Statistical Methods in Medical Research 2008; **17**: 5–32

Classical latent variable models for medical research

Sophia Rabe-Hesketh Graduate School of Education and Graduate Group in Biostatistics, University of California, Berkeley, USA and Institute of Education, University of London, London, UK and **Anders Skrondal** Department of Statistics and The Methodology Institute, London School of Economics, London, UK and Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway

Latent variable models are commonly used in medical statistics, although often not referred to under this name. In this paper we describe classical latent variable models such as factor analysis, item response theory, latent class models and structural equation models. Their usefulness in medical research is demonstrated using real data. Examples include measurement of forced expiratory flow, measurement of physical disability, diagnosis of myocardial infarction and modelling the determinants of clients' satisfaction with counsellors' interviews.

1 Introduction

Latent variable modelling has become increasingly popular in medical research. By latent variable model we mean any model that includes unobserved random variables which can alternatively be thought of as random parameters. Examples include factor, item response, latent class, structural equation, mixed effects and frailty models.

Areas of application include longitudinal analysis[1], survival analysis[2], meta-analysis[3], disease mapping[4], biometrical genetics[5], measurement of constructs such as quality of life[6], diagnostic testing[7], capture–recapture models[8], covariate measurement error models[9] and joint models for longitudinal data and dropout.[10]

Starting at the beginning of the 20th century, ground breaking work on latent variable modelling took place in psychometrics.[11–13] The utility of these models in medical research has only quite recently been recognized and it is perhaps not surprising that medical statisticians tend to be unaware of the early, and indeed contemporary, psychometric literature.

In this paper we review classical latent variable models, namely common factor models and structural equation models (with continuous observed and continuous latent

Address for correspondence: Sophia Rabe-Hesketh, 3659 Tolman Hall, University of California, Berkeley, CA 94720-1670, USA. E-mail: sophiarh@berkeley.edu

Los Angeles, London, New Delhi and Singapore
10.1177/0962280207081236

variables), item response models (with categorical observed and continuous latent variables) and latent class models (with categorical observed and categorical latent variables) and demonstrate their usefulness in medical research using real data. We focus on classical latent variable models for several reasons: 1) the classical models can be fruitfully applied in medical research without modifications, 2) familiarity with classical latent variable models can help standardize terminology and thereby facilitate communication and perhaps more importantly, prevent misguided applications, 3) familiarity can reduce the risk of 'reinventing the wheel' and wasting resources on developing programs for problems that can be readily handled by standard software and 4) the classical models are important because they typically serve as building blocks for more advanced latent variable models. Although we consider mixed effects models[14,15] to be latent variable models, we do not include them here since they are so well known in medical statistics.

The plan of the paper is as follows. In Sections 2–5, we introduce classical latent variable models. In each of these sections we present examples using real data, mention common uses of the models in medical research and recommend further reading. Finally, we close the paper with some remarks regarding recent developments. A brief overview of useful software for latent variable modelling is given in an appendix.

2 Common factor models

2.1 Unidimensional factor models

Consider $j = 1, \ldots, N$ independent subjects. The *classical measurement model* from psychometric test theory,[16] also called the *parallel measurement model*, assumes that repeated measurements y_{ij} of the same *true score* $\beta + \eta_j$ for subject j are conditionally independent given η_j with conditional expectation $\beta + \eta_j$,

$$y_{ij} = \beta + \eta_j + \epsilon_{ij}, \qquad \mathrm{E}(\eta_j) = 0, \quad \mathrm{E}(\epsilon_{ij}) = 0, \quad \mathrm{Cov}(\eta_j, \epsilon_{ij}) = 0,$$

where β is the population mean true score, η_j are independently distributed random deviations of subjects' true scores from the population mean with variance ψ and ϵ_{ij} are independently distributed measurement errors with constant variance θ. The corresponding standard deviation $\sqrt{\theta}$ is called the standard error of measurement.

This model is appropriate if the repeated measurements are exchangeable. However, if y_{ij} are not merely replicates, but measurements using different instruments or raters i, it is likely that the instruments or raters use different scale origins and units. This situation is accommodated by the so-called *congeneric measurement model*[17]

$$y_{ij} = \beta_i + \lambda_i \eta_j + \epsilon_{ij}, \tag{1}$$

where the measure-specific mean, scale and measurement error variance are given by β_i, λ_i and θ_{ii}, respectively. The reliability ρ_i, the fraction of true score variance to total variance, for a particular instrument i becomes

$$\rho_i \equiv \frac{(\lambda_i)^2\psi}{(\lambda_i)^2\psi + \theta_{ii}}$$

Without any parameter constraints the model is not identified (several sets of parameter values can produce the same probability distribution) because multiplying the standard deviation $\sqrt{\psi}$ of the common factor by an arbitrary positive constant can be counteracted by dividing all factor loadings by the same constant. Identification is achieved either by anchoring, where the first factor loading is fixed to one, $\lambda_1 = 1$, or by factor standardization, where the factor variance is set to one, $\psi = 1$.

There are several important special cases of the congeneric measurement model. The *essentially tau-equivalent* measurement model is obtained if the scales of the instruments are identical $\lambda_i = 1$, the *tau-equivalent* measurement model if $\lambda_i = 1$ and the origins are identical $\beta_i = \beta$, and the parallel measurement model if $\lambda_i = 1$, $\beta_i = \beta$ and the measurement error variances are identical $\theta_{ii} = \theta$.

It should be noted that the congeneric measurement model is a unidimensional factor model with *factor loading* λ_i, *common factor* η_j and *unique factors* ϵ_{ij} for *indicators* y_{ij}. Spearman[11] introduced this model, arguing that intelligence is composed of a general factor, common to all subdomains such as mathematics, music, etc., and specific factors for each of the subdomains. The common factor can represent any hypothetical construct, a concept that cannot even in principle be directly observed, intelligence and depression[18] being prominent examples. In this case the measures i are typically questions or items of a questionnaire or structured interview.

The expectation of the vector of responses $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})'$ for subject j is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ and the model-implied covariance matrix, called the *factor structure*, becomes

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\boldsymbol{y}_j) = \boldsymbol{\Lambda}\psi\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2}$$

where $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\boldsymbol{\Theta}$ is a diagonal matrix with the θ_{ii} on the diagonal.

2.1.1 Estimation, goodness-of-fit and factor scoring

For maximum likelihood estimation, it is invariably assumed that the common and unique factors are normally distributed, implying a multivariate normal distribution for $\boldsymbol{y}_j$. Replacing the $\boldsymbol{\beta}$ in the likelihood by their maximum likelihood estimates, the sample means $\bar{\boldsymbol{y}}_.$, gives a profile likelihood

$$l^M(\boldsymbol{\Lambda}, \psi, \boldsymbol{\Theta}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp(-\frac{1}{2}\sum_{j=1}^{N}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_.)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_.)$$

In the case of complete data, the empirical covariance matrix $\mathbf{S}$ of $\boldsymbol{y}_j$ is the sufficient statistic for the parameters. It can be shown[19] that instead of maximizing the likelihood we can equivalently minimize the fitting function

$$\mathrm{F}_{\mathrm{ML}} = \log|\boldsymbol{\Sigma}| + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - n, \tag{3}$$

with respect to the unknown free parameters. The fitting function is non-negative and zero only if there is a perfect fit in the sense that the fitted $\boldsymbol{\Sigma}$ equals $\mathbf{S}$. When there are missing data, the 'case-wise' likelihood is maximized giving consistent estimates if data are missing at random (MAR).[20]

Treating the unstructured covariance matrix as the saturated model, a likelihood ratio goodness-of-fit test can be used. Under the null hypothesis that the restricted model of interest is correct (and suitable regularity conditions), the likelihood ratio statistic or deviance has a chi-square distribution with $n(n+1)/2 - p$ degrees of freedom, where p is the number of parameters in the restricted model (after eliminating the intercepts), equal to $2n$ in the congeneric measurement model. Since most parsimonious and hence appealing models are rejected in large samples, more than a hundred goodness-of-fit indices[21] have been proposed that are relatively insensitive to sample size. Some of these indices (e.g., root mean square error of approximation (RMSEA)[22]) are based on a comparison with the saturated model, but most (e.g., comparative fit index (CFI)[23], Tucker-Lewis index (TLI)[24]) are based on a comparison with the unrealistic null model of uncorrelated items, and are thus analogous to the coefficient of determination in linear regression. Arbitrarily, fit is typically considered 'good' if the RMSEA is below 0.05 and the CFI and TLI are above 0.90.

Often the purpose of factor modelling is to investigate the measurement properties of different raters, instruments or items i by comparing the fit of congeneric, tau-equivalent and parallel models and estimating reliabilities. Another purpose is measurement itself, that is assigning factor scores to individual subjects j after estimating the model. The most common approach is the so-called regression method,

$$\widetilde{\eta}_j = \widehat{\psi}\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_j - \widehat{\boldsymbol{\beta}}).$$

This can motivated as the expectation of the posterior distribution of the common factor given the observed responses with parameter estimates plugged in, known as the empirical Bayes predictor or shrinkage estimator in statistics. An alternative approach is the Bartlett method,

$$\widehat{\eta}_j = (\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Theta}}^{-1}\widehat{\boldsymbol{\Lambda}})^{-1}\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Theta}}^{-1}(\mathbf{y}_j - \widehat{\boldsymbol{\beta}}),$$

which maximizes the likelihood of the responses given the common factor with parameter estimates plugged in.

In the classical measurement model, factor scores using either the regression or Bartlett method are perfectly correlated with the simple sum score $\sum_i y_{ij}$. For this model, Cronbach's α, a commonly used measure of internal consistency of a scale, can be interpreted as the reliability of the sum score.

2.1.2 Example: Forced expiratory flow

In the Health Survey of England 2004 (National Centre for Social Research and University College London, Department of Epidemiology and Public Health[25]), a sample of children between 7 and 15 years of age had their lung function assessed using a

Vitalograph Micro Spirometer. Here, we consider forced expiratory flow which was measured five times for 89 of the children. The nurse also recorded at each occasion whether or not the technique used was satisfactory. If not satisfactory, we treat the measurement as missing.

Table 1 shows maximum likelihood estimates for the classical, essentially tau-equivalent, tau-equivalent and congeneric measurement models. Using likelihood ratio tests at the 5% level, the essentially tau-equivalent measurement model is selected. This model also has the best fit indices, but the RMSEA is larger than the desired 0.05. Note that the estimated mean increases over time are consistent with a practice effect and that the estimated measurement error variance is very low at the third measurement. The estimated reliabilities range between 0.82 at the first measurement and 0.96 at the third.

2.2 Multidimensional factor models

Hypothetical constructs are often multidimensional, comprising several related aspects. An example is fatigue, for which Nisenbaum *et al.*[26] argue that, there are three correlated aspects or common factors defined as 'fatigue-mood-cognition', 'flu-type', and 'visual-impairment'.

Table 1 Forced expiratory flow–Maximum likelihood estimates for measurement models

	Parallel/ Classical		Tau-equivalent		Essentially Tau-equivalent		Congeneric	
	Est	(SE)	Est	(SE)	Est	(SE)	Est	(SE)
Intercepts								
β_1	257	(11)	261	(11)	231	(12)	232	(12)
β_2	257	(11)	261	(11)	250	(12)	250	(12)
β_3	257	(11)	261	(11)	262	(11)	263	(11)
β_4	257	(11)	261	(11)	271	(12)	271	(12)
β_5	257	(11)	261	(11)	272	(12)	272	(12)
Factor loadings								
λ_1	1		1		1		1	
λ_2	1		1		1		1.1	(0.08)
λ_3	1		1		1		1.1	(0.07)
λ_4	1		1		1		1.2	(0.08)
λ_5	1		1		1		1.1	(0.08)
Factor variance								
ψ	10500	(1638)	10773	(1660)	10934	(1685)	9024	(1711)
Meas. error variances								
θ_{11}	1772	(143)	3519	(616)	2429	(440)	2372	(423)
θ_{22}	1772	(143)	1743	(325)	1527	(289)	1548	(295)
θ_{33}	1772	(143)	386	(144)	435	(141)	445	(142)
θ_{44}	1772	(143)	1302	(259)	1193	(239)	1125	(237)
θ_{55}	1772	(143)	1958	(364)	1818	(336)	1783	(335)
Goodness-of-fit statistics								
Log-likelihood	−2182.7		−2162.8		−2142.7		−2140.4	
Deviance (d.f.)	106.97 (17)		67.05 (13)		26.95 (9)		22.35 (5)	
CFI	0.85		0.91		0.97		0.97	
TLI	0.91		0.93		0.97		0.94	
RMSEA	0.24		0.22		0.15		0.20	

Denoting the common factors as $\boldsymbol{\eta}_j = (\eta_{1j}, \eta_{2j}, \ldots, \eta_{mj})'$, the multidimensional common factor model can be written as

$$\boldsymbol{y}_j = \boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\epsilon}_j, \tag{4}$$

where $\boldsymbol{\Lambda}$ is now a $n \times m$ matrix of factor loadings with element pertaining to item i and latent variable l denoted λ_{il} and $\boldsymbol{\epsilon}_j$ a vector of unique factors. We define $\boldsymbol{\Psi} \equiv \mathrm{Cov}(\boldsymbol{\eta}_j)$ and assume that $\mathrm{E}(\boldsymbol{\eta}_j) = 0$, $\mathrm{E}(\boldsymbol{\epsilon}_j) = 0$, and $\mathrm{Cov}(\boldsymbol{\eta}_j, \boldsymbol{\epsilon}_j) = \mathbf{0}$. Note that the structure of the model is similar to a linear mixed model where $\boldsymbol{\Lambda}$ is replaced by a covariate matrix $\mathbf{Z}_j$.[27] The results presented in Section 2.1 from Equation (2) apply also to the multidimensional factor model after substituting $\boldsymbol{\Psi}$ for ψ.

It is important to distinguish between two approaches to common factor modelling; exploratory factor analysis and confirmatory factor analysis (CFA).

Exploratory factor analysis

Exploratory factor analysis (EFA)[12,13] is an inductive approach for 'discovering' the number of common factors and estimating the model parameters, imposing a minimal number of constraints for identification. The standard identifying constraints are that $\boldsymbol{\Psi} = \mathbf{I}$, and that $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$ are both diagonal. Although mathematically convenient, the one-size-fits-all parameter restrictions imposed in EFA, particularly the specification of uncorrelated common factors, are often not meaningful from a subject matter point of view.

Methods of estimation include maximum likelihood using the fitting function in (3) and principal axis factoring. However, due to confusion between factor analysis and principal component analysis, the latter is often used. In a second step, the restrictions imposed for identification are relaxed by rotating the factors to achieve equivalent models that are more interpretable by different criteria, such as varimax.[28]

Confirmatory factor analysis

In CFA,[17,29] restrictions are imposed based on substantive theory or research design. This approach was illustrated for the unidimensional factor model in Section 2.1.2. In the multidimensional case an important example of a restricted model is the *independent clusters* or *sets of congeneric measures* model where $\boldsymbol{\Lambda}$ has many elements set to zero such that each indicator measures one and only one factor. Such a configuration makes sense if one set of indicators is designed to measure one factor and another set of indicators to measure another factor. A path diagram of a two-factor independent clusters model with three variables measuring each factor is given in Figure 1. Circles represent latent variables, rectangles represent observed variables, arrows connecting circles and/or rectangles represent regressions and short arrows pointing at circles or rectangles represent residuals. Curved double-headed arrows connecting two variables indicate that they are correlated.

For identification the scales of the common factors are fixed either by factor standardization or by anchoring. Elements of $\boldsymbol{\epsilon}_j$ may be correlated, but special care must be exercised in this case to ensure that the model is identified. It is typically assumed that the common and unique factors have multivariate normal distributions to allow maximum likelihood estimation as outlined in Section 2.1.

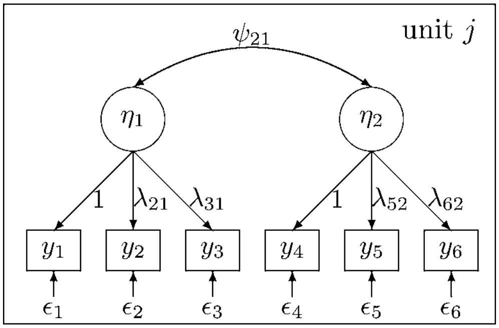

Figure 1 Independent clusters factor model.

Although routinely used, the standard likelihood ratio test is invalid for testing dimensionality since the null hypothesis is on the border of parameter space and thus violates the regularity conditions.[30,31] For this reason fit indices become important.

For further reading on classical test theory and common factor models we recommend Streiner and Norman,[32] Dunn[33–36] and McDonald.[28,37]

2.2.1 *Example: Brain scans*

It has been observed in a number of studies that people suffering of schizophrenia have enlarged brain ventricular volumes. Computerized Axial Tomography scans of heads of 50 psychiatric patients were performed to determine the ventricle-brain ratio from measurements of the perimeter of the ventricle and the perimeter of the inner surface of skull.[36,38] In a method comparison study a standard method involving a hand-held planimeter on a projection of the X-ray image was compared with a new method using an automated pixel count based on a digitized image. Both methods were replicated twice. The log-transformed measurements for the standard method are denoted 'Plan1' (y_1) and 'Plan3' (y_2) and the log-transformed measurements for the new method 'Pix1' (y_3) and 'Pix3' (y_4).

The analyses presented here are similar to those by Dunn.[36] Initially, a one-factor model (1) was considered giving the first set of estimates in Table 2. The factor loadings and measurement error variances for the Plan measurements seem to be different from the loadings and measurement error variances for the Pix measurements. We also note that the estimated measurement error variances are small for all four measures and perhaps not surprisingly, smaller for the automated Pix measurements than for the standard method. The deviance statistic is 29.36 with 2 degrees of freedom leading to rejection which is reinforced by small and large TLI and RMSEA, respectively.

The Pix measurements are sometimes consistent with each other but very different from the Plan measurement which suggests that there may be some 'gross error'. Dunn and Roberts[34] discuss such phenomena in method comparison data and refer to them as random matrix effects or method by subject interactions. To investigate this, we extend the one-factor model to include correlated measurement errors for the Pix measurements (see left panel of Figure 2), giving the second sets of estimates in Table 2. The parameter θ_{43}, representing the covariance between the measurement errors ϵ_3 and ϵ_4, is estimated

Table 2 Brain scans – Maximum likelihood estimates (with standard errors) for common factor models

	One-factor							
	One-factor		Corr. meas. errors		Two-factor			
Intercepts								
β_1 [Plan1]	1.862	(0.056)	1.862	(0.056)	1.862	(0.056)		
β_2 [Plan3]	1.711	(0.061)	1.711	(0.061)	1.711	(0.061)		
β_3 [Pix1]	1.402	(0.073)	1.402	(0.073)	1.402	(0.073)		
β_4 [Pix3]	1.413	(0.074)	1.413	(0.074)	1.413	(0.074)		
Factor loadings					Factor 1		Factor 2	
λ_1 [Plan1]	1	–	1	–	0	–	0	–
λ_2 [Plan3]	1.349	(0.315)	1.377	(0.208)	1.377	(0.208)	0	–
λ_3 [Pix1]	2.186	(0.423)	1.213	(0.208)	1.213	(0.208)	1	–
λ_4 [Pix3]	2.197	(0.425)	1.209	(0.210)	1.209	(0.210)	1	–
Factor variances								
ψ_{11}	0.056	(0.024)	0.098	(0.031)	0.098	(0.031)		
ψ_{22}					0.123	(0.029)		
ψ_{21}					0	–		
Measurement error variance								
θ_{11} [Plan1]	0.103	(0.021)	0.060	(0.016)	0.060	(0.016)		
θ_{22} [Plan3]	0.086	(0.017)	0.001	(0.019)	0.001	(0.019)		
θ_{33} [Pix1]	0.001	(0.003)	0.122	(0.029)	[a]−0.001	(0.003)		
θ_{44} [Pix3]	0.002	(0.003)	0.127	(0.029)	0.004	(0.003)		
Measurement error covariance								
θ_{43} [Pix1], [Pix3]			0.123	(0.029)				
Goodness-of-fit statistics								
Log-likelihood	11.56		24.95		24.95			
Deviance (d.f.)	29.36 (2)		2.57 (1)		2.57 (1)			
CFI	0.91		1.00		1.00			
TLI	0.73		0.97		0.97			
RMSEA	0.52		0.18		0.18			

[a] Heywood case

as 0.123, corresponding to an extremely high measurement error correlation of 0.988. Interestingly, the measurement error variances are now much larger for the Pix measurements than for the Plan measurements. The deviance statistic for this model is 2.57 with 1 degree of freedom, giving a P-value of 0.10. Although the RMSEA remains somewhat high, the CFI and TLI are now very satisfactory. Thus, it seems reasonable to retain the measurement model with correlated measurement errors for the Pix measurements.

Another way of capturing the notion of gross error is to introduce a ‘bias factor’ η_{2j} which induces extra dependence among the Pix measurements. The bias factor has factor loadings of one for the two Pix measurements and zero for the two Plan measurements and is uncorrelated with the original factor (see right panel of Figure 2). Maximum likelihood estimates with standard errors for this model are presented in the last columns of Table 2. Note that the estimated measurement error variance for [Pix1] is inadmissible since it is negative, a so-called Heywood case, which casts some doubt on the validity of this specification. Interestingly, the deviance statistic for this model is identical to that for the measurement model with correlated measurement errors. It can be shown that

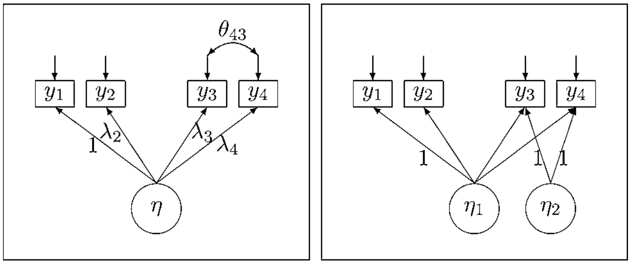

Figure 2 Brain scans – Path diagrams for one-factor model with correlated measurement error for Pix measurement (left panel) and model with 'bias factor' (right panel).

the two models are equivalent[39] in the sense that the models are one-to-one reparameterizations of each other, and there is consequently no way to empirically distinguish between the models.

2.2.2 *Some applications of factor analysis in medical research*

Classical test theory is used implicitly every time a test-retest or inter-rater reliability is reported. Classical test theory and factor models have been used for psychiatric rating scales for a long time.[40] Bland and Altman[41] popularized these ideas in medicine, but without referring to the vast measurement literature. It is standard practice to report Cronbach's α whenever forming a sum score to assess the 'internal consistency' or inter-correlatedness of the items. This is done regardless of the measurement level of the items and without being aware of the underlying model assumptions.

EFA and CFA are often used to explore the dimensionality of constructs, particularly in psychiatry[42,43] and related fields. In EFA the factor structure is usually simplified into an independent clusters model by effectively setting small factor loadings to zero (after rotation) without considering the deterioration in fit. In contrast, arbitrary but 'well established' thresholds applied to a couple of goodness-of-fit indices usually serve to purge uncertainty from model selection in CFA.

3 Item response theory (IRT) models

The term item response theory was coined by Lord[44] who was instrumental in developing statistical models for ability testing.[44–46] In this setting, the observed variables are exam questions or items, y_{ij} is '1' if examinee j answered item i correctly and '0' otherwise, and η_j represents the continuous unobserved ability of the examinee. In medical research η_j would of course be another latent trait such as physical functioning.

3.1 The one-parameter IRT model

In the one-parameter logistic (1-PL) model, the conditional response probability for item i, given ability η_j, is specified as

$$\Pr(y_{ij} = 1 \mid \eta_j) = \frac{\exp(\beta_i + \eta_j)}{1 + \exp(\beta_i + \eta_j)},$$

where the responses are assumed to be conditionally independent given ability η_j, a property often called *local independence*. This model is called a *one–parameter model* because there is one parameter, the *item difficulty* $-\beta_i$, for each item.

Ability η_j can either be treated as a latent random variable or as an unknown fixed parameter, giving the so-called Rasch model.[47] However, estimating the 'incidental parameters' η_j jointly with the 'structural parameters' $\boldsymbol{\beta}$ produces inconsistent estimators for $\boldsymbol{\beta}$, known as the incidental parameter problem.[48] Inference can instead be based on the conditional likelihood[49,50] constructed by conditioning on the sufficient statistic for η_j, which is the sum score $\sum_{i=1}^{n} y_{ij}$. This conditional logistic regression approach is also commonly used in matched case control studies.[51]

Alternatively, if η_j is viewed as a random variable, typically with $\eta_j \sim N(0, \psi)$, inference can be based on the marginal likelihood obtained by integrating out the latent variable

$$l^{M}(\boldsymbol{\beta}, \psi) = \prod_{j=1}^{N} \Pr(\boldsymbol{y}_j; \boldsymbol{\beta}, \psi) = \prod_{j=1}^{N} \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Pr(y_{ij} \mid \eta_j)\, g(\eta_j; \psi)\, \mathrm{d}\eta_j$$

$$= \prod_{j=1}^{N} \int_{-\infty}^{\infty} \prod_{i=1}^{n} \frac{\exp(\beta_i + \eta_j)^{y_{ij}}}{1 + \exp(\beta_i + \eta_j)}\, g(\eta_j; \psi)\, \mathrm{d}\eta_j,$$

where $-\,\boldsymbol{\beta}$ is the vector of item difficulties and $g(\cdot; \psi)$ is the normal density with zero mean and variance ψ.

An appealing feature of one-parameter models is that items and examinees can be placed on a common scale (according to the difficulty and ability parameters) so that the probability of a correct response depends only on the amount $\eta_j - (-\beta_i)$ by which the examinee's position exceeds the item's position. Differences in difficulty between items are the same for all examinees, and differences in abilities of two examinees are the same for all items, a property called *specific objectivity* by Rasch.

One of the main purposes of item response modelling is to derive scores from the item responses. The most common approach is to use the expectation of the posterior distribution of η_j given $\boldsymbol{y}_j$ with parameter estimates plugged in. This empirical Bayes approach is analogous to the regression method for common factor models and is referred to as expected a posterior (EAP) scoring in IRT.

3.1.1 Example: Physical functioning

The 1996 Health Survey for England (Joint Health Surveys Unit of Social and Community Planning Research and University College London[52]) administered the Physical Functioning PF-10 subscale of the SF-36 Health Survey[53] to adults aged 16 and above. The 15592 respondents were asked:

'The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?'

1) Vigorous activities: Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports

2) Moderate activities: Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf
3) Lift/Carry: Lifting or carrying groceries
4) Several stairs: Climbing several flights of stairs
5) One flight stairs: Climbing one flight of stairs
6) Bend/Kneel/Stoop: Bending, kneeling, or stooping
7) Walk more mile: Walking more than a mile
8) Walk several blocks: Walking several blocks
9) Walk one block: Walking one block
10) Bathing/Dressing: Bathing or dressing yourself

The possible responses to these questions are ‘yes, limited a lot’, ‘yes, limited a little’ and ‘no, not limited at all’. Here we analyse the dichotomous indicator for not being limited at all. Then $-\beta_i$ can be interpreted as the difficulty of activity i and η_j as the physical functioning of subject j.

The ten most frequent response patterns and their frequencies are shown in the first two columns of Table 3. Note that about a third of subjects are not limited at all in any of the activities whereas 6% are limited in all activities. Using conditional maximum likelihood (CML) estimation with the identifying constraint $\beta_1 = 0$ gives the difficulty estimates shown under ‘CML’ in Table 4.

For marginal maximum likelihood estimation it seems unrealistic to assume a normal distribution for physical functioning since such a large proportion of people is not limited at all on any of the items. We therefore leave the physical functioning distribution unspecified using so-called nonparametric maximum likelihood (NPML) estimation.[54–58] The NPML estimator of the latent variable distribution is discrete with number of mass-points determined to maximize the marginal likelihood. Here six points appeared to be needed but convergence was not achieved, so we present the discrete five-point solution instead.

Figure 3 shows the item characteristic curves, the probabilities of not being limited at all as a function of physical functioning, for four of the items. The value of physical functioning corresponding to a probability of 0.5 is the difficulty of the task. Not

Table 3 Physical functioning – Observed and expected counts for the ten most frequent response patterns with signed Pearson residuals in parentheses

	Observed	Discrete 1-PL		Normal 1-PL		Normal 2-PL	
Responses	Count	Count	(Resid)	Count	(Resid)	Count	(Resid)
1111111111	5384	5137	(3.4)	5217	(2.3)	5036	(4.9)
0111111111	1884	1670	(5.2)	1632	(6.2)	2091	(−4.5)
0000000000	970	1052	(−2.5)	813	(5.5)	978	(−0.3)
0110111111	646	518	(5.6)	463	(8.5)	529	(5.1)
0111101111	441	275	(10.0)	240	(13.0)	347	(5.0)
0000000001	343	294	(2.9)	298	(2.6)	376	(−1.7)
0110101111	252	205	(3.3)	190	(4.5)	218	(2.3)
0110110111	237	137	(8.5)	126	(9.9)	123	(10.3)
0000000011	220	197	(1.6)	229	(−0.6)	193	(1.9)
1110111111	209	264	(−3.4)	228	(−1.3)	309	(−5.7)

Table 4 Physical functioning – Estimates of item difficulties for the 1-PL model using conditional maximum likelihood (CML), marginal maximum likelihood with discrete five-point distribution (DMML) and parametric marginal maximum likelihood (PMML) assuming a normal ability distribution

		CML		DMML		PMML	
		Est	(SE)	Est	(SE)	Est	(SE)
Vigorous activities	$-\beta_1$	0		1.23	(0.04)	0.86	(0.05)
Moderate activities	$-\beta_2+\beta_1$	−3.50	(0.06)	−3.50	(0.05)	−3.65	(0.05)
Lift/Carry	$-\beta_3+\beta_1$	−3.06	(0.04)	−3.06	(0.04)	−3.22	(0.04)
Several stairs	$-\beta_4+\beta_1$	−1.85	(0.04)	−1.84	(0.04)	−1.97	(0.04)
One flight stairs	$-\beta_5+\beta_1$	−4.31	(0.05)	−4.30	(0.04)	−4.43	(0.05)
Bend/Kneel/Stoop	$-\beta_6+\beta_1$	−2.48	(0.04)	−2.48	(0.04)	−2.63	(0.04)
Walk more mile	$-\beta_7+\beta_1$	−2.88	(0.04)	−2.87	(0.04)	−3.03	(0.04)
Walk several blocks	$-\beta_8+\beta_1$	−4.27	(0.05)	−4.27	(0.05)	−4.40	(0.05)
Walk one block	$-\beta_9+\beta_1$	−5.68	(0.06)	−5.68	(0.06)	−5.69	(0.06)
Bathing/Dressing	$-\beta_{10}+\beta_1$	−6.34	(0.06)	−6.35	(0.06)	−6.28	(0.06)

surprisingly, bathing and dressing oneself are easier than vigorous activities. Wilson[59] advocates plotting a so-called Wright map, a histogram of the ability scores on the same vertical scale as positions of the items. Such a graph is given together with the estimated ability distribution in Figure 4.

Table 3 gives the observed and expected frequencies for the ten most frequent response patterns for the model with a discrete latent variable discussed here, the 1-PL model assuming a normal distribution, and the 2-PL model discussed in the next section. The signed Pearson residuals given in parentheses suggest that none of the models fit well.

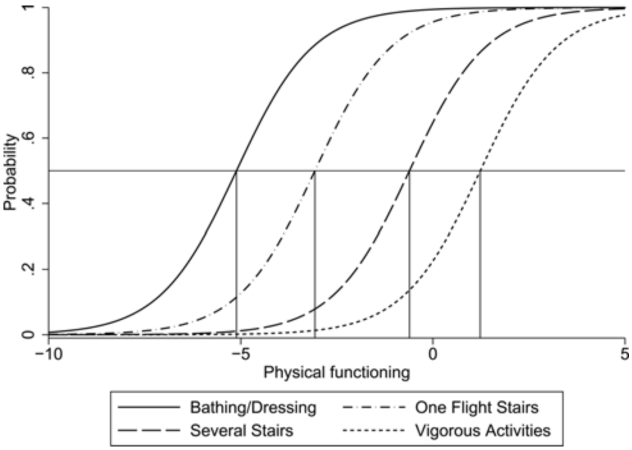

Figure 3 Physical functioning – Item characteristic curves for 1-PL with discrete five-point latent variable distribution.

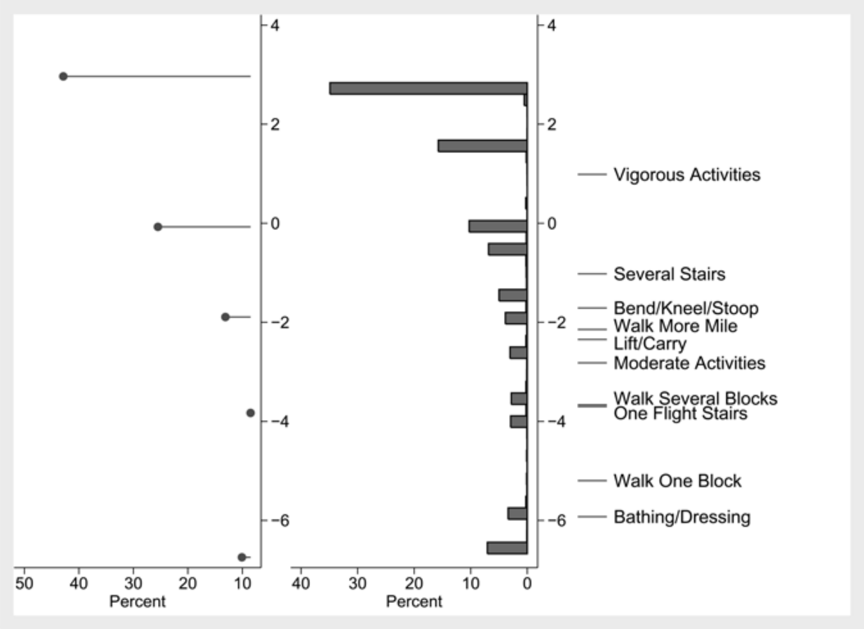

Figure 4 Physical functioning – Estimated distribution (1-PL with five masspoints), distribution of empirical Bayes scores and difficulties of items.

3.2 Two-parameter IRT model

Although the 1-PL model is parsimonious and elegant it often does not fit the data. A more general model is the two-parameter logistic (2-PL) model

$$\Pr(y_{ij} = 1 \mid \eta_j) = \frac{\exp(\beta_i + \lambda_i \eta_j)}{1 + \exp(\beta_i + \lambda_i \eta_j)},$$

where the factor loading λ_i is referred to as the discrimination parameter because items with a large λ_i are better at discriminating between subjects with different abilities. The traditional parametrization of the conditional log-odds $\beta_i + \lambda_i \eta_j$ is $\lambda_i(\eta_j - b_i)$, where $b_i = -\beta_i/\lambda_i$ is called the difficulty parameter because the probability of a correct response is 0.5 if $\eta_j = b_i$ as in the one-parameter model. The model is called a *two-parameter model* because there are two parameters for each item, the discrimination and difficulty parameter.

Importantly, a conditional likelihood can no longer be constructed since there is no sufficient statistic for η_j and the marginal likelihood assuming that $\eta_j \sim N(0, 1)$ is hence used. Since there is no closed form of the marginal likelihood, Bock and Lieberman[60] introduced the use of numerical integration by Gauss–Hermite quadrature. Here, we use a refinement of this approach called adaptive quadrature.[61]

For the PF-10 example, item characteristic curves for the 2-PL model, assuming a normal distribution for physical functioning, are shown in Figure 5 for the same four items as in Figure 3. For the items shown here, *Vigorous Activities* is the least discriminating

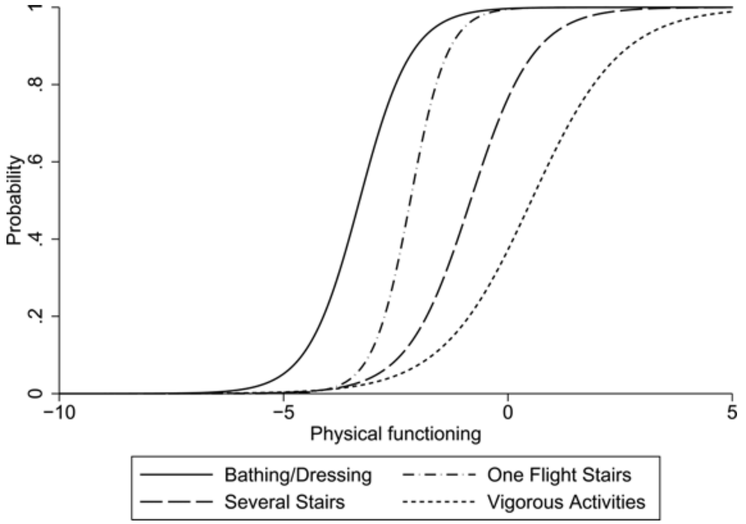

Figure 5 Physical functioning – Item characteristic curves for 2–PL.

whereas *One Flight Stairs* is the most discriminating. Note that the 2-PL model does not share the specific objectivity property of the 1-PL model. An item can be easier than another item for low abilities but more difficult than the other item for higher abilities due to the item-subject interaction $\lambda_i \eta_j$. This property has caused some psychometricians to reject the use of discrimination parameters because they 'wreak havoc with the logic and practice of measurement'.[62]

For dichotomous responses, it is often instructive to formulate models using a latent continuous response y^*_{ij} underlying the dichotomous response with

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Specifying a unidimensional factor model for y^*_{ij}

$$y^*_{ij} = \beta_i + \lambda_i \eta_j + \epsilon_{ij}, \quad \eta_j \sim \mathrm{N}(0, \psi),\ \epsilon_{ij} \sim \mathrm{N}(0, 1),\ \mathrm{Cov}(\eta_j, \epsilon_{ij}) = 0,$$

leads to the so-called normal-ogive model[45,63] for the observed y_{ij},

$$\Phi^{-1}[\Pr(y_{ij} = 1 \mid \eta_j)] = \beta_i + \lambda_i \eta_j$$

Here $\Phi(\cdot)$ is the cumulative standard normal distribution function and $\Phi^{-1}(\cdot)$ is the probit link. Replacing the probit link by a logit link yields the 2-PL model which can also be derived by specifying logistic distributions for the specific factors ϵ_{ij} in the latent response formulation.

For ordinal responses, proportional odds and adjacent category logit versions of the IRT models described above have been proposed under the names graded response model[64] and partial credit model,[65] respectively.

We recommend the books by Hambleton *et al.* [66] and Embretson and Reise[67] for further reading on IRT.

3.2.1 Some applications of IRT in medical research

IRT is sometimes used in psychiatry for investigating the measurement properties of rating scales, for instance for affective disorder[68] and depression.[69,70] More recently there has been a surge of interest in IRT for quality of life and related constructs. For example, the 2004 Conference on Outcome Research organized by the US National Cancer Institute was dedicated to IRT. In 2003 there was a special issue of *Quality of Life Research* on the measurement of headache which was dominated by papers using the Rasch model.

Papers advocating the use of IRT in health outcome measurement include Revicki and Cella[71] and Hays *et al.*[72] in *Quality of Life Research* and *Medical Care*, respectively, journals that regularly publish papers using IRT. However, the use of IRT for quality of life has been criticized because items, such as those concerning impairment, may be better construed as ‘causing’ quality of life than ‘reflecting’ quality of life.[6]

Item response models have also been used for other purposes than measurement such as capture–recapture modelling[73] for estimation of prevalences, typical examples being diabetes or drug use.

4 Latent class models

4.1 Exploratory latent class model

In latent class models[74,75] both the latent and observed variables are categorical. The C categories of the latent variable can be thought of as labels $c = 1, \ldots, C$ classifying subjects into distinct sub-populations. For simplicity, we consider binary response variables $y_{ij}, i = 1, \ldots, n$. The conditional response probability for variable i, given latent class membership c, is given by

$$\Pr(y_{ij} = 1 \mid c) \;=\; \pi_{i|c}, \tag{5}$$

where $\pi_{i|c}$ are free parameters and different responses y_{ij} and $y_{i'j}$ for the same subject are conditionally independent given class membership. The probability that subject j belongs to latent class c is also a free parameter, denoted π_c. The model is called exploratory because no restrictions are imposed on either the $\pi_{i|c}$ or the π_c.

The marginal likelihood becomes

$$l^{M}(\boldsymbol{\pi}) = \prod_{j=1}^{N} \Pr(\boldsymbol{y}_j; \boldsymbol{\pi}) = \prod_{j=1}^{N} \sum_{c=1}^{C} \pi_c \prod_{i=1}^{n} \Pr(y_{ij} \mid c) = \prod_{j=1}^{N} \sum_{c=1}^{C} \pi_c \prod_{i=1}^{n} \pi_{i|c}^{y_{ij}} (1 - \pi_{i|c})^{1-y_{ij}},$$

where $\boldsymbol{\pi}' = (\pi_1, \pi_{1|1}, \ldots, \pi_{n|1}, \ldots, \pi_C, \pi_{1|C}, \ldots, \pi_{n|C})$. It is evident that the latent class model is a multivariate finite mixture model with C components.

For further reading on latent class models we recommend the book by McCutcheon[76] and the survey papers by Clogg[77] and Formann and Kohlmann,[78] the latter with special emphasis on applications in medicine.

4.1.1 Example: Diagnosis of myocardial infarction

Rindskopf and Rindskopf[7] analyse data from a coronary care unit in New York City where patients were admitted to rule out myocardial infarction (MI) or 'heart attack'. Each of 94 patients was assessed on four diagnostic criteria rated as 1 = present and 0 = abscent:

- [Q-wave]–Q-wave in the ECG
- [History]–Classical clinical history
- [LDH]–Having a flipped LDH
- [CPK]–CPK-MB

The data are shown in Table 5. For clarity we label the classes $c = 1$ for MI and $c = 0$ for no MI. Then π_0 is the probability or 'prevalence' of not having MI, which can be parameterized as

$$\text{logit}(\pi_0) = \varrho_0,$$

and $\pi_1 = 1 - \pi_0$. The conditional response probabilities can be specified as

$$\text{logit}[\Pr(y_{ij} = 1|c)] = e_{ic}$$

The probabilities $\Pr(y_{ij} = 1|c = 1)$ represent the sensitivities of the diagnostic tests (the probabilities of a correct diagnosis for subjects with the illness), whereas $1 - \Pr(y_{ij} = 1|c = 0)$ represent the specificities (the probabilities of a correct diagnosis for subjects without the illness).

Table 5 Diagnosis of myocardial infarction data

[Q-wave] ($i = 1$)	[History] ($i = 2$)	[LDH] ($i = 3$)	[CPK] ($i = 4$)	Observed count	Expected count	Probability of MI ($c = 1$)
1	1	1	1	24	21.62	1.000
0	1	1	1	5	6.63	0.992
1	0	1	1	4	5.70	1.000
0	0	1	1	3	1.95	0.889
1	1	0	1	3	4.50	1.000
0	1	0	1	5	3.26	0.420
1	0	0	1	2	1.19	1.000
0	0	0	1	7	8.16	0.044
1	1	1	0	0	0.00	0.017
0	1	1	0	0	0.22	0.000
1	0	1	0	0	0.00	0.001
0	0	1	0	1	0.89	0.000
1	1	0	0	0	0.00	0.000
0	1	0	0	7	7.78	0.000
1	0	0	0	0	0.00	0.000
0	0	0	0	33	32.11	0.000

Source: Rindskopf and Rindskopf (1986).[7]

Maximum likelihood estimates are given in Table 6. The estimates $\widehat{e}_{10}$ for Q-wave and $\widehat{e}_{41}$ for [CPK] correspond to conditional response probabilities very close to 0 and 1, respectively, giving a so-called boundary solution. In these regions, large changes in the logit correspond to minute changes in the probability, leading to a flat likelihood and thus large standard errors for these parameters. Comparing the expected counts with the observed counts in Table 5, the model appears to fit well.

From Table 6, the prevalence of MI is estimated as 0.46. The specificity of [Q-wave] is estimated as 1, implying that all patients without MI will have a negative result on that test. [History] has the lowest specificity of 0.70. The estimated sensitivities range from 0.77 for Q-wave to 1.00 for [CPK], so that 77% of MI cases test positively on [Q-wave] and 100% on [CPK].

We can obtain the posterior probabilities (similar to positive predictive values) of MI given the four test results using Bayes theorem,

$$\Pr(c = 1|\mathbf{y}_j) = \frac{\pi_1 \prod_i \Pr(y_{ij}|c = 1)}{\pi_0 \prod_i \Pr(y_{ij}|c = 0) + \pi_1 \prod_i \Pr(y_{ij}|c = 1)}$$

These probabilities are presented in the last column of Table 5. For almost all patients the posterior probabilities are close to 0 or 1 making the diagnosis clear cut.

4.1.2 Some applications of latent class analysis in medical research

An important application of latent class models is in medical diagnosis[7] as demonstrated in the example. Latent class models are also often used to investigate rater agreement[78,79] such as agreement among pathologists in their classification of tumors, and for scaling,[80] for instance for investigating the stages and pathways of drug involvement.[81]

Extensions of latent class models for longitudinal data include latent transition or latent Markov models[82–84] for discovering latent states and estimating transition probabilities between them. Finite mixtures of regression models or linear mixed models, known as mixture regression models[56,85,86] and growth mixture models,[87] respectively, are used for discovering 'latent trajectory classes'.

Table 6 Myocardial infarction – Maximum likelihood estimates for latent class model with two classes

	Class 0 ('No MI')			Class 1 ('MI')		
Parameter	Est	(SE)	Probability	Est	(SE)	Probability
			1-Specificity			Sensitivity
e_{1c} [Q-wave]	−17.58	(953.49)	0.00	1.19	(0.42)	0.77
e_{2c} [History]	−1.42	(0.39)	0.30	1.33	(0.39)	0.79
e_{3c} [LDH]	−3.59	(1.01)	0.03	1.57	(0.47)	0.83
e_{4c} [CPK]	−1.41	(0.41)	0.20	16.86	(706.04)	1.00
			1-Prevalence			Prevalence
ϱ_0 [Cons]	0.17	0.22	0.54	–	–	0.46

Source: Skrondal and Rabe-Hesketh (2004).[111]

5 Structural equation models (SEM) with latent variables

The geneticist Wright[88] introduced path analysis for observed variables and Jöreskog[89–91] combined such models with common factor models to form structural equation models with latent variables.

There are several ways of parameterizing SEMs with latent variables. The most common is the LISREL parametrization[89–91] but we will use a parametrization suggested by Muthén[92] here because it turns out to be more convenient for the subsequent application.

The measurement part of the model is the confirmatory factor model (4) specified in Section 2.2. The structural part of the model specifies regressions for the latent variables on other latent and observed variables

$$\boldsymbol{\eta}_j = \boldsymbol{\alpha} + \mathrm{B}\boldsymbol{\eta}_j + \boldsymbol{\Gamma}\boldsymbol{x}_j + \boldsymbol{\zeta}_j \qquad (6)$$

Here $\boldsymbol{\eta}_j$ is a vector of latent variables with corresponding lower-triangular parameter matrix $\mathbf{B}$ governing the relationships among them, $\boldsymbol{\alpha}$ is a vector of intercepts, $\boldsymbol{\Gamma}$ a regression parameter matrix for the regression of the latent variables on the vector of observed covariates $\boldsymbol{x}_j$, and $\boldsymbol{\zeta}_j$ is a vector of disturbances. We define $\boldsymbol{\Psi} \equiv \mathrm{Cov}(\boldsymbol{\zeta}_j)$ and assume that $\mathrm{E}(\boldsymbol{\zeta}_j)=0$, $\mathrm{Cov}(\boldsymbol{x}_j, \boldsymbol{\zeta}_j) = 0$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_j, \boldsymbol{\zeta}_j) = 0$.

Substituting from the structural model into the measurement model, we obtain the reduced form,

$$\boldsymbol{y}_j = \boldsymbol{\beta} + \boldsymbol{\Lambda}(\mathrm{I} - \mathrm{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{x}_j + \boldsymbol{\zeta}_j) + \boldsymbol{\epsilon}_j \qquad (7)$$

The mean structure for $\boldsymbol{y}_j$, conditional on the covariates $\boldsymbol{x}_j$, becomes

$$\boldsymbol{\mu}_j \equiv \mathrm{E}(\boldsymbol{y}_j|\boldsymbol{x}_j) = \boldsymbol{\beta} + \boldsymbol{\Lambda}(\mathrm{I} - \mathrm{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{x}_j),$$

and the covariance structure of $\boldsymbol{y}_j$, conditional on the covariates, becomes

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\boldsymbol{y}_j|\boldsymbol{x}_j) = \boldsymbol{\Lambda}(\mathrm{I} - \mathrm{B})^{-1}\boldsymbol{\Psi}[(\mathrm{I} - \mathrm{B})^{-1}]'\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \qquad (8)$$

An important special case is the Multiple-Indicator-MultIple-Cause (MIMIC) model[93] which imposes the restriction $\mathbf{B} = \mathbf{I}$ in the structural model (6),

$$\boldsymbol{\eta}_j = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{x}_j + \boldsymbol{\zeta}_j$$

As for common factor models, the marginal likelihood of a SEM can be expressed in closed form if we assume that $\boldsymbol{\zeta}_j$ and $\boldsymbol{\epsilon}_j$ and hence $\boldsymbol{y}_j$, are multivariate normal. Instead of maximizing the marginal likelihood, we can minimize a fitting function similar to (3), with respect to the unknown free parameters of the SEM.

For further reading on structural equation modelling we recommend the books by Loehlin,[94] Bollen[95] and Kaplan.[96]

5.1 Example: Clients' satisfaction with counsellors' interviews

Alwin and Tessler[97] and Tessler[98] described and analysed data from an experiment to investigate the determinants of clients' satisfaction with counsellors' initial interviews.

Three experimental factors were manipulated: 1) 'Experience (E)' x_{1j}, clients' information regarding the length of time the counsellor has acted in his professional capacity (no experience versus full-fledged counsellor), 2) 'Value similarity (VS)' x_{2j}, the degree to which the client perceives the counsellor as similar in values and life-style preferences (sharply different philosophy of life versus high communality), and 3) 'Formality (F)' x_{3j}, the extent to which the counsellor exercises the maximum level of social distance permitted by norms governing a counselling relationship (informal versus formal). Ninety-six female subjects were randomly assigned to the two levels of each of the experimental factors in a full factorial design. All subjects were exposed to the same male counsellor.

It was important to assess the degree to which the experimental manipulations had been accurately perceived by the clients. In the measurement part of model, three latent variables, each corresponding to an experimental factor, were thus considered as 'manipulation checks': 'Perceived experience (P-E)' η_{1j}, measured by the items y_{1j}, y_{2j} and y_{3j}; 'Perceived value similarity (P-VS)' η_{2j}, measured by the items y_{4j}, y_{5j} and y_{6j}; and 'Perceived formality (P-F)' η_{3j}, measured by the items y_{7j}, y_{8j} and y_{9j}. An independent clusters confirmatory factor model (akin to that shown in Figure 1 but with three latent variables) was thus specified for the items, only allowing the items to measure the latent variables they were supposed to measure.

Clients' satisfaction was construed as a two-dimensional latent variable:

- 'Relationship-centered satisfaction (RS)' η_{4j}, representing the client's sense of closeness to the counsellor, measured by four items each scored from 0 to 6:
  - 'Personalism' $y_{10,j}$: A likert item (from 'agree strongly' to 'disagree strongly') with question wording 'I think that the counsellor is one of the nicest persons I've ever met'
  - 'Warmth' $y_{11,j}$: A semantic differential format (from 'cold' to 'warm')
  - 'Friendliness' $y_{12,j}$: A semantic differential format (from 'friendly' to 'unfriendly')
  - 'Concern' $y_{13,j}$: A semantic differential format (from 'unconcerned' to 'concerned')

- 'Problem-centered satisfaction (PS)' η_{5j}, representing the client's perception of the counsellor's ability to help, measured by four items each scored from 0 to 6:
  - 'Thoroughness' $y_{14,j}$: A Likert item (from 'agree strongly' to 'disagree strongly') with question wording "The counsellor was very thorough. I was left with the feeling that nothing important had been overlooked"
  - 'Skillfulness' $y_{15,j}$: A semantic differential format (from 'unskilled' to 'skilled')
  - 'Impressiveness' $y_{16,j}$: A semantic differential format (from 'impressive' to 'unimpressive')
  - 'Success in bringing clarity to the problem' $y_{17,j}$: A Likert item (from 'agree strongly' to 'disagree strongly') with question wording 'I felt that the nature of my problem had been clarified, that is that the counsellor had helped me to understand exactly what was troubling me'

Assuming an independent clusters confirmatory factor model for client's satisfaction, the factor loading matrix for all five common factors becomes

$$\boldsymbol{\Lambda}' = \begin{bmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{42} & \lambda_{52} & \lambda_{62} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{73} & \lambda_{83} & \lambda_{93} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{10,4} & \lambda_{11,4} & \lambda_{12,4} & \lambda_{13,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{14,5} & \lambda_{15,5} & \lambda_{16,5} & \lambda_{17,5} \end{bmatrix}$$

and the covariance matrix $\boldsymbol{\Theta}$ of the unique factors is diagonal.

In the structural part of model, Alwin and Tessler specified the following configuration for the parameter matrices:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & 0 & 0 \\ 0 & \gamma_{22} & 0 \\ 0 & 0 & \gamma_{33} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ b_{41} & b_{42} & b_{43} & 0 & 0 \\ b_{51} & b_{52} & b_{53} & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 \\ 0 & 0 & \psi_{33} & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & \psi_{45} \\ 0 & 0 & 0 & \psi_{54} & \psi_{55} \end{bmatrix}$$

This specification encodes a number of hypotheses regarding the investigated processes. The $\boldsymbol{\Gamma}$ matrix prescribes that a given experimental factor affects only the corresponding perception and not the perception of other experimental factors. Moreover, the experimental factors are not permitted to have direct effects on the the satisfaction constructs. The $\mathbf{B}$ matrix prescribes that there are no relations among the perception factors but that all three perception factors are allowed to affect both satisfaction factors. Finally, the $\boldsymbol{\Psi}$ matrix prescribes that the disturbances for the perception factors are uncorrelated but that the disturbances for the satisfaction factors may be correlated. A path diagram for the structural part of the SEM is shown in Figure 6.

Maximum likelihood estimates for the structural part of the model are presented in Table 7 (the estimates for the measurement part are omitted). Interestingly, the estimates suggest that the major determinant of relationship-centered satisfaction (RS) is the formality of the interview situation, whereas similarity with the counsellor has only a moderate effect and experience of the counsellor has a negligible effect. In contrast, the experience of the counsellor appears to be the major determinant of problem-centered satisfaction (PS) whereas similarity has only a moderate effect and formality a negligible

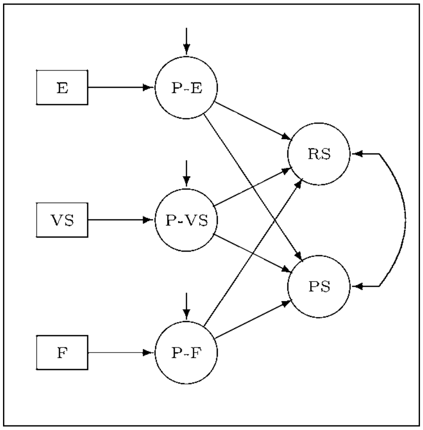

Figure 6 Clients' satisfaction with counsellors' interviews – Path diagram for structural part of model (measurement part omitted).

effect. The estimates thus indicate that there are two distinct kinds of client centered satisfaction in initial interviews and that they depend on the features of the interview situation.

The maximized log-likelihood is −1674.09 and the deviance is 285.22 with 160 degrees of freedom, suggesting rejection of the model. However, the CFI is 0.94, the TLI is 0.93 and the RMSEA is 0.09, indicating reasonable fit. The validity of the restrictions imposed by Alwin and Tessler can be investigated by assessing the improvement in fit obtained by relaxing them.

5.2 Some applications of SEM in medical research

An important application of SEM is for covariate measurement error.[58,99,100] Unfortunately, covariate measurement error models are usually not recognized as SEM and a restrictive classical measurement model is implicitly assumed.

More complex SEM with paths between several latent variables are commonly used in areas such as psychiatry,[101] addiction[102] and social medicine[103] and sometimes in public

Table 7 Clients' satisfaction with counsellors' interviews – Maximum likelihood estimates for structural part of SEM. Estimated regression parameters with standard errors in left panel and estimated residual (co)variances with standard errors in right panel

Path	Parameter	Est	(SE)	Parameter	Est	(SE)
E → P-E	γ_{11}	0.98	(0.02)	ψ_{11}	0.03	(0.01)
VS → P-VS	γ_{22}	0.99	(0.01)	ψ_{22}	0.02	(0.01)
F → P-F	γ_{33}	0.98	(0.02)	ψ_{33}	0.01	(0.00)
P-E → RS	b_{41}	0.00	(0.07)	ψ_{44}	0.39	(0.11)
P-VS → RS	b_{42}	0.11	(0.07)	ψ_{55}	0.19	(0.08)
P-F → RS	b_{43}	−0.19	(0.07)	ψ_{54}	0.13	(0.05)
P-E → PS	b_{51}	0.31	(0.08)			
P-VS → PS	b_{52}	0.06	(0.06)			
P-F → PS	b_{53}	0.01	(0.06)			

health and epidemiology.[104] Structural equation modelling is also a standard tool in biometrical genetics.[5] Here, common factors represent additive and dominant genetic and shared environment effects on observed characteristics or phenotypes and produce the covariance structure predicated by Mendelian genetics. The models become more complex when the phenotypes are latent.[105] Latent growth curve models for multivariate longitudinal data are used in areas such as child development[106] and ageing.[107]

6 Concluding remarks

We have reviewed classical latent variable models and demonstrated how they can used to address research problems in medicine.

There has recently been considerable work on unifying and extending the classical models within a general framework.[108–112] We have not discussed generalizations such as SEM for categorical and mixed responses,[92,113] multilevel latent variable models,[110,114] models with nonlinear functions among latent variables,[115] latent class structural equation models[116,117] and models including both continuous and discrete latent variables.[109] Such complex models are useful for estimating complier average causal effects in clinical trials,[118] joint modelling of longitudinal data and dropout or survival,[10] multiprocess survival models,[119,120] and many other problems. Skrondal and Rabe-Hesketh[27] provide a recent survey of advanced latent variable modelling.

Acknowledgements

We wish to thank The Research Council of Norway for a grant supporting our collaboration. We also thank the UK Data Archive for making the Health Survey for England, 1996 and 2004 data available. The original data creators, depositors, copyright holders, or funders of the Data Collections and the UK Data Archive bear no responsibility for our analysis or interpretation.

We appreciate the information provided by Richard C. Tessler regarding our SEM example.

References

1 Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**: 963–74.

2 Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine* 1988; **7**: 1121–37.

3 DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**: 1777–88.

4 Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**: 671–81.

5 Neale MC, Cardon LR. *Methodology for Genetic Studies of Twins and Families*. Kluwer, 1992.

6 Fayers PM, Hand DJ. Causal variables, indicator variables, and measurement scales.

Journal of the Royal Statistical Society, Series A 2002; **165**: 233–61.

7 Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5**: 21–7.

8 Chao A, Tsay PK, Lin SH, Shau WY, Chao DU. Tutorial in biostatistics: The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* 2001; **20**: 3123–57.

9 Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology* 1990; **132**: 734–45.

10 Hogan JW, Laird NM. Model-based approaches to analyzing incomplete repeated measures and failure time data. *Statistics in Medicine* 1997; **16**: 259–71.

11 Spearman C. General intelligence, objectively determined and measured. *American Journal of Psychology* 1904; **15**: 201–93.

12 Thurstone LL. *The vectors of mind*. University of Chicago Press, 1935.

13 Thomson GH. *The factorial analysis of human ability*. University of London Press, 1938.

14 Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. Springer, 2000.

15 Rabe-Hesketh S, Skrondal A. Generalized linear mixed effects models. In Fitzmaurice G, Davidian M, Molenberghs G, Verbeke G eds. *Advances in longitudinal data analysis: a handbook of modern statistical methods*. Chapman & Hall/CRC, 2007.

16 Gulliksen H. *Theory of mental tests*. Wiley, 1950.

17 Jöreskog KG. Statistical analysis of sets of congeneric tests. *Psychometrika* 1971; **36**: 109–33.

18 Dunn G, Sham PC, Hand DJ. Statistics and the nature of depression. *Journal of the Royal Statistical Society, Series A* 1993; **156**: 63–87.

19 Jöreskog KG. Some contributions to maximum likelihood factor analysis. *Psychometrika* 1967; **32**: 443–82.

20 Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–92.

21 Tanaka JS. Multifaceted conceptions of fit in structural equation models. In Bollen KA, Long JS eds. *Testing structural equation models*. Sage, 1993; 10–39.

22 Browne MW, Cudeck R. Alternative ways of assessing model fit. In Bollen KA, Long JS eds. *Testing structural equation models*. Sage, 1993; 136–62.

23 Bentler PM. Comparative fit indices in structural models. *Psychological Bulletin* 1990; **107**: 238–46.

24 Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 1973; **38**: 1–10.

25 National Centre for Social Research and University College London, Department of Epidemiology and Public Health. *Health Survey for England, 2004 [computer file]*. UK Data Archive [Distributor], 2006.

26 Nisenbaum R, Reyes M, Mawle AC, Reeves WC. Factor analysis of unexplained severe fatigue and interrelated symptoms: overlap with criteria for chronic fatigue syndrome. *American Journal of Epidemiology* 1998; **148**: 72–7.

27 Skrondal A, Rabe-Hesketh S. Latent variable modelling: a survey. *Scandinavian Journal of Statistics*, in press.

28 McDonald RP. *Factor analysis and related methods*. Erlbaum, 1985.

29 Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 1969; **34**: 183–202.

30 Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics* 1994; **50**: 1171–7.

31 Dominicus A, Skrondal A, Gjessing HK, Pedersen N, Palmgren J. Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics* 2006; **36**: 331–40.

32 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford University Press, 2003.

33 Dunn G. Design and analysis of reliability studies. *Statistical Methods in Medical Research* 1992; **1**: 123–57.

34 Dunn G, Roberts C. Modelling method comparison data. *Statistical Methods in Medical Research* 1999; **8**: 161–79.

35 Dunn G. *Statistics in psychiatry*. Arnold, 2000.

36 Dunn G. *Statistical evaluation of measurement errors: design and analysis of reliability studies, Second edition*. Arnold, 2004.

37 McDonald RP. *Test theory: a unified treatment*. Erlbaum, 1999.

38 Turner SW, Toone BK, Brett-Jones JR. Computerized tomographic scan changes in early schizophrenia. *Psychological Medicine* 1986; **16**: 219–5.

39 Rabe-Hesketh S, Skrondal A. Parameterization of multivariate random effects models for categorical data. *Biometrics* 2001; **57**: 1256–64.

40 Jasper HH. The measurement of depression–elation and relation to a measure of extraversion–introversion. *Journal of Abnormal Social Psychology* 1930; **25**: 307–18.

41 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–10.

42 McKay D, Danyko S, Neziroglu F, Yaryuratobias JA. Factor structure of the Yale-Brown obsessive-compulsive scale – A 2-dimensional measure. *Behaviour Research and Therapy* 1995; **33**: 865–69.

43 Jacob KS, Everitt BS, Patel V, Weich S, Araya R, Lewis GH. The comparison of latent variable models of nonpsychotic psychiatric morbidity in four culturally diverse populations. *Psychological Medicine* 1998; **28**: 145–52.

44 Lord FM. *Applications of item response theory to practical testing problems*. Erlbaum, 1980.

45 Lord FM. *A theory of test scores*. Psychometric Monograph 7, Psychometric Society, 1952.

46 Lord FM, Novick MR. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

47 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut, 1960.

48 Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrica* 1948; **16**: 1–32.

49 Andersen EB. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* 1970; **32**: 283–301.

50 Andersen EB. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 42–54.

51 Breslow NE, Day N. *Statistical methods in cancer research*. Volume I – *The analysis of case-control studies*. IARC, 1980.

52 Joint Health Surveys Unit of Social and Community Planning Research and University College London. *Health Survey for England, 1996 [computer file] (Third Edition)*. Colchester, Essex: UK Data Archive [Distributor], 2001.

53 Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health survey. Manual and interpretation guide*. The Health Institute, New England Medical Center, 1993.

54 Laird NM. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* 1978; **73**: 805–11.

55 Lindsay BG. *Mixture models: theory, geometry and applications*, volume 5 of *NSF-CBMS regional conference series in probability and statistics*. Institute of Mathematical Statistics, 1995.

56 Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 1999; **55**: 117–28.

57 Böhning D. *Computer-assisted analysis of mixtures and applications. Meta-Analysis, disease mapping and others*. Chapman & Hall, 2000.

58 Rabe-Hesketh S, Pickles A, Skrondal A. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling* 2003; **3**: 215–32.

59 Wilson M. *Constructing measures. An item response modeling approach*. Erlbaum, 2005.

60 Bock RD, Lieberman M. Fitting a response model for n dichotomously scored items. *Psychometrika* 1970; **33**: 179–97.

61 Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* 2005; **128**: 301–23.

62 Wright BD. Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 1977; **14**: 97–116.

63 Lawley DN. On problems connected with item selection and test construction. In *Proceedings of the Royal Society of Edinburgh*, Volume 61. 1943; 273–87.

64 Samejima F. *Estimation of latent trait ability using a response pattern of graded scores*. Psychometric Monograph 17, Psychometric Society, 1969.

65 Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**: 561–73.

66 Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Sage, 1991.

67 Embretson SE, Reise SP. *Item response theory for psychologists*. Erlbaum, 2000.

68 Bech P. Rating scales for affective disorders. Their validity and consistency. *Acta Psychiatrica Scandinavica* 1981; **64**: 1–101.

69 Bech P, Allerup P, Gram IF, Reisby N, Rosenberg R, Jacobsen O, Nagy A. The Hamilton depression scale. Evaluation of objectivity using logistic models. *Acta Psychiatrica Scandinavica* 1981; **63**: 290–9.

70 Licht RW, Qvitzau S, Allerup P, Bech P. Validation of the Bech-Rafaelsen melancholia scale and the Hamilton depression scale in patients with major depression; is the total score a valid measure of illness severity? *Acta Psychiatrica Scandinavica* 2005; **111**: 144–9.

71 Revicki DA, Cella DF. Health status assessment for the twenty-first century. Item response theory, item banking and computer adaptive testing. *Quality of Life Research* 1997; **6**: 595–600.

72 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Medical Care* 2000; **38**: 28–42.

73 Bartolucci F, Forcina A. Analysis of capture-recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. *Biometrics* 2001; **57**: 714–9.

74 Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. In Stouffer SA, Guttmann L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA eds. *Studies in social psychology in world war II*, Volume 4, measurement and prediction. Princeton University Press, 1950; 362–412.

75 Goodman LA. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – A modified latent structure approach. *American Journal of Sociology* 1974; **79**: 1179–1259.

76 McCutcheon AL. *Latent class analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage, 1987.

77 Clogg CC. Latent class models. In Arminger G, Clogg CC, Sobel ME eds., *Handbook of Statistical Modelling for the Social and Behavioral Sciences*. Plenum Press, 1995; 311–59.

78 Formann AK, Kohlmann T. Latent class analysis in medical research. *Statistical Methods in Medical Research* 1996; **5**: 179–211.

79 Agresti A. *Categorical data analysis*, Second Edition. Wiley, 2002.

80 Clogg CC, Sawyer DO. A comparison of alternative models for analyzing the scalability of response patterns. In: Leinhardt S ed., *Sociological Methodology* 1981. Jossey-Bass, 1981: 240–80.

81 Pedersen W, Skrondal A. Alcohol and sexual victimization: A longitudinal study of Norwegian girls. *Addiction* 1996; **91**: 565–81.

82 Collins LM, Wugalter SE. Latent class models for stage-sequential dynamic latent-variables. *Multivariate Behavioral Research* 1992; **27**: 131–57.

83 Langeheine R. Latent variable Markov models. In von Eye A, Clogg CC eds., *Latent variable analysis. Applications to developmental research*. Sage, 1994; 373–98.

84 Vermunt JK, Langeheine R, Böckenholt U. Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics* 1999; **24**: 179–207.

85 Wedel M, DeSarbo W. Mixture regression models. In *Applied latent class analysis*. Cambridge University Press, 2002; 366–82.

86 Vermunt JK. An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica* 2004; **58**: 220–33.

87 Muthén BO, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang CP, Kellam SG, Carlin JB, Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics* 2002; **3**: 459–75.

88 Wright S. On the nature of size factors. *Genetics* 1918; **3**: 367–74.

89 Jöreskog KG. A general method for estimating a linear structural equation system. In Goldberger AS, Duncan OD eds., *Structural equation models in the social sciences*. Seminar, 1973; 85–112.

90 Jöreskog KG. Analysis of covariance structures. In Krishnaiah PR ed. *Multivariate analysis*, Volume III. Academic, 1973; 263–85.

91 Jöreskog KG. Structural equation models in the social sciences. Specification, estimation and testing. In Krishnaiah PR ed. *Applications of Statistics*. North-Holland, 1977; 265–87.

92 Muthén BO. A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika* 1984; **49**: 115–32.

93 Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 1975; **70**: 631–39.

94 Loehlin JC. Latent variable models. *An introduction to factor, path, and structural equation analysis*. Erlbaum, 2003.

95 Bollen KA. *Structural equations with latent variables*. Wiley, 1989.

96 Kaplan D. *Structural equation modeling. Foundations and extensions*. Sage, 2000.

97 Alwin DF, Tessler RC. Causal models, unobserved variables, and experimental data. *American Journal of Sociology* 1974; **80**: 58–86.

98 Tessler RC. Clients' reactions to initial interviews. Determinants of relationship-centered and problem-centered satisfaction. *Journal of Counseling Psychology* 1975; **22**: 187–191.

99 Plummer M, Clayton DG. Measurement error in dietary assessment. An investigation using covariance structure models. Part I. *Statistics in Medicine* 1993; **12**: 925–35.

100 Plummer M, Clayton DG. Measurement error in dietary assessment. An investigation using covariance structure models. Part II. *Statistics in Medicine* 1993; **12**: 937–48.

101 Drake RJ, Pickles A, Bentall RP, Kinderman P, Haddock G, Tarrier N, Lewis SW. The evolution of insight, paranoia and depression during early schizophrenia. *Psychological Medicine* 2004; **34**: 285–92.

102 Newcombe MD, Bentler PM. *Consequences of adolescent drug use: Impact on the lives of young adults*. Sage, 1988.

103 Johnson RJ, Wolinsky FD. The structure of health status among older adults: disease, disability, functional limitation, and perceived health. *Journal of Health and Social Behavior* 1993; **34**: 105–21.

104 De Stavola BLD, Nitsch D, dos Santos Silva I, McCormack V, Hardy R, Mann V, Cole TJ, Morton S, Leon DA. Statistical issues in life course epidemiology. *American Journal of Epidemiology* 2006; **163**: 84–96.

105 Simonoff E, Pickles A, Hervas A, Silberg JL, Rutter M, Eaves L. Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction. *Psychological Medicine* 1998; **28**: 825–37.

106 McArdle JJ, Epstein D. Latent growth curves within developmental structural equation models. *Child Development* 1987; **58**: 110–33.

107 McArdle JJ, Hamgami F, Jones K, Jolesz F, Kikinis R, Spiro A, Albert MS. Structural modeling of dynamic changes in memory and brain structure using longitudinal data from the normative aging study. *Journals of Gerontology Series B – Psychological Sciences and Social Sciences* 2004; **59**: 294–304.

108 Bartholomew DJ, Knott M. *Latent variable models and factor analysis*. Arnold, 1999.

109 Muthén BO. Beyond SEM: general latent variable modeling. *Behaviormetrika* 2002; **29**: 81–117.

110 Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. *Psychometrika* 2004; **69**: 167–90.

111 Skrondal A, Rabe-Hesketh S. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, 2004.

112 Rabe-Hesketh S, Skrondal A. Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika* 2007; **72**: 123–140.

113 Skrondal A, Rabe-Hesketh S. Multilevel logistic regression for polytomous data and rankings. *Psychometrika* 2003; **68**: 267–87.

114 Vermunt JK. Multilevel latent class models. In Stolzenberg RM ed. *Sociological Methodology* 2003, Volume 33. Blackwell, 2003; 213–39.

115 Arminger G, Muthén BO. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* 1998; **63**: 271–300.

116 Dayton CM, MacReady GB. Concomitant variable latent class models. *Journal of the American Statistical Association* 1988; **83**: 173–78.

117 Hagenaars JAP. *Loglinear Models with Latent Variables*. Sage University Paper

Series on Quantitative Applications in the Social Sciences. Sage, 1993.

118 Jo B. Model misspecification sensitivity in estimating causal effects of interventions with noncompliance. *Statistics in Medicine* 2002; **21**: 3161–81.

119 Lillard LA. Simultaneous-equations for hazards-marriage duration and fertility timing. *Journal of Econometrics* 1993; **56**: 189–217.

120 Steele F, Goldstein H, Browne W. A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling* 2004; **4**: 145–59.

121 Rabe-Hesketh S, Skrondal A, Pickles A. Gllamm manual. Technical Report 160, U.C. Berkeley Division of Biostatistics Working Paper Series, 2004. Downloadable from http://www.bepress.com/ucbbiostat/paper160/.

122 Rabe-Hesketh S, Skrondal A. *Multilevel and longitudinal modeling using stata*. Stata Press, 2005.

123 Muthén LK, Muthén BO. *Mplus user's guide* Third edition). Muthén & Muthén, 2004.

124 Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. *BUGS 0.5 Examples, Volume 1*. MRC-Biostatistics Unit, 1996.

125 Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. *BUGS 0.5 Examples, Volume 2*. MRC-Biostatistics Unit, 1996.

126 Congdon P. *Bayesian statistical modelling*, Second Edition. Wiley, 2006.

127 Hardouin JB. Rasch analysis: Estimation and tests with raschtest. *The Stata Journal* 2007; 7: 22–44.

128 Fox J. Structural equation modeling with the sem package in R. *Structural Equation Modeling* 2006; **13**: 465–86.

129 Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software* 2006; **17**: 1–25.

130 Waller NG. LCA 1.1: an R package for exploratory latent class analysis. *Applied Psychological Measurement* 2004; **28**: 141–2.

131 Hatcher L. *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. SAS Press, 1994.

132 Wolfinger RD. Fitting non-linear mixed models with the new NLMIXED procedure. Technical report, SAS Institute, Cary, NC, 1999.

133 Vermunt JK, Magidson J. *Technical guide for latent GOLD 4.0: basic and advanced*. Statistical Innovations, 2005.

134 Jöreskog KG, Sörbom D, Du Toit SHC, Du Toit M. *LISREL 8: new statistical features*. Lincolnwood, IL: Scientific International, 2001.

135 Bentler PM. *EQS structural equation program manual*. Multivariate Software, 1995.

136 Neale MC, Boker SM, Xie G, Maes HH. *Mx: statistical modeling*, sixth edition. Virginia Commonwealth University, Department of Psychiatry, 2002. Downloadable from `http://www.vipbg.vcu.edu/mxgui/`.

137 Arbuckle JL, Wothke W. *Amos 5.0 update to the Amos user's guide*. SmallWaters Corporation, 2003.

138 Du Toit M ed. *IRT from SSI*. Scientific Software International, 2003.

139 Linacre JM ed. *A User's Guide to the WINSTEPS and MIMISTEP Masch-model computer programs*. winsteps.com, 2006.

140 Wu ML, Adams RJ, Wilson MR. *ConQuest [Computer software and manual]*. Australian Council for Educational Research, 1998.

Appendix

Some software for classical latent variable modelling

The examples in this paper were estimated using gllamm[121,122] (Tables 3–6) and Mplus[123] (Tables 1,2,7).

Table A1 lists some software packages and indicates whether each can be used to fit CFA, IRT models, latent class (LC) models and structural equation models (SEM). WinBUGS[124,125] can be used for Markov chain Monte Carlo estimation of all these models,[126] but requires expertise for setting up the models and monitoring convergence.

Table A1 Some software for classical latent variable modelling

	Model types			
Software	CFA	IRT	LC	SEM
Stata	gllamm[121]	gllamm raschtest[127]•	gllamm	gllamm
R	sem[128]⋆	ltm[129]	lca[130]	sem⋆
SAS	CALIS[131]⋆	NLIMXED[132]	TRAJ	CALIS⋆
Mplus[123]	√	√	√	√
LatentGOLD[133]	◇	√	√	
LISREL[134]	√	†		√
EQS[135]	√	†		√
Mx[136]	√	†		√
Amos[137]	√			√
BILOG-MG[138]		•		
WINSTEPS[139]		•		
ConQuest[140]		•		

√, Model type is accommodated; †, Only normal-ogive (probit link) models; ⋆, No general missing data patterns; •, Only one-parameter models; ◇, Only uncorrelated factors.