Nerv.Ment.Dis.*, **197**, 215–224.

van Lang, N., Ferdinand, R., Ormel, J. & Verhulst, F. (2006). Latent class analysis of anxiety and depressive symptoms of the Youth Self-Report in a general population sample of young adolescents. *Behav.Res.Ther.*, **44**, 849–860.

Vermunt, J.K. (2007). *Multilevel mixture item response theory models: an application in education testing*. http://spitswww.uvt.nl/~vermunt/isi2007_vermunt.pdf, accessed: 30th July 2012.

Vermunt, J.K. & Magidson, J. (2002). Latent class cluster analysis. In J.A. Hagenaars & A.L. McCutcheon, eds., *Applied latent class analysis*, 89–106, Cambridge University Press, Cambridge.

Viroli, C. (2012). Using factor mixture analysis to model heterogeneity, cognitive structure, and determinants of dementia: an application to the aging, demographics, and memory study. *Statistics in Medicine*, **31**, 2110–2122.

Wadsworth, M.E., Hudziak, J.J., Heath, A.C. & Achenbach, T.M. (2001). Latent class analysis of child behavior checklist Anxiety/Depression in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, **40**, 106–114.

Walton, K., Ormel, J. & Krueger, R. (2011). The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models. *Journal of Abnormal Child Psychology*, **39**, 553–561.

Watson, D. (2005). Rethinking the mood and anxiety disorders: A quantitative hierarchical model for DSM-V. *J.Abnorm.Psychol.*, **114**, 522–536.

Weich, S., McBride, O., Hussey, D., Exeter, D., Brugha, T. & McManus, S. (2011). Latent class analysis of co-morbidity in the adult psychiatric morbidity survey in england 2007: Implications for DSM-5 and ICD-11. *Psychological Medicine*, **41**, 2201–2212.

Widiger, T. & Samuel, D. (2005). Diagnostic categories or dimensions? A question for the diagnostic and statistical manual of mental disorders-fifth edition. *J.Abnorm.Psychol.*, **114**, 494–504.

Wittchen, H., Beesdo-Baum, K., Gloster, A.T., Höfler, M., Klotsche, J., Lieb, R., Beauducel, A., Bühner, M. & Kessler, R.C. (2009). The structure of mental disorders re-examined: Is it developmentally stable and robust against additions? *International Journal of Methods in Psychiatric Research*, **18**, 189–203.

World Health Organization (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva.

Wu, L.T., Woody, G.E., Yang, C., Pan, J.J. & Blazer, D.G. (2011). Abuse and dependence on prescription opioids in adults: a mixture categorical and dimensional approach to diagnostic classification. *Psychological Medicine*, **41**, 653–664.

Wu, P.C. & Huang, T.W. (2010). Person heterogeneity of the BDI-II-C and its effects on dimensionality and construct validity: Using mixture item response models. *Measurement and Evaluation in Counseling and Development*, **43**, 155–167.

Zheng, M. & Baker, L.A. (2007). Investigating ADHD subtypes: An application of factor mixture modeling to twin data. In *Behavior Genetics Association 37th Annual Meeting Abstracts*, vol. 37, 6 of *Behavior Genetics*, 734–809.

Zhu, B.P., Lemeshow, S., Hosmer, D.W., Klar, J., Avrunin, J. & Teres, D. (1996). Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: A simulation study. *Crit.Care Med.*, **24**, 57–63.

Zinbarg, R., Barlow, D., Liebowitz, M., Street, L., Broadhead, E., Katon, W., Roybyrne, P., Lepine, J., Teherani, M., Richards, J., Brantley, P. & Kraemer, H. (1994). The DSM-IV field trial for mixed anxiety-depression. *Am.J.Psychiatry*, **151**, 1153–1162.

Zohar, J. & Westenberg, H.G.M. (2000). Anxiety disorders: a review of tricyclic antidepressants and selective serotonin reuptake inhibitors. *Acta Psychiatrica Scandinavica*, **101**, 39–49.