Citation:

Brown, A. & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (a volume in the Multivariate Applications Series)*. New York: Routledge/Taylor & Francis Group.

15 Scoring and estimating score precision using IRT

Anna Brown and Tim J. Croudace

Chapter Overview

The ultimate goal of measurement is to produce a score by which individuals can be assessed and differentiated. Item response theory (IRT) modeling views responses to test items as indicators of a respondent's standing on some underlying psychological attributes (van der Linden & Hambleton, 1997) – we often call them latent traits – and devises special algorithms for estimating this standing. This chapter gives an overview of methods for estimating person attribute scores using one-dimensional and multi-dimensional IRT models, focusing on those that are particularly useful with patient-reported outcome (PRO) measures.

To be useful in applications, a test score has to approximate the latent trait well, and importantly, the precision level must be known in order to produce information for decision-making purposes. Unlike classical test theory (CTT), which assumes the precision with which a test measures the same for all trait levels, IRT methods assess the precision with which a test measures at different trait levels. In the context of patient-reported outcomes measurement, this enables assessment of the measurement precision for an individual patient. Knowing error bands around the patient's score is important for informing clinical judgments, such as deciding upon significance of any change, for instance in response to treatment etc. (Reise & Haviland, 2005). At the same time, summary indices are often needed to summarize the overall precision of measurement in a research sample, population group, or in the population as a whole. Much of this chapter is devoted to methods for estimating measurement precision, including the score-dependent standard error of measurement and appropriate sample-level or population-level marginal reliability coefficients.

Patient-reported outcome measures often capture several related constructs, the

feature that may make the use of multi-dimensional IRT models appropriate and beneficial (Gibbons, Immekus & Bock, 2007). Several such models are described, including a model with multiple correlated constructs, a model where multiple constructs are underlain by a general common factor (second-order model), and a model where each item is influenced by one general and one group factor (bifactor model). To make the use of these models more easily accessible for applied researchers, we provide specialized formulae for computing test information, standard errors and reliability. We show how to translate a multitude of numbers and graphs conditioned on several dimensions into easy-to-use indices that can be understood by applied researchers and test users alike. All described methods and techniques are illustrated with a single data analysis example involving a popular PRO measure, the 28-item version of the General Health Questionnaire (GHQ28; Goldberg & Williams, 1988), completed in mid-life by a large community sample as a part of a major UK cohort study.

Common Factor Model and Item Responses

Psychometric tools measuring patient-reported outcomes or health related quality of life (HRQoL) often necessitate the capture of several related constructs (Fayers & Machin, 2007). For example, the Quality of Life Interview for the Chronically Mentally Ill differentiates seven sub-domains (Gibbons et al., 2007), each a distinct area of functioning. Hence, we will use multidimensional models as the basis (Gibbons, Immekus & Bock, 2007), and consider a one-dimensional model as the simplest special case.

Let $\boldsymbol{\theta} = \left(\theta_1, \theta_2, ..., \theta_T\right)'$ be a set of *T* psychological attributes (traits). We assume that the latent traits are normally distributed random variables; they have the mean of zero, unit variance and are freely correlated (their covariance matrix is $\boldsymbol{\Sigma}$). In a test with *m* items, each item is designed to elicit an internal response (denoted u_i^*), which depends on one or more attributes, as described by a common factor model:

$$u_i^* = \mu_i + \lambda_{1i}\theta_1 + \lambda_{2i}\theta_2 + ... + \lambda_{Ti}\theta_T + \varepsilon_i \,. \qquad (15.1)$$

In this expression, μ_i is the mean utility for the population of respondents; $\lambda_{1i}, \lambda_{2i}, ..., \lambda_{Ti}$ are factor loadings of item *i* on each of the attributes, and ε_i is the unique part (or error). The error terms are normally distributed random variables with the mean zero and are

uncorrelated with the common factors and among themselves, so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal.

The internal response u_i^* can represent, for example, the degree of agreement with the item or the value the respondent attaches to the item (item *utility*); however, this *response tendency* is not observed directly. The actual observed item response u_i will typically constitute an option that the respondent chose from available options, either dichotomous ("yes"-"no", "agree"-"disagree") or polytomous ("never"-"sometimes"-"often"-"always"; or "strongly disagree" – "disagree" – "neutral" – "agree" – "strongly agree" etc.). In IRT models, the observed response u_i is related to the latent response tendency u_i^* through a threshold process. For instance, it is assumed that the response is positive when the response tendency is over a threshold value τ_i, and the response is negative otherwise:

$$u_i = \begin{cases} 1 & \text{if } u_i^* \geq \tau_i \\ 0 & \text{if } u_i^* < \tau_i \end{cases} . \tag{15.2}$$

In what follows we give formulae for the dichotomous case, mentioning how to extend them to polytomous cases in passing. All response models can be extended to the polytomous case by considering, for example, a graded response approach (Samejima, 1969), or from the Rasch modeling perspective, a partial credit model (Masters & Wright, 1997).

Given the item-factors relationship (15.1) and the threshold relationship (15.2), the item response function for item *i* in a *threshold / loading* form is given by

$$\mathrm{P}(u_i) = \Pr(u_i = 1 | \boldsymbol{\theta}) = \Phi\left(\frac{-\tau_i + \lambda_{1i}\theta_1 + \ldots + \lambda_{Ti}\theta_T}{\sqrt{\psi_i^2}} \right), \tag{15.3}$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function evaluated at *x*, and ψ_i^2 is the residual variance for item *i*. Without loss of generality, we use the normal-ogive link here. Alternatively, the logistic link function $\mathrm{L}(x) = 1/(1+e^{-x})$ can be used (Reckase, 2009).

Because the continuous response tendency u_i^* is unobserved (only its dichotomization u_i is observed), the variance ψ_i^2 cannot be identified in typical single-group models. It is customary to parameterize IRT models (e.g. McDonald, 1999) by letting

$$\alpha_i = \frac{-\tau_i}{\sqrt{\psi_i^2}} \quad \text{and} \quad \beta_{ij} = \frac{\lambda_{ij}}{\sqrt{\psi_i^2}} \ . \tag{15.4}$$

With this, the item response function (IRF) can be written in an *intercept / slope* form as

$$P_i = \mathrm{P}\left(u_i \middle| \theta\right) = \Phi\left(\alpha_i + \beta_{1i}\theta_1 + \ldots + \beta_{Ti}\theta_T\right). \tag{15.5}$$

The multidimensional response model is flexible in terms of how many attributes can influence a particular item. For instance, every item in a multi-trait test might measure one trait only and therefore will have only one non-zero slope (we say that items "possess an independent-clusters basis"; McDonald, 1999, page 179). When this is the case, the IRF is a familiar s-shaped item characteristic curve. The IRF becomes a surface when the item depends on two attributes. In these two cases of low dimensionality, the IRF can be easily plotted.

Latent Trait Estimation

Latent trait scores can be estimated by treating the model parameters (item intercepts and slopes, trait covariances, etc.) as if they were known. This is reasonable if model parameters have been accurately estimated. When the model parameters are fixed to their estimated values, the probability of every item response depends only on the latent traits, and the fundamental approach to estimating the trait scores is to search for values that maximize the joint likelihood of each response pattern $\mathbf{u} = \left(u_1, u_2, \ldots u_m\right)$. The maximum likelihood (ML) approach assumes local independence, that is, the item responses are independent after conditioning on the latent traits. The ML scores are found iteratively by maximizing the mode of the likelihood function

$$l\left(\mathbf{u} \middle| \boldsymbol{\theta}\right) = \prod_{i=1}^{m} P_i\left(\boldsymbol{\theta}\right)^{u_i} \left(1 - P_i\left(\boldsymbol{\theta}\right)\right)^{1-u_i}. \tag{15.6}$$

Maximum likelihood estimation uses information from the item responses only and is philosophically uncontroversial (McDonald, 2011). The score, however, is undefined for some response patterns, notably for "perfect" patterns (for instance by answering "yes" to **all** items or selecting the top rating category such as "strongly agree"). The likelihood of such a pattern increases infinitely when the latent trait score increases. The score is also undefined when non-keyed responses are given to all items (for instance by answering "no" or selecting the bottom rating category such as "strongly disagree"). Such extreme patterns can be quite

common in PRO measures, for example in screening tests.

An alternative approach is to use prior information about the trait distribution in addition to the actual item responses. With this Bayesian approach, the likelihood of a score given the observed response pattern (posterior distribution) is computed from known density of the score in the population (prior distribution, usually standard normal), and the likelihood of the observed response pattern given the score (joint likelihood value used in ML estimation). Two alternative methods using this approach are *expected a posteriori* (EAP) estimation, which computes the mean of the posterior distribution; and *maximum a posteriori* (MAP) estimation, which maximizes the mode of the posterior distribution (Embretson & Reise, 2000). By adding information from the prior distribution, the problem of undefined trait scores for "perfect" patterns is overcome, that is, a score is always defined when the posterior approach is used.

The EAP estimation is most often used when only one trait θ is measured. The EAP score is computed as the ratio between the integral of the posterior function weighted by the latent trait θ, and the unweighted integral of the posterior function. The EAP scores are approximated using numerical integration

$$\mathrm{EAP}\left(\mathbf{u}|\theta\right) \approx \frac{\sum_q \theta_q \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}{\sum_q \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}, \tag{15.7}$$

where q are quadrature points selected along the latent trait continuum. For a discussion regarding the number of points necessary for precise estimation see Thissen and Orlando (2001). It is evident that an EAP score is non-iterative and easy to compute when only one or two dimensions are involved. The EAP estimation, however, becomes computationally demanding when the number of traits increases, as the number of points on the multidimensional grid used for integration increases exponentially.

The MAP scores are found iteratively by maximizing the mode of the posterior distribution

$$l_P\left(\mathbf{u}|\boldsymbol{\theta}\right) = \phi\left(\boldsymbol{\theta}\right) \prod_{i=1}^{m} \left[P_i\left(\boldsymbol{\theta}\right)^{u_i} \left(1 - P_i\left(\boldsymbol{\theta}\right)\right)^{1-u_i} \right]. \tag{15.8}$$

Finding the best set of scores optimizing the multidimensional function above requires iterative procedures using gradients. For one-dimensional models this seems a complication

compared to the non-iterative EAP; however, in multidimensional models this estimation method is much more efficient because it searches for the optimal set of theta values for all traits simultaneously and the number of iterations is unaffected by the number of traits.

The posterior EAP and MAP approaches have been shown to achieve precise estimates with fewer items than the ML method; however, they are known to shrink the latent trait distribution towards the population mean. The amount of shrinkage depends on several factors, including the test length, and can be quite substantial in short tests (Thissen & Orlando, 2001).

Standard Error of Measurement and Information

Score estimates are only our best guess at estimating the true score, and inevitably, all estimation methods are associated with a certain degree of error. The likelihood of a response pattern **u**, or the posterior likelihood, is a function that is typically Gaussian in appearance and has a single peak. The mode (or the mean) provides a limited summary of the likelihood function. The width or the spread of the likelihood, on the other hand, indicates the degree of uncertainty around the estimated score – the narrower the spread, the more confident we can be that the true theta value is close to the estimated value. The spread of the likelihood for approximately Gaussian distributions is meaningfully described by its standard deviation. The standard deviation of the likelihood of a response pattern, therefore, is a measure of the **standard error** of estimation in IRT.

For the EAP estimator, it is natural to compute the standard deviation from the mean (which is the estimated EAP score) directly, by taking values of the likelihood function at the quadrature points:

$$\mathrm{SE}_{EAP}\left(\hat{\theta}\right) \approx \sqrt{\frac{\sum_q \left(\theta_q - \mathrm{EAP}[\theta]\right)^2 \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}{\sum_q \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}}. \qquad (15.9)$$

For the methods maximizing the mode of the likelihood function (ML and MAP), the variance of the score estimate is obtained as the inverse of the expected value of the second derivative of $l\left(\mathbf{u}|\boldsymbol{\theta}\right)$ or $l_P\left(\mathbf{u}|\boldsymbol{\theta}\right)$ with respect to $\boldsymbol{\theta}$ (Fisher, 1921). The expected value of the second derivative of likelihood function is called the *Fisher information*

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right), \tag{15.10}$$

and the variance of the likelihood function (the squared standard error of the score) is the reciprocal of the Fisher information

$$SE^2(\boldsymbol{\theta}) = \frac{1}{\mathcal{I}(\boldsymbol{\theta})}. \tag{15.11}$$

The Fisher information is inversely related to the error variance and therefore provides a measure of the test precision. Applying the general formula (15.11) to the ML estimator, we obtain for the one-dimensional case $SE_{ML}(\hat{\theta}) = 1/\sqrt{\mathcal{I}(\hat{\theta})}$. The standard error of the MAP estimator involves the information of the posterior likelihood function instead:

$$SE_{MAP}(\hat{\theta}) = \frac{1}{\sqrt{\mathcal{I}_P(\hat{\theta})}}. \tag{15.12}$$

For the multidimensional case, the standard error of the MAP-estimated score for the latent trait θ_a involves the directional posterior information evaluated at the point-estimates of the trait scores $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_T)$:

$$SE^a(\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{\mathcal{I}_P^a(\hat{\boldsymbol{\theta}})}}. \tag{15.13}$$

The following section shows how to compute the Fisher information in both one-dimensional and multidimensional cases.

Item Information Function (IIF)

When responses to test items depend on one latent trait, the maximum likelihood Item Information Function (IIF) for item *i* is given by

$$\mathcal{I}_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)\left[1 - P_i(\theta)\right]}, \tag{15.14}$$

where $P_i'(\theta)$ denotes the first derivative of the item response function (McDonald, 1999).

For items with graded response categories, the item information can be derived from category response functions as follows (Samejima, 1969; Dodd, DeAyala & Koch, 1995):

$$\mathcal{I}_i(\theta) = \sum_{x=0}^{k} \frac{\left[P_{ix}'(\theta)\right]^2}{P_{ix}(\theta)}. \qquad (15.15)$$

A challenging feature of many PRO measures is their potentially multi-dimensional structure, whereby item responses may depend on two or more latent traits. When this is the case, the *direction* of information must also be considered (Ackerman, 2005; Reckase, 2009). The definition of item information is generalized to accommodate the change in slope with direction $\boldsymbol{\alpha}$ taken from a point in the latent trait space

$$\mathcal{I}_i^{\alpha}(\boldsymbol{\theta}) = \frac{\left[\nabla_{\alpha} P_i(\boldsymbol{\theta})\right]^2}{P_i(\boldsymbol{\theta})\left[1 - P_i(\boldsymbol{\theta})\right]}, \qquad (15.16)$$

where $P_i(u_i = 1|\boldsymbol{\theta})$. Let $\boldsymbol{\alpha}$ be a vector of angles to all T axes that defines the direction from a point $\boldsymbol{\theta}$, and $\nabla_{\alpha} P_i(\boldsymbol{\theta})$ be the gradient in direction $\boldsymbol{\alpha}$, which is given by (Reckase, 2009):

$$\nabla_{\boldsymbol{\alpha}} P_i(\boldsymbol{\theta}) = \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_1}\cos(\alpha_1) + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_2}\cos(\alpha_2) + \ldots + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_T}\cos(\alpha_T). \qquad (15.17)$$

For each item, we might consider one or more directions of information. Often, these would correspond to the direction(s) of the trait(s) in which we are interested. When computing the information along the axis θ_a, the angle to θ_a is 0° and therefore $\cos(\alpha_a) = 1$, whereas the angle to any other axis, say θ_b, is determined by the correlation between θ_a and θ_b so that $\cos(\alpha_b) = \text{corr}(\theta_a, \theta_b)$ (Bock, 1975). It follows that the information in direction of the trait θ_a is given by:

$$\nabla_a P_i(\boldsymbol{\theta}) = \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_1}\text{corr}(\theta_1, \theta_a) + \ldots + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_a} + \ldots + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_T}\text{corr}(\theta_T, \theta_a). \qquad (15.18)$$

It follows that in instruments measuring uncorrelated traits, only the partial derivative in the direction of the trait itself contributes to the information. However, for instruments involving correlated traits, partial derivatives in directions of other traits might contribute, adding information proportionately to the strengths of their relationships with the focus trait. In applications sections of this chapter we will provide specific information functions for

models commonly used to score responses to PRO measures.

Test Information Function (TIF)

When local independence holds, the Test Information Function (TIF) is simply the sum of the item information functions. In one-dimensional cases, the TIF is $\mathcal{I}(\theta) = \sum_{i=1}^{m} \mathcal{I}_i(\theta)$. In multi-dimensional cases, the total information in direction $\boldsymbol{\alpha}$ is the sum of all item information functions (15.16). Thus, the test information about trait θ_a is:

$$\mathcal{I}^a(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathcal{I}_i^a(\boldsymbol{\theta}). \tag{15.19}$$

So far, we have taken into account information provided by the item responses, i.e. the maximum likelihood (ML) information. When Bayesian estimators are used, information given by the prior distribution will also contribute to the latent trait estimation. The so-called *posterior test information* function, $\mathcal{I}_P(\boldsymbol{\theta})$, reflects contributions provided by the item responses and the prior distribution. For a single latent trait, the standard normal prior distribution adds unity to the information across the latent trait continuum (Du Toit, 2003),

$$\mathcal{I}_P(\theta) = \mathcal{I}(\theta) + 1. \tag{15.20}$$

For multiple latent traits, the multivariate standard normal prior with correlation matrix $\boldsymbol{\Sigma}$ adds to the ML test information function as follows

$$\mathcal{I}_P^a(\boldsymbol{\theta}) = \mathcal{I}^a(\boldsymbol{\theta}) + \left[\boldsymbol{\Sigma}^{-1}\right]_a, \tag{15.21}$$

where $\left[\boldsymbol{\Sigma}^{-1}\right]_a$ is the a^{th} diagonal element of the inverted latent trait correlation matrix (for proof, see Appendix B in Brown & Maydeu-Olivares, 2011). When all traits are uncorrelated, the term added to the ML test information in equation (15.21) equals unity. When the traits are correlated positively, the term is greater than unity (and therefore the prior distribution contributes more information for the trait estimation).

Reliability

While the standard error functions (15.12) to (15.13) characterize the precision of measurement provided by an instrument completely, it may be convenient in applications to

offer a single index of precision. Such a single index is more likely to appeal to, and be understood by, the test user, and it enables a quick evaluation of a PRO measure's overall measurement quality by third parties, such as regulatory authorities. Provision of simple indices is particularly beneficial for multi-trait measures, where the main challenge is to present and summarize the information functions in a comprehensible fashion. Contour plots and "clam shell" plots (see Ackerman, 2005; and Reckase, 2009) are already complicated when just two dimensions are involved, but with more dimensions, they become quite impractical.

One such index, *marginal reliability*, was suggested by Green and colleagues (1984):

$$\rho = \frac{\sigma^2 - \bar{\sigma}^2_{error}}{\sigma^2} = 1 - \frac{\bar{\sigma}^2_{error}}{\sigma^2} . \tag{15.22}$$

This coefficient follows the fundamental definition of reliability as the proportion of variance in the observed score due to true score (equal to observed minus error variance). Provided the standard error function is relatively uniform, the marginal reliability estimate will be representative of the test's overall reliability, and may be used for comparison with classical reliability statistics. In addition, the reliability is an important quantity because it describes how closely the estimated trait resembles the latent (true) trait score:

$$\text{corr}(\theta, \hat{\theta}) = \sqrt{\rho} . \tag{15.23}$$

There are two ways to compute the marginal reliability coefficient. One way, referred to as *theoretical reliability* (du Toit, 2003), involves averaging the squared standard error for all trait values **in the theoretical distribution** to obtain the error variance in (15.22):

$$\bar{\sigma}^2_{error} = \int_{-\infty}^{\infty} SE^2(\theta)\phi(\theta)d\theta . \tag{15.24}$$

Once the error variance has been estimated (usually by approximating the integral (15.24)), the marginal reliability is computed by assuming the latent trait score variance in (15.22) $\sigma^2 = 1$, giving

$$\rho = 1 - \bar{\sigma}^2_{error} . \tag{15.25}$$

With multiple traits, the theoretical reliability procedure becomes unattractive since it involves multidimensional integration. An alternative approach, *empirical reliability*, involves averaging the squared standard errors of estimated trait scores **in a sample** to obtain

the error variance in (15.22). That is, the observed and the error variance are computed from the estimated trait scores $\hat{\theta}$ and their standard errors in a sample of *N* respondents:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \bar{\hat{\theta}}\right)^2, \qquad \hat{\bar{\sigma}}^2_{error} = \frac{1}{N}\sum_{j=1}^{N} SE^2\left(\hat{\theta}_j\right). \tag{15.26}$$

The standard error is computed as appropriate for the estimator, using the standard deviation of the posterior likelihood (15.9) for the EAP estimator, the inverse of the Fisher information for the ML estimator, and the inverse of the posterior information (15.12) for the MAP estimator. Conveniently, some software packages such as TESTFACT (Bock et al., 2003) or Mplus (Muthen & Muthen, 1998-2012) compute the standard error of a score as part of the score estimation process[1].

The empirical reliability is very easy to compute for a single trait, but its computational simplicity becomes especially advantageous when dealing with multi-trait measures. The sample-based empirical reliability approach essentially uses the estimated scores $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_{1j}, \hat{\theta}_{2j}, \ldots, \hat{\theta}_{Tj})$ for every respondent *j* as a sample of points on the multidimensional grid.

Typically, single reliability indices provide good summaries of measurement precision for instruments with relatively uniform test information functions. For instruments with peaked information functions, summary indices are unlikely to be representative indicators of precision for any ranges of the latent trait. In addition, Bayesian estimators tend to shrink the scores towards the population mean, particularly when the number of items is small. When this is the case, the observed variance $\hat{\sigma}^2$ computed using (15.26) may be low, and the empirical index might underestimate the actual reliability.

Scoring and Estimating Measurement Precision in Applications

In the following sections, we show how test scores and their standard errors are estimated for a range of item response models popular in patient-reported outcomes applications. Four models – the unidimensional model, the correlated traits model, the second-order model, and the bifactor model – are included in the discussion. We illustrate all points made with the same data analysis example.

---

[1] Mplus 7 provides standard errors for the estimated scores only when EAP estimator is used.

Data Example. The 28-item 'scaled' General Health Questionnaire (GHQ-28) in a general population birth cohort sample (the NSHD)

The 28-item version of Goldberg's General Health Questionnaire (GHQ-28; Goldberg, 1972; Goldberg and Hillier, 1979; Goldberg & Williams, 1988) is a popular instrument that was initially developed as a screening questionnaire for detecting psychiatric disorders in community settings and non-psychiatric clinical settings, such as primary care or general practice (Croudace et al, 2003; Goldberg et al. 1997). A vast literature of clinical and general population studies has indicated that the 28-item version of the GHQ[2] performs well as a screening instrument for minor psychiatric disturbance and as an indicator of psychiatric morbidity. Although the GHQ-28 was developed to detect potential psychiatric cases by measuring level of non-specific psychiatric morbidity by probing for signs and symptoms, it is often considered as a potentially multidimensional measure. It can be used as a population-wide measure of psychological distress, for example in the first stage of a two-stage epidemiological study of mental health. It is also a candidate outcome measure for clinical trials, as a primary or secondary measure of emotional distress.

Respondents completing the GHQ questionnaire are asked to think about their health in general and any medical complaints they have had over **the past few weeks**. The version used here consists of 28 items designed to measure four **a priory** facets of mental health variation (four subscales): Somatic Symptoms, Social Dysfunction, Anxiety / Insomnia, and Severe Depression / Hopelessness. Each facet is measured with seven questions (for item content see Table 15.1), and responses are given by selecting one of four categories. These are indicated by a set of 'verbal anchors'. On the copyrighted forms, no numbered scoring is displayed for each anchor and there is subtle variation in their wording, which varies according to the phrasing of the questions. The response categories refer to either absence of the symptom ('*not at all'*), or presence of the symptom in relation to the "usual" amount – i.e. *'no more than usual'*, '*rather more than usual'*, '*much more than usual'*, for instance:

"Have you recently lost much sleep over worry?"

(Not at all - No more than usual - Rather more than usual - Much more than usual)

For some questions, response categories are unique but still follow the same pattern of

[2] copyrighted instrument distributed by GL Assessment, see FAQ at www.gl-assessment.co.uk/health_and_psychology/resources/general_health_questionnaire/faqs.asp?css=1

comparison with the "usual" state of things, for instance:

"Have you recently felt capable of making decisions about things?"

(More so than usual - Same as usual - Less so than usual - Much less capable)

The GHQ-28 has 20 items that are worded in terms of "psychological problems / distress" (i.e. negatively worded) as the former example question, and 8 items that are worded in terms of "psychological health" (positively worded) as the latter question. Regardless of whether the question is worded positively or negatively, the response categories are always presented in the order of increasing problems or symptomatic distress. Thus, the last category represents the most severe symptoms or the greatest *psychological / emotional distress* for all items.

Our dataset comprises responses from the ongoing Medical Research Council National Survey of Health and Development (NSHD), also known as the British 1946 birth cohort. We refer to the wave of interviewing undertaken in 1999 when the participants were aged 53 (Wadsworth et al, 2003). At that time, 2091 respondents (1422 men and 1479 women, about 99% of the total number of study participants) provided answers on all items of the GHQ-28, and this is the full-response-set sample we consider here.

As would be expected, in this predominantly healthy mid-life community sample, most respondents selected either the first or the second response categories for most questions (corresponding to an absence or only a small amount of symptoms or distress). All item responses were positively skewed, particularly for the rather extreme items of the Severe Depression facet. Despite the large size of the dataset, the last response category yielded only single frequency figures for some items, and to avoid instability of parameter estimates we collapsed the two top response categories prior to any analyses, effectively resulting in the item coding **0-1-2-2**.

We start by exploring the factor structure of these data in M*plus*[3] (Muthén & Muthén, 1998-2010) using diagonally weighted least squares (DWLS) estimation (the software default) based on the matrix of polychoric correlations (Muthén, du Toit & Spisic, 1997). Considering that the instrument was designed to measure four facets of psychological distress, we expect to find that four moderately correlated factors underlie the data. However, there is also a clear possibility of one general "psychological distress" construct underpinning the associations between these four related facets.

---

[3] Analyses can also be performed using freeware FACTOR program, available on http://psico.fcep.urv.es/utilitats/factor/index.html

An exploratory analysis yielded five eigenvalues greater than one (12.9, 2.6, 2.2, 1.5 and 1.1). As the ratio of the first to the second eigenvalue is very large, a strong general factor is evident, together with three or four further factors. Goodness of fit of this exploratory solution is $\chi^2 = 3917$, $df = 272$, $p < 0.001$, RMSEA = 0.068, CFI = 0.959.

A target rotation of the four-factor solution to a hypothesized structure (where items designed to measure a facet have zero loadings on all other facets) yielded what we consider a well-behaved solution. Items designed to measure a facet had large loadings on the corresponding factor and mostly near-zero loadings on the other three factors (see Table 15.1). As expected, these four factors were moderately correlated (their inter-correlations ranged from 0.33 to 0.56).

INSERT TABLE 15.1 ABOUT HERE

Based on theoretical considerations governing the design of the GHQ-28 and the above exploratory solution, any psychometric model for this data should reflect the interrelated nature of the measured facets. These relationships could be modelled through a factor capturing variation between individuals on the primary psychological adjustment / distress continuum, and possibly further factors, that may exist over and above the general factor. In what follows, we consider four alternative models for the GHQ-28 data.

The first model is a **unidimensional** model, in which one common factor explains all variation in the data (Figure 15.1a). This is the most basic IRT model and we use it here for illustration of the basic scoring principles.

The second model is a **correlated traits** model, in which four freely correlated facets are indicated by their respective items (Figure 15.1b). This model is useful if the focus of measurement is on the four facets.

The third model is a **second-order** model, in which a second-order factor explains common variance in the four facets (see Figure 15.1c). This model is useful when scores on the general factor as well as scores on its facets are of interest.

The fourth and the last is a **bifactor** model, in which every item is influenced by two factors – the general factor and a group factor (Figure 15.1d). The general factor accounts for the common variance shared by all items. Group factors are “residual” dimensions, which account for any remaining common variance specific to the items forming the facets. The bifactor model is a useful representation of data where the general factor is of particular, but

not sole interest.

INSERT FIGURE 15.1 ABOUT HERE

Scoring under the Unidimensional IRT Model

When the instrument items depend on one latent trait, the ML item information function (15.14) is calculated from the normal ogive item response function $P_i(\theta) = \Phi(\alpha_i + \beta_i\theta)$ as follows (McDonald, 1999; page 284):

$$\mathcal{I}_i(\theta) = \frac{\beta_i^2\left[\phi(\alpha_i + \beta_i\theta)\right]^2}{\Phi(\alpha_i + \beta_i\theta)\left[1 - \Phi(\alpha_i + \beta_i\theta)\right]}. \qquad (15.27)$$

Here, $\phi(x)$ denotes a standard normal density function evaluated at *x*. Having computed the item information functions for all items, the ML test information is computed by adding all item contributions, and the posterior information is computed using (15.20).

When the logistic link $P_i(\theta) = \mathrm{L}(\mathrm{D}\alpha_i + \mathrm{D}\beta_i\theta)$ is used[4], the item information function (15.14) amounts to

$$\mathcal{I}_i(\theta) = (\mathrm{D}\beta_i)^2\,\mathrm{L}(\mathrm{D}\alpha_i + \mathrm{D}\beta_i\theta)\left[1 - \mathrm{L}(\mathrm{D}\alpha_i + \mathrm{D}\beta_i\theta)\right]. \qquad (15.28)$$

All formulae for item information involving the normal-ogive link function given further in this chapter can be adopted for the logistic link by using the general expression above.

Illustration: GHQ-28 general factor scored under the unidimensional model

It is important to note that the above formulae for unidimensional IRT models cannot be applied if the key assumption of local independence does not hold in the data. In particular, when local independence does not hold the test information cannot be decomposed into the sum of item information functions. We know that dependencies exist between items within the four facets of GHQ-28, however, for purposes of illustration we proceed as if the GHQ-28 really did measure only one underlying trait, "psychological distress". This analysis, using the simplifying assumption of local independence, will be later compared with analyses

[4] D=1.7 is the scaling constant used with the intercept and slope parameters from the normal ogive metric

using other models, and we shall investigate the extent to which the simplification affects the results.

Item parameters were estimated according to the single factor model depicted in Figure 15.1a, using the DWLS estimator in M*plus* software. We report the chi-square and other fit indices for this and subsequent models in Table 15.2. As expected, the model fitted the data poorly, with RMSEA and CFI/TLI indicating unsatisfactory fit. Specifically, this model failed to reproduce correlations between the items within the four facets.

INSERT TABLE 15.2 ABOUT HERE

The individual trait scores were estimated using two Bayesian methods - the EAP and the MAP. The EAP scores ranged from -2.96 to 3.52 (mean 0.03 and standard deviation 0.90). The MAP scores range from -2.87 to 3.42 (mean 0.03 and standard deviation 0.89).The two sets of scores correlate almost perfectly, with the correlation being different from 1 in the fifth decimal place, and in the rest of this chapter we therefore use the EAP score only.

The standard errors of the EAP score estimates were computed according to equation (15.9). Figure 15.2 shows the EAP standard errors plotted against the estimated trait scores for all individuals in the NSHD sample. For comparison, we also plot the standard errors computed from the Fisher information of the posterior likelihood function using equation (15.20). It can be seen that the standard error function based on the Fisher information traces the errors of the EAP scores almost perfectly. Figure 15.2 shows that for the most of the latent score range (between -3 and 3) the precision of measurement is good, with the standard errors below 0.5.

INSERT FIGURE 15.2 ABOUT HERE

The empirical reliability of the MAP sample scores (computed from Fisher information) is **0.899**, and empirical reliability of EAP sample scores computed from the squared EAP standard errors is also **0.908**. In this case, the two empirical estimators give similar results. We, however, suspect that this high reliability values are an overestimation, due to ignoring the local dependencies between items measuring the same facets in this simple model.

Scoring under the Correlated Traits IRT Model

In measures where items each item indicates only one of T correlated facets f_1, f_2,…, f_T (the items possess an independent-clusters basis), such as our hypothesized structure for the GHQ-28 depicted in Figure 15.1b, the one-dimensional theory applies when it comes to computing the ML item information and test information functions. Because each item response is conditional on a single trait, there is no difference between estimating trait scores by using a model with multiple correlated traits and by using separate unidimensional models, provided that the item parameters in both models are identical.

There is, however, a difference between estimating the trait scores in univariate and multivariate fashion when Bayesian estimators are used. In a model with correlated traits, the latent trait correlations will be taken to account as a feature of the multivariate prior distribution. When the traits are associated, the population-based likelihood of having a certain score on one trait will vary depending on the levels of other traits. For example, in the GHQ-28 with its positively correlated facets, a combination of four scores that are all 1 standard deviation below the population mean is more likely than a combination of scores where two facet scores are 1 standard deviation above and two facet scores are 1 standard deviation below the population mean. When the appropriate prior population density is used in Bayesian estimation, it increases the estimation precision by "borrowing" information from related trait scores. To quantify how much information is added by related scores, we refer to equation (15.21).

Illustration: GHQ-28 facets scored under the correlated traits model

The correlated traits model (see Figure 15.1b) was estimated using the DWLS estimator. Goodness-of-fit indices are reported in Table 15.2. As expected, the model fitted much better than the unidimensional model. After the model parameters were estimated, the individual responses were scored using the MAP estimator (since the EAP estimator is very computationally heavy with as many as four dimensions). Model-based correlations between the latent traits are given in Table 15.3. These correlations are used to assess the contributions from the prior multivariate normal distribution to estimation of each subscale score. Because the four facets are positively correlated, the prior information provided by each facet in (15.21) is greater than 1. For instance, the prior distribution adds impressive 3.33 to the information function for Anxiety / Insomnia subscale.

INSERT TABLE 15.3 ABOUT HERE

The standard errors of the facet scores estimated in the multivariate fashion as described above are given in Figure 15.3 (a-d). Corresponding standard errors of the univariate score estimation (if the scores were MAP-estimated without the knowledge of the related facets) are shown for comparison. It can be seen that the multivariate scoring yields higher accuracy than the scale-by-scale univariate scoring when traits are correlated.

Marginal reliabilities computed for each estimated facet are given in Table 15.4. It can be seen that the Severe Depression score shows particularly high shrinkage towards the population mode, producing low observed score variance (0.64). The empirical reliabilities for Somatic Symptoms and Anxiety subscales are above 0.8, but fall short of 0.7 for Social Dysfunction and Severe Depression.

INSERT FIGURE 15.3 AND TABLE 15.4 ABOUT HERE

Scoring under the Bifactor IRT Model

In this section, we use the general formulation for response functions and multidimensional information given earlier in this chapter to provide formulae necessary to compute the item and test information for bifactor models. Let g be a *general factor* measured by the instrument, and s_1, s_2,…, s_T be *T group factors* (for instance, related to facets). The set of factors underlying item responses is therefore $\mathbf{g}^* = \left(g, s_1, s_2, ..., s_T\right)'$. At this point, it is important to emphasize the difference in conceptual meaning between subscales / facets and "group factors". In bifactor models, group factors do **not** have the same interpretation as facets in correlated traits models. Statistically, they are residuals left after accounting for the general factor, and therefore represent unique features common to groups of items that are not explained by the general factor. The group factors, therefore, cannot be thought of as 'Anxiety' or 'Depression' or whatever meaning we associate with facets – instead, they represent common features of 'Anxiety' or 'Depression' items that are not captured by 'Psychological Distress'.

The general factor and the group factors are assumed mutually uncorrelated, have zero means and unit variances so that their distribution is multivariate standard normal

$\mathbf{g}^* \sim N_{T+1}(0, \mathbf{I})$, where **I** denotes the identity covariance matrix. In the bifactor model, the item response tendency is influenced by two factors – the general factor *g* and one group factor, say s_a.

$$u_i^* = \alpha_i + \beta_{0i} g + \beta_{ai} s_a + \varepsilon_i . \tag{15.29}$$

In the above expression, α_i is the intercept, β_{0i} is the slope parameter for the general factor *g*, β_{ai} is the slope for the group factor s_a, and ε_i is the error. Assuming the error is distributed normally with the mean 0 and unit variance, the item response function given in the intercept / slope parameterization is, therefore, the two-dimensional normal ogive

$$P_i(\mathbf{g}^*) = P(u_i = 1 | \mathbf{g}^*) = \Phi(\alpha_i + \beta_{0i} g + \beta_{ai} s_a) . \tag{15.30}$$

Because all the factors in this model are uncorrelated, the gradient in direction of the general factor *g* in equation (15.17) becomes

$$\nabla_g P_i(\mathbf{g}^*) = \frac{\partial P_i(\mathbf{g}^*)}{\partial g} = \beta_{0i} \phi(\alpha_i + \beta_{0i} g + \beta_{ai} s_a), \tag{15.31}$$

where $\phi(x)$ is the normal density function evaluated at *x*. Thus, the IIF in direction of the general factor *g* for the bifactor model is

$$\mathcal{I}_i^g(\mathbf{g}^*) = \frac{(\beta_{0i})^2 \left[\phi(\alpha_i + \beta_{0i} g + \beta_{ai} s_a)\right]^2}{\Phi(\alpha_i + \beta_{0i} g + \beta_{ai} s_a)\left[1 - \Phi(\alpha_i + \beta_{0i} g + \beta_{ai} s_a)\right]} . \tag{15.32}$$

Since the item information function is conditioned on both the general and the group factor, local independence holds and the IIFs are additive. The test information about the general factor *g* is the sum of all IIFs in the direction of *g*. This summation, however, will make the test information function for the general factor **conditional on all group factors**, although each item information function is only conditional on **one** group factor (in addition to the general factor).

When prior knowledge about the population distribution is used for score estimation, the prior information must be added to the ML test information to compute the posterior test information. In the bifactor model, the latent covariance matrix is **Σ** = **I** and the information added in (15.21) is simply 1:

$$\mathcal{I}_P^g(\mathbf{g}^*) = \mathcal{I}^g(\mathbf{g}^*) + \left[\mathbf{I}^{-1}\right]_g = \sum_i \mathcal{I}_i^g(\mathbf{g}^*) + 1 \tag{15.33}$$

Illustration: GHQ-28 general and group factors scored under the bifactor model

The bifactor model illustrated in Figure 15.1d, with the latent factor variances set to 1 and the item uniqunesses set to 1, involved estimating 28 slope parameters β_{0i} related to the general factor, and 28 slope parameters $\beta_{Si}, \beta_{Ai}, \beta_{SDi}, \beta_{Di}$ related to the four group factors s_S, s_A, s_{SD}, s_D. In addition, 56=28*2 category thresholds were estimated (two thresholds for each 3-category item). Goodness-of-fit indices for this model are reported in Table 15.2. The model shows acceptable CFI/TLI, and it is an improvement compared to the correlated traits model. The person parameters (general and group scores) for the GHQ-28 were estimated using the MAP estimator.

The properly conditioned and therefore additive **item** information functions (IIFs) for this model depend on two factors only (one general and one group factor) and can be plotted as surfaces. However, the **test** information function (TIF) in the direction of the general factor depends on five factors (the general factor as well as four group factors), and therefore cannot be presented easily.

We use estimated scores for each person j $\left(\widehat{\mathbf{g}}_j, \hat{\mathbf{s}}_{Sj}, \hat{\mathbf{s}}_{Aj}, \hat{\mathbf{s}}_{SDj}, \hat{\mathbf{s}}_{Dj}\right)$ as sampled points on the multidimensional grid to summarize high-dimensional information functions, in the following manner. For each sampled point, item information functions will yield scalar values, the TIF for the general factor can be easily computed by summing the scalar IIF values. Finally, the standard errors can be computed and plotted against estimated values of the general factor for all individuals in the sample. Figure 15.4 illustrates such a scatter for the GHQ-28 data. It can be seen that the standard errors for most participants were below 0.4, and only two participants out of 2901 had the standard errors above 0.5. These larger standard errors were associated with low scores on the general factor (θ = -1.78 and θ = -2.48). The empirical reliability of MAP scores calculated from these standard errors is **0.901**, very close to the estimate given by the unidimensional model.

INSERT FIGURE 15.4 ABOUT HERE

Scoring under the Second-Order IRT Model

In measures with a second-order factorial structure (such as the model depicted in Figure 15.1c), common variance in T facets f_1, f_2,…, f_T is explained by one general factor g,

and every item response tendency depends on one first-order factor (facet), say f_a:

$$u_i^* = \alpha_i + \beta_{ai} f_a + \varepsilon_i \,. \tag{15.34}$$

Here, α_i is the item intercept, β_{ai} is the slope for the facet f_a, and ε_i is the error distributed with the mean 0 and unit variance (thus we keep the standard IRT intercept/slope parameterization for the measurement part of the model).

Each facet factor, in turn, is an indicator of the general factor, g:

$$f_a = \lambda_a g + d_a \,, \tag{15.35}$$

where d_a is a residual term, also called "disturbance" in structural equation modeling (SEM) literature. The residual terms are uncorrelated with the general factor and with each other. The structural part of the model in equation (15.35) is parameterized so that the general factor g and the facet factors f_1, f_2,…, f_T have unit variances, i.e. the factor loadings λ_a are standardized. It is convenient to work with standardized facet scores, and use the established formulae for the standard normal ogive etc. In order for the facet scores to be standardized, the disturbance terms must have variances $\mathrm{var}\left(d_a\right) = 1 - \lambda_a^2$.

It follows that the set of latent factors in the measurement model (15.34) $\mathbf{f} = \left(f_1 \,, f_2 \,, ..., f_T \right)'$ is distributed $\mathbf{f} \sim N_T\left(0, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ is the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_T \\ \lambda_2\lambda_1 & 1 & \cdots & \lambda_2\lambda_T \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_T\lambda_1 & \lambda_T\lambda_2 & \cdots & 1 \end{pmatrix}. \tag{15.36}$$

In the measurement model (15.34), item response tendencies depend on facet factors only and do not depend on the general factor. The item response function can be written as the standard normal ogive

$$P_i\left(\mathbf{f}\right) = P\left(u_i = 1 \middle| \mathbf{f}\right) = \Phi\left(\alpha_i + \beta_{ai} f_a\right). \tag{15.37}$$

Because the item response tendency u_i* depends on one facet f_a only, and does not depend on the general factor, the partial derivatives with respect to the general factor and irrelevant facets are zero, and the gradient in direction of the general factor g in equation (15.17) becomes

$$\nabla_g P_i(\mathbf{f}) = \frac{\partial P_i(\mathbf{f})}{\partial g}\text{corr}(g,g) + \sum_{k=1}^{T}\left(\frac{\partial P_i(\mathbf{f})}{\partial f_k}\text{corr}(f_k, g)\right) = \frac{\partial P_i(\mathbf{f})}{\partial f_a}\text{corr}(f_a, g). \quad (15.38)$$

The partial derivative with respect to f_a in (15.38) is

$$\frac{\partial P_i(\mathbf{f})}{\partial f_a} = \beta_{ai}\phi(\alpha_i + \beta_{ai} f_a), \quad (15.39)$$

where $\phi(x)$ is the normal density function evaluated at x. It follows from (15.35) that the correlation between the general factor g and the facet factor f_a is $\text{corr}(f_a, g) = \lambda_a$, and the gradient is

$$\nabla_g P_i(\mathbf{f}) = \beta_{ai}\lambda_a \ \phi(\alpha_i + \beta_{ai} f_a). \quad (15.40)$$

Finally, the IIF **in the direction of the general factor *g*** for an item measuring the facet factor f_a in the second-order model, parameterized as described above, is

$$\mathcal{I}_i^g(\mathbf{f}) = \frac{(\beta_{ai}\lambda_a)^2\left[\phi(\alpha_i + \beta_{ai} f_a)\right]^2}{\Phi(\alpha_i + \beta_{ai} f_a)\left[1 - \Phi(\alpha_i + \beta_{ai} f_a)\right]}. \quad (15.41)$$

In the second-order model, local independence holds and the item information functions are additive. The TIF about the general factor g is the sum of all IIFs in the direction of g. This summation, just as it was the case with the bifactor model, makes the test information function for the general factor **conditional on all facet factors**, although each item information function is only conditional on one facet factor.

When Bayesian methods are used for score estimation, the prior information must be added to the ML test information to compute the posterior test information. In the second-order model, the general factor is estimated from continuous facet indicators, and the amount of information added by the univariate normal prior is 1

$$\mathcal{I}_P^g(\mathbf{f}) = \mathcal{I}^g(\mathbf{f}) + 1. \quad (15.42)$$

An added advantage of the second-order model is that along with the scores for the general factor, scores for the facets can also be estimated. It follows from equation (15.37) that the ML item information **in direction of the facet factor f_a** is

$$\mathcal{I}_i^a(\mathbf{f}) = \frac{\beta_{ai}^2\left[\phi(\alpha_i + \beta_{ai} f_a)\right]^2}{\Phi(\alpha_i + \beta_{ai} f_a)\left[1 - \Phi(\alpha_i + \beta_{ai} f_a)\right]}. \quad (15.43)$$

The TIF about the facet factor f_a is the sum of all IIFs in the direction of f_a. Just like in the model with correlated traits, relationships between the facet factors described by (15.36) will be taken to form the covariance matrix for the multivariate normal distribution used as a prior. To quantify how much information is added by related scores, we refer to equation (15.21).

The Second-Order Model as a Special Case of the Bifactor Model

If we substitute the expression for the facet factor (15.35) into the expression for the item response tendency (15.34), we obtain an expression of the response tendency as a direct function of the general factor and the residual (disturbance) term:

$$u_i^* = \alpha_i + \beta_{ai}\lambda_a g + \beta_{ai} d_a + \varepsilon_i \,. \tag{15.44}$$

The only difference between this expression and the bifactor equation (15.29) is the naming convention for the residual / disturbance term, and choice of regression parameters (slopes). We keep notation for the disturbance terms d_a in the second-order model distinct from the group terms s_a in the bifactor model only to emphasize that they have different variances in our chosen parameterization. Recall that while variances of the group terms in the bifactor model are set to 1, variances of the disturbance terms in the second-order factor model are set to 1 minus squared factor loading on the general factor ($1-\lambda_a^2$). Otherwise, these entities have equivalent roles in the models and equivalent meaning.

It follows that any second-order model can be equivalently parameterized as a bifactor model with some restrictions on slope parameters (Rindskopf & Rose; 1988). That is, the slope on the general factor in a bifactor representation of the second-order model is proportionate to the respective slope on the group factor, with one common proportion coefficient λ_a for all items related to the same group factor s_a:

$$\beta_{0i} = \beta_{ai}\lambda_a \,. \tag{15.45}$$

Consequently, the bifactor model estimates twice as many slope parameters as there are items, while the second-order model estimates as many slopes on the group factors as there are items plus as many common facet coefficients as there are facets. Put simply, the second-order model is nested within the bifactor model.

It is now clear that the expression for the IIF in the direction of the general factor for the second-order model (15.41) is identical to the expression for the bifactor model (15.32),

when the two models are parameterized as described. That is, in the bifactor model we set variances for the general, group and unique factors equal 1. In the second-order model, we set variances for the general, facet and unique factors equal 1 (this means that variances for the facet residuals/disturbances are constrained to 1 minus squared factor loading on the general factor). This parameterization ensures a familiar z-score metric interpretation, and scale, for all model-estimated factors.

Illustration: GHQ-28 general factor and facets scored under the second-order model

The second-order model for the GHQ-28 illustrated in Figure 15.1c, with the general and facet factors' variances set to 1 (and recall that item uniqunesses are set to 1), estimates 4 facet loadings on the general factor, and 28 item slope parameters relating to the four facet factors f_S, f_A, f_{SD}, and f_D. There are also 56=28*2 category threshold parameters (two thresholds for each 3-category item). Goodness-of-fit indices for this model are reported in Table 15.2. This model fitted much better than the unidimensional model and almost as well as the correlated traits model. The scores on the general factor and the facets were estimated using the MAP estimator.

The resulting test information function (TIF) in direction of the general factor depends on all four facets, although information functions for each item depend on one facet only. This means that the properly conditioned and therefore additive **item** information functions can be plotted as curves; however, the 4-dimensional **test** information function cannot be plotted. Again, we use the sample-based approach to working with information functions here, as follows.

Once the factor scores $\widehat{\mathbf{f}}_j = \left(\hat{f}_{Sj}, \hat{f}_{Aj}, \hat{f}_{SDj}, \hat{f}_{Dj}\right)$ have been estimated for each individual j, they are used as sampled points. For each point, the item information functions yield scalar values, and the TIF for the general factor is easily computed by summing the IIFs. The prior contribution is then added as described by (15.42). Finally, the standard errors can be computed and plotted against the estimated values of the general factor for all individuals in the sample. Figure 15.5 illustrates this scatter for our GHQ-28 data example. It can be seen that the standard errors for most participants were below 0.4, and only one participant out of 2901 had the standard error above 0.5. This larger standard error (SE = 0.54) was associated with a low score on the general factor (θ = -2.32). The empirical reliability of MAP scores calculated from these standard errors is **0.891**, lower than the estimate for the unidimensional

model, and the bifactor model.

INSERT FIGURE 15.5 ABOUT HERE

The standard errors of the facet scores were computed using the information function (15.43). The resulting standard error functions were very similar to the respective functions estimated under the correlated traits model. The empirical reliabilities for each facet are given in Table 15.4. It can be seen that they are very similar to the respective reliability coefficients for the correlated traits model.

Evaluation of the Alternative Scoring Methods for GHQ-28 Data Example

A range of scoring options for a typical, ordinal-response patient-reported outcome measure with a general-group structure was discussed and illustrated with the GHQ-28 dataset. The unidimensional model, violating the assumption of local independence, showed poor fit to the data (see Table 15.2). The other models, which accommodated the multidimensional faceted structure of the GHQ-28 fitted the data nearly acceptably, and the bifactor model showed the best fit.

Table 15.5 reports correlations between the **general psychological distress** scores estimated with the alternative scoring models. It can be seen that all methods yielded highly similar scores (the correlations ranged from .960 to .987). The lowest correlation was observed between our worst fitting model – the unidimensional model – and our best fitting model – the bifactor model. Figure 15.6 presents scatterplots for the general factor scores estimated under different models. It can be seen that there was slight departure from linearity in the relationship between the unidimensional scores and the bifactor scores, and the unidimensional scores and the second-order scores. The second-order model yielded general factor scores that were very close to the bifactor scores, and the relationship between these two sets of scores was linear.

INSERT FIGURE 15.6 ABOUT HERE

Reliability estimates obtained using the marginal (empirical) coefficients, which were fully conditioned on all latent traits in the multidimensional models, reached around .9 for all models. These estimates, however, are likely to be overly optimistic for the unidimensional

model, which ignored local dependencies between the items within facets and therefore likely overestimated the test information and underestimated the standard errors.

We expect the bifactor reliability estimates to be an accurate reflection of the real measurement precision provided by this well-fitting scoring model. To examine how accurate the estimated scores really were for all models, we conducted a simulation study, where the bifactor model was the generating model for item responses. The estimated item parameters from the GHQ-28 bifactor model were used to generate true scores and categorical item responses for N = 2901 cases in 100 replications. The simulated item responses were scored using the same methods we applied to the real GHQ-28 dataset. The correlation between the true and the estimated scores on the general factor were computed. Recall that according to (15.23), the squared correlation between the true and the estimated score is a measure of test reliability.

Results of this simple simulation study are reported in the last column of Table 15.4. The highest squared correlation between the true and the estimated score (the test reliability) is for the bifactor model (.869) and the lowest for the unidimensional model (.824). Comparing these figures with the reliability estimates in the same table, we can see that the test information methods overestimated the reliability slightly for the bifactor and the second-order models (by 3.8% and 4.6% respectively), and more substantially for the unidimensional model (by 9.1%). The slight overestimation with the well-fitting bifactor model is likely due to remaining local dependencies between items within the Somatic Symptoms subscale.

Overall, we conclude that the bifactor model is the most suitable for these data, and that it produces scores with good measurement precision that can be reasonably accurately estimated using the described sample-based methodology involving multidimensional information. When assessment on the general factor is desired, the second-order model is the second best from the fit and the measurement precision point of view; and it may be preferred to the bifactor model in situations when scoring on the subscales is also required.

Conclusion and Discussion

Which model is the most suitable for a particular measurement instrument should be a decision based largely on conceptual grounds, not merely model fit. Furthermore, particular measurement focus (e.g. whether the general factor or the facet factors are of interest) imposes practical requirements on the scoring model. Approaching the model choice from the theoretical perspective governing the instrument's design, the GHQ-28 used as an example

patient-reported outcome measure was developed to measure four conceptually distinct constructs that were expected to show moderate to strong associations in most samples (clinical or general population). The correlated traits model is clearly suitable from this point of view. However, this model fails to capture any common variance existing among the subscales, and therefore is only suitable for measuring the facets but not the overall psychological distress.

The second-order model, on the other hand, captures not only the facets but also the general factor underlying them. The researcher may consider the four health facets to be manifestations of the common construct (psychological distress), which represent specific features that go beyond this common view/summary. For example, the individual facets may have differential prognostic significance. In such a framework, it certainly makes sense to estimate and report individual scores on the four "subscales", and the second-order model is well suited to this purpose.

In the bifactor model, no familiar 'subscale' scores are estimated in addition to the general factor. Instead, group factors are estimated, which capture residual variance after the general factor has been accounted for. Although not commonly used, these residual scores can be also considered in applied research. We expect to see them exploited in longitudinal studies with repeated PRO measures, in parallel with examining a process of change in general score values.

The second-order model would be the model of choice for understanding the GHQ-28 from the perspective that gave rise to its four-subscale structure and the perspective of its classical scoring. If, however, the researcher believes that all individual variation is caused by the general psychological distress continuum, and common content shared by the items within each subscale is in fact a "method" factor related to specific manifestation of distress symptoms, the bifactor model is more suitable. It might also be the case that only *some* of the group factors are interpretable as method factors.

Acknowledgements

The authors are grateful to Marcus Richards of the MRC Unit for Lifelong Health and Ageing for allowing us to use the GHQ-28 data for the numerical example.

Work on this chapter was supported by a grant from the Isaac Newton Trust (grant number RG63087) awarded to the first author, and by a grant from the Wellcome Trust (grant number 088869/Z/09/Z) awarded to the second author. At the time of writing, both authors were working at the University of Cambridge, Department of Psychiatry.

References

Ackerman, T.A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. McArdle. (Eds.).*Contemporary Psychometrics* (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum.

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bock, R.D., Gibbons, R., Schilling, S.G., Muraki, E., Wilson, D.T., & Wood, R. (2003). *TESTFACT 4.0 user's guide.* Chicago, IL: Scientific Software International.

Gibbons, R.D., Bock, R.D., Hedeker, D., Weiss, D.J., Segawa, E., Bhaumik, D.K., Kupfer, D.J., Frank, E., Grochocinski, V.J. & Stover, A. (2007). Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement, 31*, 4–19.

Gibbons, R.D., Immekus, J.C. & Bock, R.D. (2007). The Added Value of Multidimensional IRT Models. Didactic workbook. Retrieved on 1 June 2011 from http://outcomes.cancer.gov/areas/measurement/multidimensional_irt_models.pdf

Brown, A. & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71(3)*, 460-502.

Croudace, T.J., Evans, J., Harrsion, G., Sharp, D.J., Wilkinson, E., McCann, G., Spence, M., Crilly, C. & Brindle, L. (2003). Impact of the ICD-10 Primary Health Care (PHC) diagnostic and management guidelines for mental disorders on detection and outcome in primary care: cluster randomised controlled trial. *British Journal of Psychiatry*, *182*, 20-30.

Dodd, B.G., De Ayala, R.J. & Koch, W.R. (1995). Computerized adaptive testing with

polytomous items. *Applied Psychological Methods, 19*, 5-22.

Du Toit, M. (Ed.). (2003). *IRT from SSI*. Chicago: Scientific Software International.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Fayers, P.M. & Machin, D. (2007). *Quality of Life: The assessment, analysis and interpretation of patient-reported outcomes*. *Second edition*. John Wiley & Sons.

Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. *Philosophical transactions A, 222*, 309-368.

Green, B.F., Bock, R., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. Oxford University Press: London.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine, 27*, 191-197.

Goldberg, D. P., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine, 9*, 139-145.

Goldberg, D.P. & Williams, P. (1988). *A user's guide to the general health questionnaire*. NFER Nelson: Windsor.

McDonald, R.P. (1999). *Test theory. A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (2011). Measuring Latent Quantities. *Psychometrika, 76 (4)*, 511-536.

Masters, G.N. & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden and R. Hambleton (Eds): *Handbook of modern item response theory*. New York: Springer-Verlag.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Newbury Park, CA: Sage.

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript. College of Education, UCLA. Los Angeles, CA.

Muthén, L.K. & Muthén, B.O. (1998-2010). *Mplus User's guide. Sixth edition*. Los Angeles,

CA: Muthén & Muthén.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

Reise, S. & Haviland, M. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84(3)*, 228-238.

Rindskopf, D. & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*, 51–67.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum.

Van der Linden, W.J. & Hambleton, R. (1997). Item Response Theory: brief history, common models and extensions. In W.J. van der Linden and R. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.

Wadsworth M.E., Butterworth, S.L., Hardy, R.J., Kuh, D.J., Richards, M., Langenberg, C., Hilder, W.S. & Connor, M. (2003). The life course prospective design: an example of benefits and problems associated with study longevity. *Social Science and Medicine, 57*, 2193-2205.

Tables

Table 15.1. Standardized factor loadings in the oblique four-factor solution for GHQ-28.

Item	Wording	Factor 1 Somatic Symptoms	Factor 2 Anxiety	Factor 3 Social Dysfunction	Factor 4 Depression
S1	good health	0.53		0.39	
S2	need good tonic	0.62	0.36		
S3	run down	0.72	0.32		
S4	felt ill	0.69			
S5	pains in head	0.97			
S6	pressure in head	0.97			
S7	hot and cold spells	0.39			
A1	lost sleep over worry		0.79		
A2	difficulty staying asleep		0.67		
A3	constant strain		0.73		
A4	edgy		0.58		
A5	panicky/scared		0.49		0.33
A6	everything on top of you		0.69		
A7	nervous		0.59		0.36
SD1	manage to keep busy			0.52	
SD2	take longer time			0.55	
SD3	doing things well			0.95	
SD4	satisfied with tasks			0.86	
SD5	play a useful part			0.76	
SD6	making decisions			0.61	
SD7	enjoy daily activities		0.35	0.41	
D1	worthless				0.61
D2	hopeless				0.72
D3	not worth living				0.79
D4	thoughts of suicide				0.95
D5	nerves too bad				0.53
D6	wishing dead				0.86
D7	suicidal ideas				0.92

Note. Loadings below 0.3 in magnitude are omitted. S = Somatic Symptoms, A = Anxiety / Insomnia, SD = Social Dysfunction, D = Depression.

Table 15.2. Goodness of fit for alternative GHQ-28 models.

Model	Chi-square	df	RMSEA	CFI	TLI
Unidimensional	15314	350	.121	.831	.818
Second-order	5920	346	.075	.937	.931
Correlated traits	5870	344	.074	.938	.932
Bifactor	4142	323	.064	.957	.950

Table 15.3. Estimated latent trait correlations under the Correlated traits model

	S	A	SD	D
(S) Somatic Symptoms	1			
(A) Anxiety / Insomnia	.70 (.01)	1		
(SD) Social Dysfunction	.59 (.01)	.60 (.01)	1	
(D) Severe Depression	.54 (.02)	.76 (.01)	.57 (.02)	1

Note. Standard errors are in parentheses. S = Somatic Symptoms, A = Anxiety / Insomnia, SD = Social Dysfunction, D = Depression.

Table 15.4. Marginal (empirical) reliabilities under alternative GHQ-28 scoring models

Model	Estimator	S	A	SD	D	**General**	$\left[\text{corr}(\theta, \hat{\theta})\right]^2$
Unidimensional	MAP					**.899**	.824
	EAP					**.908**	
Correlated traits	MAP	.817	.871	.666	.639	**-**	**-**
Second-order	MAP	.813	.867	.665	.607	**.891**	.852
Bifactor	MAP	-	-	-	-	**.902**	.869

Note. S = Somatic Symptoms, A = Anxiety / Insomnia, SD = Social Dysfunction, D = Depression.

Table 15.5. Correlations between GHQ-28 general distress scores estimated using alternative scoring models.

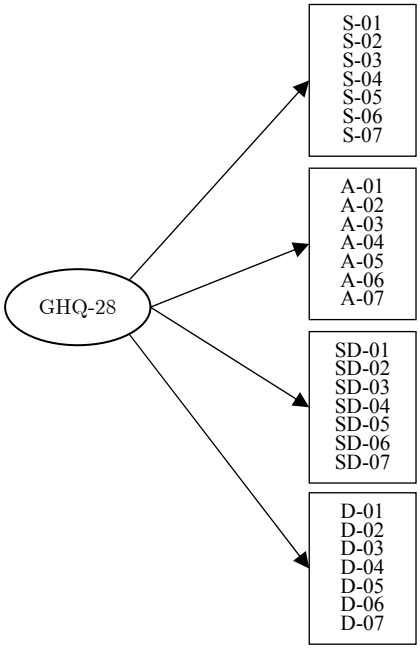
	Second-order	Bifactor
Unidimensional	.987	.960
Second-order		.982

Notes: Both EAP and MAP scoring for the unidimensional model yield identical correlations with the other models.

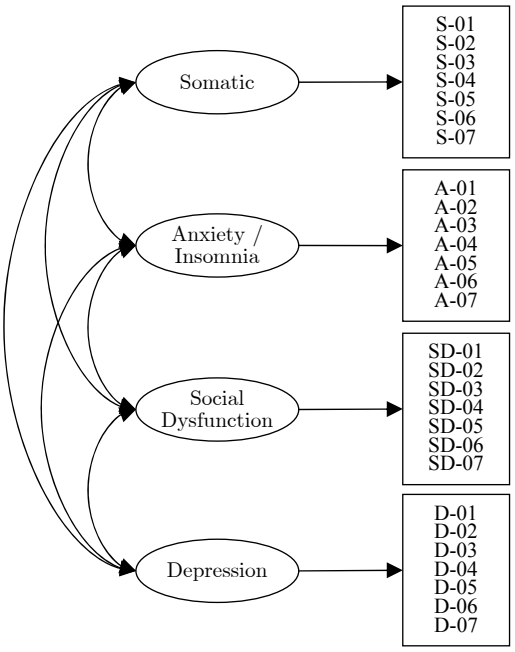
Figures

Figure 15.1. Four alternative models for GHQ-28 item responses

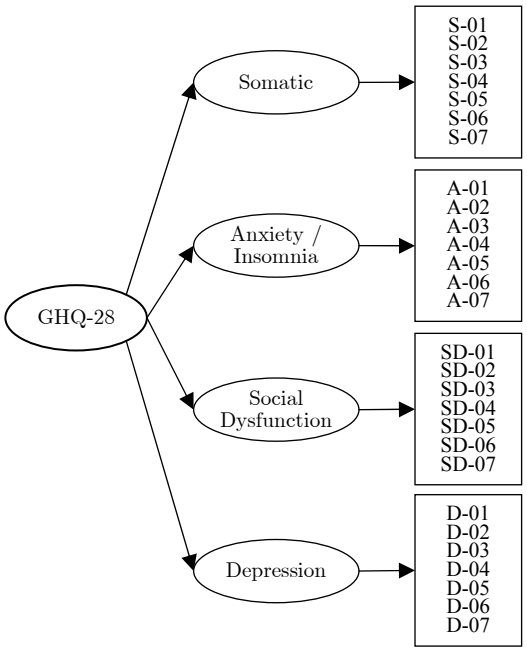
a. Unidimensional model



b. Correlated traits model



c. Second-order model



d. Bifactor model

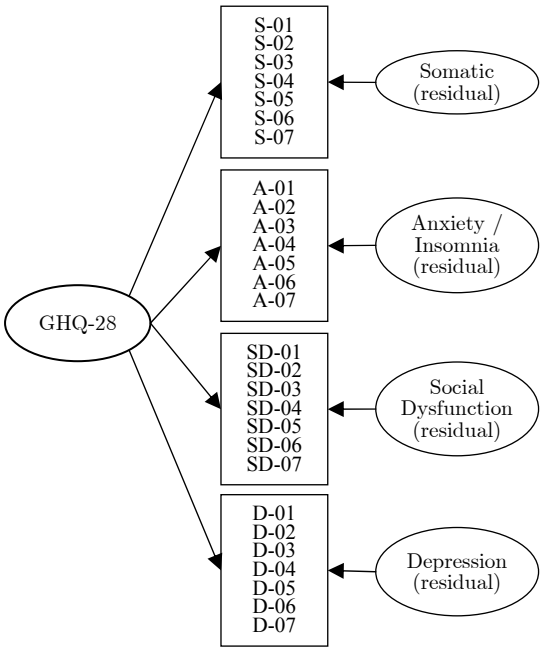
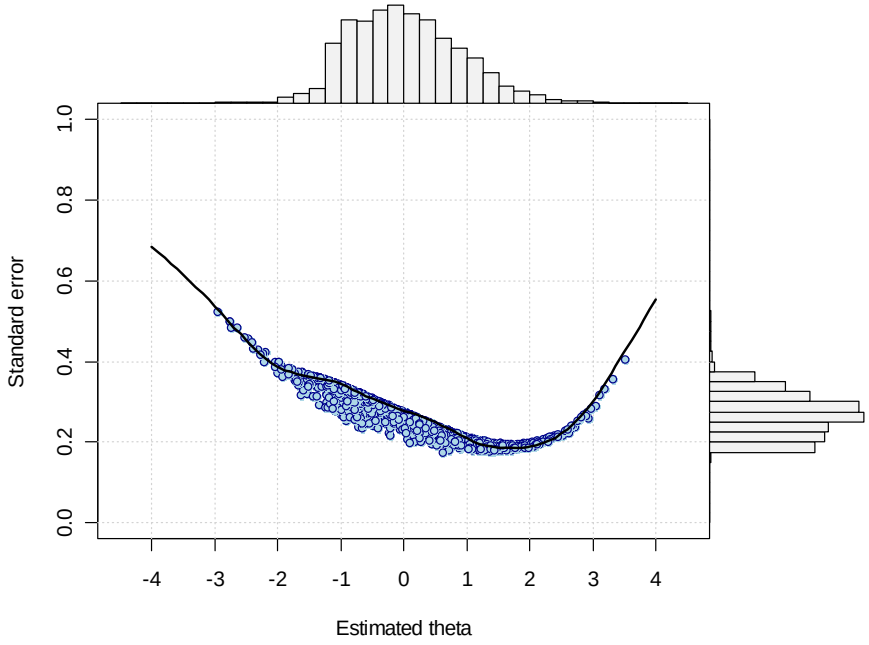
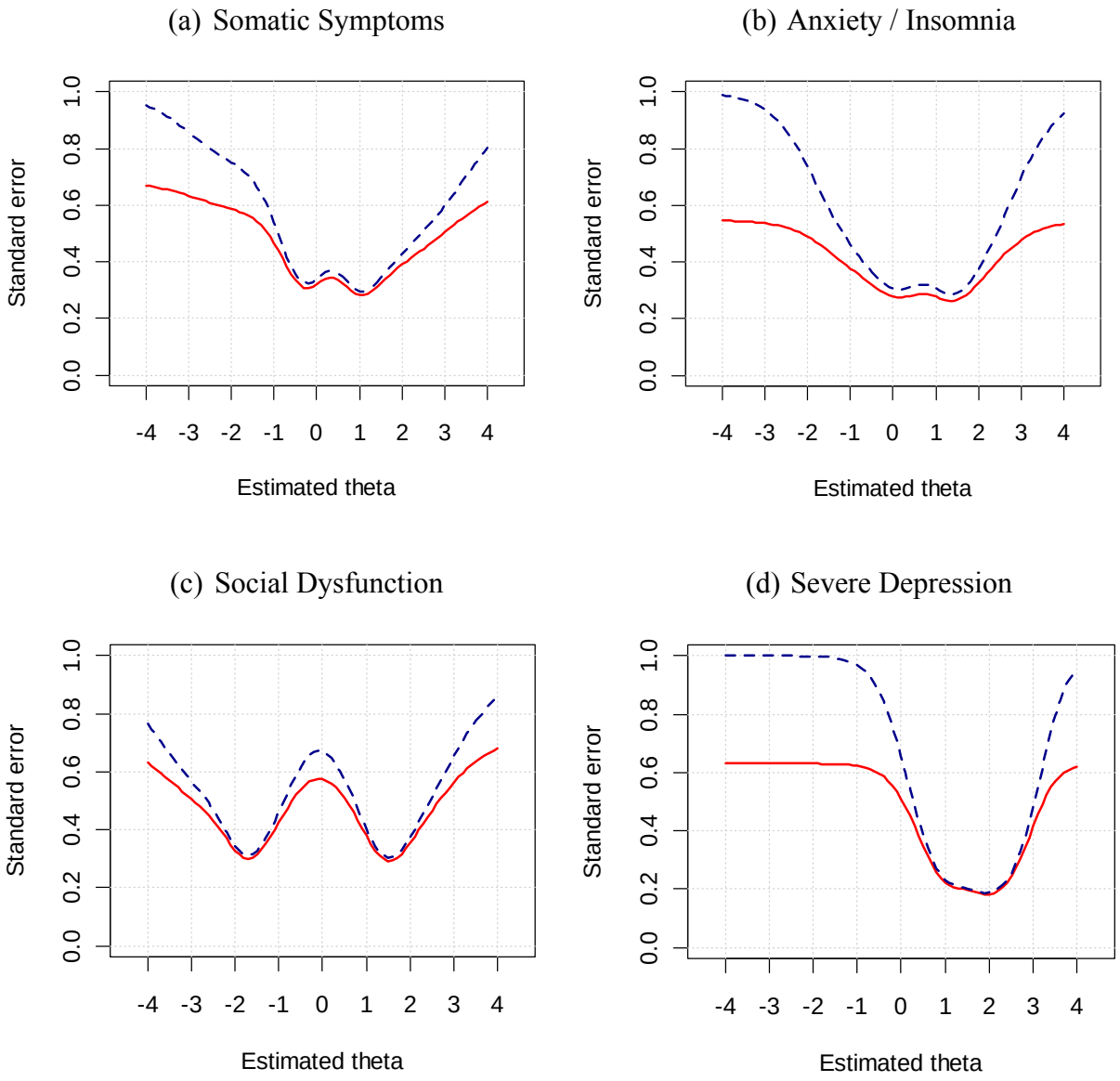


Figure 15.2. Standard errors for the general psychological distress factor under the unidimensional model, with distributions of estimated scores and standard errors.



Note. Points represent standard errors of individual EAP scores; solid line is the standard error function derived from the Fisher posterior information.

Figure 15.3. Standard error functions for the four subscales in the GHQ-28 data example; univariate versus multivariate scoring using the MAP estimator.



Note. Solid line is the standard error function for the multivariate estimation; dashed line is the standard error function for the univariate (scale-by-scale) estimation.

Figure 15.4. Standard errors of MAP general factor scores under the bifactor model in the GHQ-28 data example; fully conditioned on all latent factors.

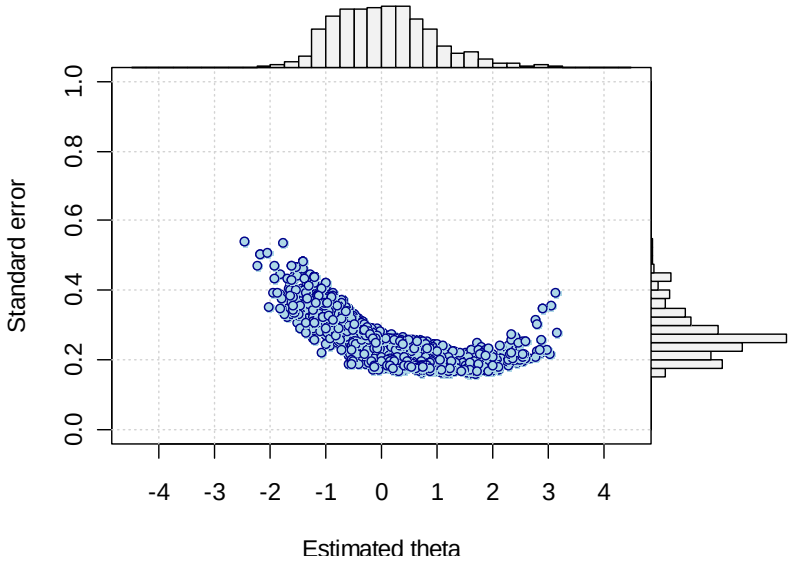


Figure 15.5. Standard errors of MAP general factor scores under the second-order factor model in the GHQ-28 data example; fully conditioned on all latent factors.

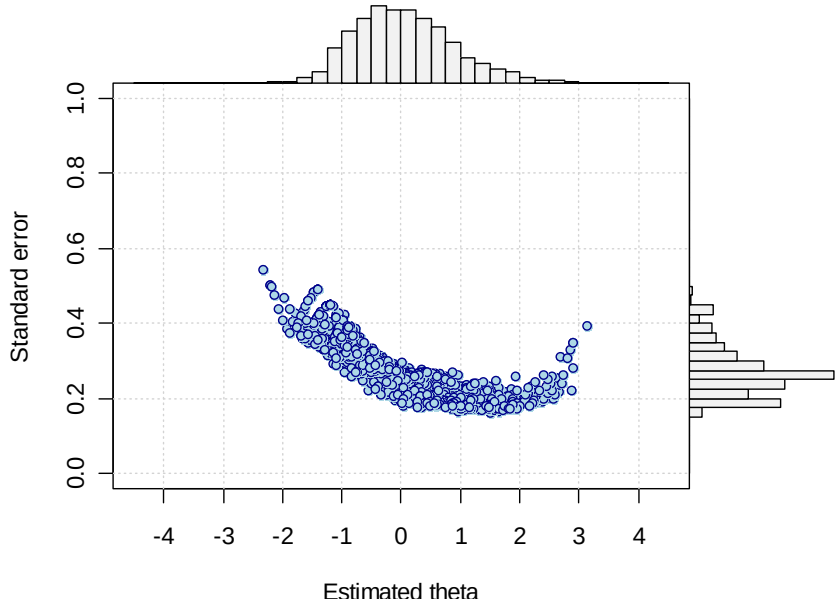


Figure 15.6. Comparison of the GHQ-28 general factor score estimates under different scoring models.