Application of Item Response Theory Models for Intensive Longitudinal Data

Donald Hedeker, Robin J. Mermelstein, and Brian R. Flay
University of Illinois at Chicago

Summary

Item Response Theory (IRT) models, also called latent trait models, have been extensively used in educational testing and psychological measurement. While an IRT model is essentially a mixed-effects regression model for longitudinal categorical data, use of IRT has been somewhat limited outside of education and psychology. In this chapter, we describe IRT models and some key developments in the IRT literature, and illustrate their application for intensive longitudinal data. In particular, we relate IRT models to mixed models and indicate how software for the latter can be used to estimate the IRT model parameters. Using Ecological Momentary Assessment (EMA) data from a study of adolescent smoking, we describe how IRT models can be used to address key questions in smoking research.

1 Introduction

Item response theory (IRT) or latent trait models provide a statistically-rich class of tools for analysis of educational test and psychological scale data. In the simplest case these data are comprised of a sample of subjects responding dichotomously to a set of test or scale items. Interest is in estimation of characteristics of the items and subjects. These methods were largely developed in the 1960s through 1980s, though, as Bock [1997] notes in his brief historical review of IRT, the seeds for these models began with Thurstone in the 1920s [Thurstone, 1925, 1926, 1927]. A seminal reference on IRT is the book by Lord and Novick [1968], and in particular the chapters written by Birnbaum in this book. More recent texts and collections include Embretson and Reise [2000], Hambleton et al. [1991], Heinen [1996], Hulin et al. [1983], Lord [1980], van der Linden and Hambleton [1997].

Prior to the development of IRT, classical test theory [Spearman, 1904, Novick, 1966] was used to estimate an individual's score on a test. IRT models overcome several limitations of classical test theory for analysis of such data. In classical test theory, the test was considered the unit of analysis, whereas IRT analysis focuses on the test item. A major challenge for classical test theory is how to score individuals who complete different versions of a test. By focusing on the item as the unit of analysis, IRT effectively solved this problem.

Whereas IRT was originally developed for dichotomous items, extensions for polytomous items, ordinal [Samejima, 1969] and nominal [Bock, 1972] for instance,

soon emerged. Thissen and Steinberg [1986] present a taxonomy of many ordinal and nominal IRT models, succinctly describing the various ways in which these models relate to each other. A similar synthesis of IRT models for polytomous items can be found in Mellenbergh [1995]. Also, the collection by van der Linden and Hambleton [1997] contains several articles describing IRT models for polytomous items.

In its original form, an IRT model assumes that an individual's score on the test is a unidimensional latent "ability" variable or trait, often denoted as θ. Of course, for psychological scales this latent variable might be better labeled as "mood" or "severity," depending on what the scale is intended to measure. Some examples of using IRT with psychological scales can be found in Thissen and Steinberg [1988] and Schaeffer [1988]. This latent variable is akin to a factor in a factor analysis model for continuous variables, and so IRT and factor analysis models are also very much related; a useful source for this is Bartholomew and Knott [1999]. Because a test or scale can measure more than one latent factor, multidimensional IRT models have also been developed [Bock et al., 1988, Gibbons and Hedeker, 1992]; the collection by van der Linden and Hambleton [1997] presents some of these developments.

Typically, a sample of N subjects respond to n_i items at one occasion, though the amount of actual time that one occasion represents can vary. In some situations, subjects are assessed at multiple occasions, perhaps with the same set or a related set of items. Several articles have described applications and developments of IRT modeling to such longitudinal situations in psychology [Adams et al., 1997a,

Anderson, 1985, Embretson, 1991, 2000, Fischer and Pononcy, 1994].

In this chapter the basic IRT model for dichotomous items will be described. It will be shown how this model can be viewed as a mixed model for dichotomous data, which can be fit using standard software. Illustration of IRT modeling will be done using Ecological Momentary Assessment (EMA) data from a study of adolescent smoking. Our example will be a somewhat non-traditional IRT illustration, in the sense that we will not examine responses to test or questionnaire items, per se. Instead, as described more fully later, we will examine whether a subject records a smoking report (that is, smokes a cigarette) in specific time periods defined by day-of-week and hour-of-day.

The basic premise that we are examining with these data is that patterns of smoking behavior, over different times of day and days of the week, may be early markers of the development of nicotine dependence. In all, we will examine thirty-five time periods based on data collected over one week (seven days crossed with five time-of-day intervals). Thus, our "items" are these thirty-five time periods, and our dichotomous response is whether or not a subject smokes, which is ascertained for each of these periods. A subject's latent "ability" can then be construed as their underlying level of smoking during this one-week reporting period, and our interest is to see how this behavior relates to these day-of-week and time-of-day periods. Our aim is to show how IRT models can be applied to analysis of intensive longitudinal data to address key research questions. Specifically, we hypothesized that both early

morning and mid-week smoking would be key determinants of the development of dependence or smoking level.

2 IRT model

To set notation, let Y_{ij} be the dichtomous response of subject i ($i = 1, 2, \ldots N$ subjects) to item j ($j = 1, 2, \ldots n_i$ items). Note that subject i is measured on n_i items, so we don't necessarily assume that all subjects are measured on all items. Also, although the notation might imply that items are nested within subjects, in a typical testing situation it is more common for subjects and items to be crossed, because all N subjects are given the same n items. An exception to this is in computerized adaptive testing where the items that a subject receives are selectively chosen from a large pool of potential items based on their sequential item responses. Here, we will simply assume that there is a set of n items in total, but that not all subjects necessarily respond to all of these items.

Denote a "correct" or positive response as $Y_{ij} = 1$ and an "incorrect" or negative response as $Y_{ij} = 0$. A popular IRT model is the one-parameter logistic model, which is commonly referred to as the Rasch model [Rasch, 1960, Wright, 1977, Thissen, 1982]. This model specifies the probability of a correct response to item j ($Y_j = 1$) conditional on the "ability" of subject i (θ_i) as

$$P(Y_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-a(\theta_i - b_j)]} \tag{1}$$

The parameter θ_i denotes the level of the latent trait for subject i; higher values reflect a higher level on the trait being measured by the items. Trait values are usually assumed to be normally distributed in the population of subjects with mean zero and variance one.

The parameter b_j is the item difficulty, which determines the position of the logistic curve along the ability scale. The further the curve is to the right, the more difficult the item. The parameter a is the slope or discriminating parameter, it represents the degree to which the item response varies with ability θ. In the Rasch model all items are assumed to have the same slope, and so a does not carry the j subscript in this model. In some representations of the Rasch model the common slope parameter a does not explicitly appear in the model. Notice that the function $\{1 + \exp[-(z)]\}^{-1}$ is simply the cumulative distribution function (cdf) for the logistic distribution. Thus, if a subject's trait level θ_i exceeds the difficulty of the item b_j then the probability of a correct response is greater than .5 (similarly if $\theta_i < b_j$ the probability is less than .5 of a correct response).

Figure 1 provides an illustration of the model with three items.

Insert Figure 1 about here

In this figure the values of b_j have been set equal to -1, 0, and 1 to represent a relatively easy, moderate, and difficult item, respectively. Also, the value of a equals unity for these curves. As can be seen, the item difficulty b is the trait level

where the probability of a correct response is .5. So, for example, the first item is one that does not require a high ability level (*i.e.*, -1) to yield a 50:50 chance of getting the item correct. For a test to be an effective measurement tool it is useful to have items with a range of difficulty levels.

As noted, the Rasch model assumes that all items are equally discriminating. To relax this assumption, several authors described various two-parameter models that included both item difficulty and discrimination parameters [Richardson, 1936, Lawley, 1943, Birnbaum, 1968, Bock and Aitkin, 1981]. Here, consider the two-parameter logistic model that specifies the probability of a correct response to item j ($Y_j = 1$) conditional on the ability of subject i (θ_i) as

$$P(Y_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{2}$$

where a_j is the slope parameter for item j, and b_j is the difficulty parameter for item j. Figure 2 provides an illustration of the model with three items.

Insert Figure 2 about here

In this figure the values of a_j have been set equal to .75, 1, and 1.25, respectively. As can be seen, the lower the value of a the lower the slope of the logistic curve and the less the item discriminates levels of the trait. In an extreme case, if $a = 0$, the slope is horizontal and the item is not able to discriminate any levels of the trait. These item discrimination parameters are akin to factor loadings in a factor analysis

model. They indicate the degree to which an item "loads" on the latent trait θ.

As noted by Bock and Aitkin [1981], it is convenient to represent the two-parameter model as

$$P(Y_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-(c_j + a_j\theta_i)]} \ , \tag{3}$$

where $c_j = -a_j b_j$ represents the item-intercept parameter. Similarly, the Rasch model can be written as

$$P(Y_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp[-(c_j + a\,\theta_i)]} \ , \tag{4}$$

with $c_j = -a\,b_j$. In this form, it is easy to see that these models are simply variants of mixed-effects logistic regression models. For example, the Rasch model written in terms of the log odds, or logit, of response is

$$\log\left[\frac{P(Y_{ij} = 1 \mid \theta_i)}{1 - P(y_{ij} = 1 \mid \theta_i)}\right] = c_j + a\,\theta_i \ . \tag{5}$$

2.1 IRT in mixed model form

These IRT models can also be represented using notation that is more common in mixed-effects regression models. For this, let $\boldsymbol{\lambda}_i$ represent the $n_i \times 1$ vector of logits for subject i. The Rasch model can then be written as

$$\boldsymbol{\lambda}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{1}_i\sigma_v\theta_i \tag{6}$$

where $\boldsymbol{X}_i$ is an $n_i \times n$ item indicator matrix obtained from $\boldsymbol{I}_n$ (*i.e.*, the $n \times n$ identity matrix), $\boldsymbol{\beta}$ is the $n \times 1$ vector of item difficulty parameters (*i.e.*, the b_j parameters in the IRT notation), $\mathbf{1}_i$ is a $n_i \times 1$ vector of ones, θ_i is the latent trait (*i.e.*, random effect) value of subject i, which is distributed $\mathcal{N}(0,1)$. The parameter σ_υ is the common slope parameter (*i.e.*, the a parameter in the IRT notation) which indicates the heterogeneity of the random subject effects. Note that it is common to write the random-intercepts mixed logistic model as

$$\boldsymbol{\lambda}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{1}_i\upsilon_i \tag{7}$$

where the random subject effects υ_i are distributed in the population of subjects as $N(0,\sigma_\upsilon^2)$. This is equivalent to the above since $\theta_i = \upsilon_i/\sigma_\upsilon$.

Mixed models are often described as multilevel [Goldstein, 1995] or hierarchical linear models (HLM; [Raudenbush and Bryk, 2002]). IRT models can also be thought of in this way by considering the level-1 item responses as nested within the level-2 subjects. The usual IRT notation is a bit different from the usual multilevel or HLM notation, but the Rasch model is simply a random intercept logistic regression model with item indicator variables as level-1 regressors. As such it can be estimated by several software programs for multilevel or mixed modeling (*e.g.*, SAS PROC NLMIXED, Stata [StataCorp, 1999], HLM [Raudenbush et al., 2004], MLwiN [Rasbash et al., 2004], EGRET [Corcoran et al., 1999], LIMDEP [Greene, 1998], GLLAMM [Rabe-Hesketh et al., 2001], Mplus [Muthén and Muthén, 2001],

and MIXOR [Hedeker and Gibbons, 1996]). This is in addition to the many software programs that are specifically implemented for IRT analysis (*e.g.*, BILOG [Mislevy and Bock, 1998], MULTILOG [Thissen, 1991], NOHARM [Fraser, 1988], and ConQuest [Wu et al., 1998]).

Just to be entirely explicit, because we are not assuming that all subjects provide responses to all n items, the n_i rows of the item indicator matrix $\boldsymbol{X}_i$ are obtained from $\boldsymbol{I}_n$ depending on which items subject i responded to. For example, if $n = 3$ and subject i answered items 1 and 3, then $\boldsymbol{X}_i$ equals

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Thus, the number of rows in $\boldsymbol{X}_i$ is simply the number of items for that subject. Likewise for the vectors $\boldsymbol{\lambda}_i$ and $\mathbf{1}_i$. The dimension of the parameter vector $\boldsymbol{\beta}$ is equal to the total number of items.

The two-parameter model can also be represented in matrix form as

$$\boldsymbol{\lambda}_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{X}_i'\boldsymbol{T}\theta_i \ , \tag{9}$$

where $\boldsymbol{T}$ is a vector of standard deviations,

$$\boldsymbol{T}' = [\sigma_{\upsilon_1} \ \ \sigma_{\upsilon_2} \ \ \ldots \ \ \sigma_{\upsilon_n}] \ \ . \tag{10}$$

These standard deviations correspond to the discrimination parameters of the IRT model (the Rasch model is simply a special case of this model, where these standard

deviations are all the same). As can be seen, the two-parameter model is a mixed-effects logistic regression model that allows the random effect variance terms to vary across the items at the first level. Goldstein [1995] refers to such multilevel models as having complex variance structure. Several of the aforementioned mixed or multilevel software programs can fit such models, and so can fit two-parameter IRT models.

An aspect that can be confusing concerns the different parameterizations used in IRT and mixed models. In the two-parameter IRT model, the following parameterizations are typically used (see Bock and Aitkin [1981]):

$$\sum_{j=1}^{n} b_j = 0 \quad \text{and} \quad \prod_{j=1}^{n} a_j = 1 \ . \tag{11}$$

These parameterizations center the item difficulty parameters around zero, and multiplicatively center the item discrimination parameters around one. As a result, mixed model estimates need to be rescaled and/or standardized to be consistent with IRT results. To illustrate this, we consider the oft-analyzed LSAT section 6 data published in Thissen [1982]. This dataset consists of responses from 1000 subjects to five dichotomous items from section 6 of the LSAT exam. In his article, Thissen published results for the Rasch model applied to these data. These results are presented in Table 1 (labeled as IRT). We analyzed these data using a random-intercept logistic model with item indicators as fixed regressors using MIXOR and SAS PROC NLMIXED (see the Appendix for the SAS PROC NLMIXED code).

Both programs gave near-identical results; the NLMIXED estimates are listed in Table 1 both in the raw and translated forms.

Insert Table 1 about here

To get to the IRT formulation, first the sign of the mixed model difficulty parameter estimates is reversed, and then the mean of the estimates is subtracted from each, namely,

$$b_j = -\beta_j - \frac{1}{n}\sum_{j'=1}^{n} -\beta_{j'} \quad .$$

This ensures that the mean of these transformed estimates equals zero. As can be seen, this yields item difficulty estimates nearly identical to those reported in Thissen [1982]. In either representation it is clear that relatively many subjects got item 1 correct and relatively fewer got item 3 correct. This is because in the mixed model (using the formulation described herein) a positive regression coefficient indicates increased probability of response (*i.e.*, a correct response) with increasing values of the regressor, whereas in the IRT a negative difficulty indicates an easier item (*i.e.*, more subjects got the item correct).

Transitioning between the mixed and two-parameter model estimates is slightly more complicated. Bock and Aitkin [1981] list results for a two-parameter probit analysis of these same data; these are given in Table 2 (labeled as IRT). This same model was estimated using MIXOR and SAS PROC NLMIXED with similar results.

Table 2 lists those from MIXOR, the appendix provides the PROC NLMIXED code for this run.

Insert Table 2 about here

Here are the steps necessary to go from the mixed to the IRT model estimates.

1. Transform σ_j estimates so that their product equals 1. This yields the a_j estimates, and can be done by:

$$a_j = \exp\left[\log \sigma_j - \frac{1}{n}\sum_{j'=1}^{n} \log \sigma_{j'}\right] \quad .$$

2. Reverse the sign of the β_j estimates, and transform so that $\sum -\beta_j/a_j = 0$, using the standardized a_j estimates from the previous step.

$$b_j = -(\beta_j/a_j) - \frac{1}{n}\sum_{j'=1}^{n}(-\beta_{j'}/a_{j'}) \quad .$$

As Table 2 shows, the IRT and translated mixed model estimates agree closely. Also, as can be seen from the discrimination parameter estimates, items 3 and 5 are the most and least discriminating items, respectively. Item 3 is therefore both a difficult and discriminating item.

Unlike traditional IRT models, the mixed or multilevel model formulation easily allows multiple covariates at either level (*i.e.*, items or subjects). This and other advantages of casting IRT models as multilevel models are described by Adams et al. [1997b] and Reise [2000]. In particular, multilevel models are well-suited for

examining whether item parameters vary by subject characteristics, and also for estimating ability in the presence of such item by subject interactions. Interactions between item parameters and subject characteristics, often termed item bias [Camilli and Shepard, 1994], is an area of active psychometric research.

2.2 Generalized Linear Mixed Models

IRT models have connections with the general class of models known as generalized linear mixed models (GLMMs; [McCulloch and Searle, 2001]). GLMMs extend generalized linear models (GLMs) by inclusion of random effects, and are commonly used for analysis of correlated non-normal data. An excellent and comprehensive source on GLMM applications is the text by Skrondal and Rabe-Hesketh [2004]; a more condensed review can by found in Hedeker [2005].

There are three specifications in a GLMM. First, the linear predictor, denoted as η_{ij} (as before, i and j represent subjects and items, respectively), of a GLMM is of the form

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\upsilon}_i \ , \tag{12}$$

where $\boldsymbol{x}_{ij}$ is the vector of regressors for unit ij with fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{z}_{ij}$ is the vector of variables having random effects which are denoted $\boldsymbol{\upsilon}_i$. The random effects are usually assumed to be multivariate normally distributed, that is, $\boldsymbol{\upsilon}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_\upsilon)$. Note that they could be expressed in standardized form (as is typical in

IRT models) as $\boldsymbol{\theta}_i \sim (\mathbf{0}, \boldsymbol{I})$, where $\boldsymbol{\upsilon}_i = \boldsymbol{T}\boldsymbol{\theta}_i$ and $\boldsymbol{T}$ is the Cholesky factor of the variance covariance matrix $\boldsymbol{\Sigma}_\upsilon$, namely $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{\Sigma}_\upsilon$. Here, we've also allowed for multiple random effects, whereas the IRT models considered in this article have been in terms of a single random effect.

The second specification of a GLMM is the selection of a link function $g(\cdot)$ which converts the expected value μ_{ij} of the outcome variable Y_{ij} to the linear predictor η_{ij}

$$g(\mu_{ij}) = \eta_{ij} \; . \tag{13}$$

Here, the expected value of the outcome is conditional on the random effects (*i.e.*, $\mu_{ij} = E[Y_{ij} \mid \boldsymbol{\upsilon}_i]$). Finally, a specification for the form of the variance in terms of the mean μ_{ij} is made. These latter two specifications usually depend on the distribution of the outcome Y_{ij}, which is assumed to fall within the exponential family of distributions.

The Rasch model is seen as a GLMM by specifying item indicator variables as the vector of regressors $\boldsymbol{x}_{ij}$, and by setting $\boldsymbol{z}_{ij} = 1$ and $\boldsymbol{\upsilon}_i = \upsilon_i$ for the random effects part. The link function is specified as the logit link, namely

$$g(\mu_{ij}) \; = \; \text{logit}(\mu_{ij}) \; = \; \log\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] \; = \; \eta_{ij} \; . \tag{14}$$

The conditional expectation $\mu_{ij} = E(Y_{ij} \mid \boldsymbol{\upsilon}_i)$ equals $P(Y_{ij} = 1 \mid \boldsymbol{\upsilon}_i)$, namely, the conditional probability of a response given the random effects. Since Y is dichoto-

mous, the variance function is simply $\mu_{ij}(1 - \mu_{ij})$.

Rijmen et al. [2003] present an informative overview and bridge between IRT models, multilevel models, mixed models, and GLMMs. As they point out, the Rasch model, and variants of it, belong to the class of GLMMs. However, the more extended two-parameter model is not within the class of GLMMs because the predictor is no longer linear, but includes a product of parameters.

3 Estimation

Parameter estimation for IRT models and GLMMs typically involves maximum likelihood (ML) or variants of ML. Additionally, the solutions are usually iterative ones that can be numerically quite intensive. Here, the solution is merely sketched; further details can be found in McCulloch and Searle [2001] and Fahrmeir and Tutz [2001]. Let $\mathbf{Y}_i$ denote the vector of responses from subject i. The probability of any response pattern $\mathbf{Y}_i$ (of size n_i), conditional on $\boldsymbol{\upsilon}$, is equal to the product of the probabilities of the level-1 responses:

$$\ell(\mathbf{Y}_i \mid \boldsymbol{\upsilon}_i) = \prod_{i=1}^{n_i} P(Y_{ij} = 1 \mid \boldsymbol{\upsilon}_i) \; . \tag{15}$$

The assumption that a subject's responses are independent given the random effects (and therefore can be multiplied to yield the conditional probability of the response vector) is known as the *conditional independence* assumption. The marginal density of $\mathbf{Y}_i$ in the population is expressed as the following integral of the conditional

likelihood $\ell(\cdot)$

$$h(\mathbf{Y}_i) = \int_{\boldsymbol{\upsilon}} \ell(\mathbf{Y}_i \mid \boldsymbol{\upsilon}_i)\, f(\boldsymbol{\upsilon})\, d\boldsymbol{\upsilon}\ , \tag{16}$$

where $f(\boldsymbol{\upsilon})$ represents the distribution of the random effects, often assumed to be a multivariate normal density. Whereas (15) represents the conditional probability, (16) indicates the unconditional probability for the response vector of subject i. The marginal log-likelihood from the sample of N subjects is then obtained as $\log L = \sum_i^N \log h(\mathbf{Y}_i)$. Maximizing this log-likelihood yields ML estimates (which are sometimes referred to as maximum marginal likelihood estimates) of the regression coefficients $\boldsymbol{\beta}$ and the variance-covariance matrix of the random effects $\boldsymbol{\Sigma}_\upsilon$ (or the Cholesky of this matrix, denoted $\boldsymbol{T}$).

3.1 Integration over the random-effects distribution

In order to solve the likelihood solution, integration over the random-effects distribution must be performed. As a result, estimation is much more complicated than in models for continuous normally-distributed outcomes where the solution can be expressed in closed form. Various approximations for evaluating the integral over the random-effects distribution have been proposed in the literature; many of these are reviewed in Rodríguez and Goldman [1995]. Perhaps the most frequently used methods are based on first- or second-order Taylor expansions. Marginal quasi-likelihood (MQL) involves expansion around the fixed part of the model, whereas penalized

or predictive quasi-likelihood (PQL) additionally includes the random part in its expansion [Goldstein and Rasbash, 1996]. Unfortunately, these procedures yield estimates of the regression coefficients and random effects variances that are biased towards zero in certain situations, especially for the first-order expansions [Breslow and Lin, 1995]. To remedy this, Raudenbush et al. [2000] proposed an approach that uses a combination of a fully multivariate Taylor expansion and a Laplace approximation. This method yields accurate results and is computationally fast. Also, as opposed to the MQL and PQL approximations, the deviance obtained from this approximation can be used for likelihood-ratio tests.

Numerical integration can also be used to perform the integration over the random-effects distribution [Bock and Lieberman, 1970]. Specifically, if the assumed distribution is normal, Gauss-Hermite quadrature can approximate the above integral to any practical degree of accuracy. Additionally, like the Laplace approximation, the numerical quadrature approach yields a deviance that can be readily used for likelihood-ratio tests. The integration is approximated by a summation on a specified number of quadrature points for each dimension of the integration. An issue with the quadrature approach is that it can involve summation over a large number of points, especially as the number of random-effects is increased. To address this, methods of adaptive quadrature have been developed that use a few number of points per dimension that are adapted to the location and dispersion of the distribution to be integrated [Rabe-Hesketh et al., 2002].

More computer-intensive methods, involving iterative simulations, can also be used to approximate the integration over the random effects distribution. Such methods fall under the rubric of Markov chain Monte Carlo (MCMC; [Gilks et al., 1997]) algorithms. Use of MCMC for estimation of a wide variety of models has exploded in the last ten years or so. MCMC solutions are described in Patz and Junker [1999] and Kim [2001] for IRT models, and in Clayton [1996] for GLMMs.

3.2 Estimation of random effects

In many cases, it is useful to obtain estimates of the random effects. This is particularly true in IRT, where estimation of the random effect or latent trait is of primary importance, since these are the test ability scores for individuals. The random effects $\boldsymbol{\upsilon}_i$ can be estimated using empirical Bayes methods. For the univariate case, this estimator υ_i is given by:

$$\hat{\upsilon}_i = E(\upsilon_i \mid \mathbf{Y}_i) = h_i^{-1} \int_{\upsilon} \upsilon_i \, \ell_i \, f(\upsilon) \, d\upsilon \, , \tag{17}$$

where ℓ_i is the conditional probability for subject i under the particular model and h_i is the analogous marginal probability. This is simply the mean of the posterior distribution. Similarly, the variance of the posterior distribution is obtained as

$$V(\hat{\upsilon}_i \mid \mathbf{Y}_i) = h_i^{-1} \int_{\upsilon} (\upsilon_i - \hat{\upsilon}_i)^2 \, \ell_i \, f(\upsilon) \, d\upsilon \, . \tag{18}$$

These quantities may then be used, for example, to evaluate the response probabil-

ities for particular subjects, or for ranking subjects in terms of their abilities. Embretson [2000] thoroughly describe uses and properties of these estimators within the IRT context; additionally, the edited collection of Thissen and Wainer [2001] contains a wealth of material on this topic.

4 Illustration: Adolescent Smoking Study

Data for the analyses reported here come from a longitudinal, natural history study of adolescent smoking [Mermelstein et al., in press]. Students included in the longitudinal study were either in 8th or 10th grade at baseline, and either had never smoked, but indicated a probability of future smoking, or had smoked in the past 90 days, but had not smoked more than 100 cigarettes in their lifetime. Written parental consent and student assent were required for participation. A total of 562 students completed the baseline measurement wave. The longitudinal study utilized a multi-method approach to assess adolescents at three time points: baseline, six and twelve months. The data collection modalities included self-report questionnaires, a week-long time/event sampling method via palmtop computers (Ecological Momentary Assessments), and in-depth interviews.

Data for the current analyses came from the ecological momentary assessments obtained during the baseline timepoint. Adolescents carried the hand held computers with them at all times during the one-week data collection period and were trained both to respond to random prompts from the computers and to event record

(initiate a data collection interview) smoking episodes. Immediately after smoking a cigarette, participants completed a series of questions on the hand held computers. Questions pertained to place, activity, companionship, mood, and specifics about smoking access and topography. The hand held computers date and time-stamped each entry. For inclusion in the analyses reported here, adolescents must have smoked at least one cigarette during the seven-day baseline data collection period; 152 adolescents met this inclusion criterion. In the analyses reported here, we will focus on the timing, both in terms of time of day and day of week, of the smoking reports. Thus, we are not analyzing data from the random prompts, but only from the self-initiated smoking reports.

The assessment week was divided into 35 periods, and whether an individual provided a smoking report in each of these 35 periods was determined as follows. First, we recorded the day of the week for a given cigarette report. Then, for each day, the time of day was classified into five bins: 3am to 8:59am, 9am to 1:59pm, 2pm to 5:59pm, 6pm to 9:59pm, and 10pm to 2:59am. Technically, part of the last bin (*i.e.*, the time past midnight) belonged to the next day, but we treated this entire period as a late night bin. In this way, subjects who, for example, smoked a cigarette at 1am during a Saturday night party, would be classified as having smoked the cigarette on Saturday night and not on Sunday morning. Thus, for each of these 152 subjects, we created a 35×1 dichotomous response vector representing whether or not a cigarette report was made (yes/no) for each of these 35 time periods in the

week. Some subjects did report more than one smoking event during a given time period, however this was relatively rare and so we simply treated the outcome as dichotomous (no report versus one or more smoking reports).

In terms of the IRT vernacular our "items" are these 35 time periods, and our dichotomous response is whether or not the subjects recorded smoking in these periods. A subject's latent "ability" can then be construed as their underlying degree of smoking behavior, and our interest is to see how this behavior relates to these day-of-week and time-of-day periods. An item's "difficulty" indicates the relative frequency of smoking reports during the time period, and an item's "discrimination" refers to the degree to which the time period distinguishes levels of the latent smoking behavior variable. The discrimination parameters are akin to factor loadings in a factor analysis or structural equation model. In this light, Rodriguez and Audrain-McGovern [2004] describe a factor analysis of smoking-related sensations among adolescents, while a Rasch analysis of smoking behaviors among more advanced smokers is presented in Bretelera et al. [2004]. Admittedly, our example is a bit different in that we are examining instances of smoking behavior, rather than smoking-related sensations and/or behaviors, however we feel that the IRT approach we describe addresses some very interesting smoking-related questions. Specifically, we will examine the degree to which time periods relate to the occurence of smoking behavior, and the degree to which these periods differentially distinguish subjects of varying underlying smoking behavior levels. Our hypothesis is that time- and

day-related characteristics of smoking among these beginning smokers may serve as early behavioral markers of dependence development.

Table 3 lists the proportion and number of smoking reports for each of these 35 time periods. Note that each of these proportions is based on the 152 subjects. In other words, each represents the proportion of the 152 subjects who smoked during the particular time period.

Insert Table 3 about here

As can be seen, the later week and weekend days represent relatively popular smoking days. In terms of time periods, the late afternoon and evening hours are also more common times of smoking reports. Of the 152 subjects, approximately 41%, 24%, 14%, 15%, and 6% provided one, two, three or four, five through nine, and ten or more smoking reports, respectively.

Because the data are rather sparse, it did not seem reasonable to estimate separate item parameters for each of the thirty-five cells produced by the crossing of day of week and time of day. Instead, we chose a "main effects" type of analyses in which the item parameters varied across these two factors but not their interaction. In terms of item difficulty (the location parameter), all models included effects for time of day and day of week. The referent cell was selected to be Monday from 10pm to 3am, since this was thought to be a time of low smoking reports. In the context of the present example, item difficulty primarily refers to the degree to which

smoking reports were made during the particular time periods. In other words, the difficulty estimates reflect the proportion of reports made during the time intervals. For item discrimination (or item slopes), models were sequentially fit assuming these slopes were equal (*i.e.*, a Rasch model), or varied by time of day, day of week, or both time of day and day of week. Again, in the context of the present example, item discrimination refers to the degree to which a given time period distinguishes subjects with varying levels of underlying smoking behavior.

The results of these analysis are presented in Tables 4 and 5. Table 4 lists the estimated difficulty parameters under these four models, and Table 5 lists the corresponding item discrimination parameters. These estimates correspond to the mixed model representation of the parameters.

Insert Tables 4 and 5 about here

Note that, in Table 5, for the Rasch model (first column) a single common discrimination parameter is estimated, while in the next two models separate discrimination parameters were estimated for time of the day (second column) or each day of the week (third column). In the last of these models (column four), separate discrimination parameters were estimated for each day of week and time of day period under the constraint that the Saturday and 10pm to 3am estimates were equal. This equality was chosen since the estimates for these two periods were most similar in the time-varying and day-varying models.

To choose between these models, Table 6 presents model fit statistics, including the model deviance ($-2 \log L$), AIC [Akaike, 1973], and BIC [Read and Cressie, 1988]. Lower values on these fit statistics imply better models. As can be seen, AIC would suggest the fourth model as best, whereas BIC points to the Rasch or random-intercepts model. As noted by Rost [1997] BIC penalizes overparameterization more than AIC, and so the results here are not too surprising. Because these are nested models, likelihood-ratio tests can also be used; these are presented in the bottom half of Table 6. The likelihood-ratio tests support the model with day- and time-varying slopes (*i.e.*, discrimination parameters) over the three simpler models. Thus, there is evidence that the day of week and time of day are differentially discriminating in relating to adolescents' smoking "ability" or "dependence."

The difficulty estimates in Table 4 support the notion that adolescent smoking is least frequent on Sunday and Monday, increases during the mid-week Tuesday to Thursday period, and highest on Friday and Saturday. All of the day-of-week indicators, with Monday as the reference cell, are statistically significant, based on Wald statistics, except for Sunday in all models. Figure 3 presents the difficulty estimates implied by the final model for the seven days and five time periods. In this figure, the estimates are presented using the IRT parameterization in Equation (11). Thus, lower values reflect "easier" items, or time periods where smoking behavior is relatively more common.

Insert Figure 3 about here

Turning to the time of day indicators, where 10pm to 3am is the reference cell, it is seen that more frequent smoking periods are 2pm to 6pm and 6pm to 10pm, whereas 3am to 9am is the least frequent period for smoking. The 9am to 2pm period is intermediate and statistically similar to the 10pm to 3am reference cell.

The results for the difficulty parameters agree with the observed frequency data presented in Table 3, reflecting diffferences in the frequency of smoking reports. Alternatively, the discrimination parameter estimates reveal the weighting of the days and time periods on the latent smoking level and so can suggest aspects of the data that are not so obvious. For this, Figure 4 displays the discrimination, or slope, estimates based on the final model for the seven days and five time periods. Again, these are displayed in the IRT representation of Equation (11).

Insert Figure 4 about here

These indicate that the most discriminating days are Monday, Wednesday, and Thursday, and the least discriminating days are Friday, Saturday, Sunday, and Tuesday. This suggests, in combination with the difficulty results, that although weekend smoking is relatively prevalent it is not as informative as weekday smoking in determining the level of an adolescent's smoking behavior. In terms of time of day, these estimates and their characterization in Figure 4 clearly point out the time period of 3am to 9am as the most discriminating period. This result agrees well with

the literature on smoking, because morning smoking and smoking after awakening are important markers of smoking addiction [Heatherton et al., 1991]. Interestingly, comparing Figures 3 and 4, one can see that this discriminating time period is one of very infrequent smoking for adolescents.

Another fundamental aspect of IRT modeling is the estimation of each individual's level of ability θ. In achievement testing, these indicate the relative abilities of the sample of testees. Here, these reflect the latent smoking level of the adolescents in this study. As mentioned, estimation of θ is typically done using empirical Bayes methods, whereas estimation of the item parameters is based on maximum likelihood [Bock and Aitkin, 1981]. This combination of empirical Bayes and maximum likelihood is also the usual procedure in mixed models [Laird and Ware, 1982]. The distribution of ability estimates, based on the day and time varying slopes model, is presented in Figure 5.

Insert Figure 5 about here

This plot indicates that many individuals have a low level of smoking, though not all. Given the observed frequencies in Table 3, this pattern of results for θ estimates is not surprising. Many subjects provided only one or two smoking reports, and therefore would certainly have low levels on the latent smoking variable. A concern is that the distribution is assumed to be normal in the population, however this figure indicates a positively skewed distribution. IRT methods that do not as-

sume normality for the underlying distribution of ability are described in Mislevy [1984], who indicates how non-parametric and mixture ability distributions can be estimated in IRT models. Such extensions are also described and further developed within a mixed model framework by Carlin et al. [2001]. In the present case, we considered a non-parametric representation and estimated the density at a finite number of points from approximiately -4 to 4, in addition to estimation of the usual item parameters. These results, not shown, gave similar conclusions in terms of the item difficulty and discrimination parameters. Thus, our conclusions seem relatively robust to the distributional assumption for ability.

A natural question to ask is the relationship between the IRT estimate of smoking level and the simple sum of the number of smoking reports an individual provides. Figure 6 presents a scatterplot of the number of smoking reports versus the latent smoking ability estimates, based on the day and time varying slopes model.

Insert Figure 6 about here

These two are, of course, highly correlated, $r = .96$, but the IRT estimate is potentially more informative because it includes information about when the smoking events were reported. To give a sense of this, Table 7 lists the minimum and maximum IRT ability estimate, stratified by the number of smoking reports, for subjects with 2, 3, 4, or 5 reports. For each, the day of the week and the time of the day are indicated.

Insert Table 6 about here

As can be seen, within each strata, the lower IRT estimates are more associated with what might be considering "party" smoking, namely smoking on weekend evenings and nights. Conversely, higher IRT estimates are more associated with weekday and morning smoking reports. Again, this agrees with notions in smoking research that smoking alone or outside of a social event may be more characteristic of the development of smoking dependency [Nichter et al., 2002].

5 Discussion

IRT models have been extensively developed and used in educational and psychological measurement. However, use of IRT models outside of these areas is rather limited. Part of the reason for this is that these methods have not been well understood by non-psychometricians. With the emergence and growing popularity of mixed models for longitudinal data, formulation of IRT models within this class can help to overcome these obstacles.

In this chapter, we have strived to show the connection between IRT and mixed models, and how standard software for mixed models can be used to estimate basic IRT models. Some translation of the parameter estimates is necessary to properly express the mixed model results in IRT form, and we have illustrated how this is done using an oft-analysed dataset of LSAT test items.

IRT modeling was illustrated using Ecological Momentary Assessment (EMA) data from a study of adolescent smoking. Here, we analyzed whether or not a smoking report had been made in each of 35 time periods, defined by the crossing of seven days and five time intervals within each day. The IRT model was able to identify which time periods were most associated with smoking reports; and also which time periods were the most informative, in the sense of discriminating underlying levels of smoking behavior. As indicated, weekend and evening hours yielded the most frequent smoking reports, however morning and, to some extent, mid-week reports were most discriminating in separating smoking levels.

IRT modeling provides a useful method for addressing questions about patterning of behavior beyond mere frequency reports. For example, in the case presented here, we examined whether the time of day or day of week that an adolescent smokes discriminates underlying levels of smoking behavior. This is not possible if one only considers smoking quantity alone. Although our data indicate that adolescents are more likely to smoke on weekend evenings (the stereotypic weekend party phenomena), the data also indicate that those smoking events may be a less important indicator of the underlying level of smoking behavior than smoking episodes that occur either mid-week or in the early mornings. Thus, one might consider mid-week smoking as an indicator of a more pronounced level of adolescent smoking behavior, and perhaps as an indicator of the subsequent development of smoking dependence. Indeed, these data also lead one to address questions such as whether "binge" smok-

ing episodes are less associated with underlying smoking behavior than is a pattern of smoking that is more evenly distributed over the week. The IRT models presented here are clearly useful in furthering the empirical investigations of a number of behavioral phenomena for which both quantity and patterns of behavior may be important. For example, one could easily apply similar models to addressing questions about patterns of lapses and relapses following abstinence as well as to models of escalation.

In addition, other applications of IRT modeling of EMA data are clearly possible. For example, in this chapter we have focused on dichotomous data, but IRT models for ordinal and nominal outcomes have also been developed. Additionally, we did not include any covariates in our analsyes, but these could easily be handled. Thus, we could explore whether the item parameters differed between males and females, or whether the number of friends present during a given time period influenced the probability of a smoking report.

Appendix

SAS PROC NLMIXED can be used to perform IRT analysis, and code is given below for analysis of the LSAT-6 data. This code assumes that the data are at the individual level. An individual identifier must be present in the data and here it is named `id`. The dependent variable is named `lsat6` and coded 1 for a correct response and 0 for an incorrect response. The item indicators are named `item1` to `item5`. For example, the data are as follows for an individual with `id` 1001 who did not get any of the five items correct (`id, lsat6, item1, item2, item3, item4, item5`):

1001	0	1	0	0	0	0
1001	0	0	1	0	0	0
1001	0	0	0	1	0	0
1001	0	0	0	0	1	0
1001	0	0	0	0	0	1

Because the mixed model does not need to assume an equal number of observations per individual, individuals missing a particular item would have less than five lines of data (or have a missing value code for the missed item response). In the LSAT-6 dataset, all of the 1000 subjects had responses on the five items, so the data are complete. Below is the PROC NLMIXED code to estimate, respectively, a Rasch logistic model and a two-parameter probit model. NLMIXED is somewhat slow in running these analyses. As an alternative, one can use the freeware MIXOR

program [Hedeker and Gibbons, 1996]. MIXOR syntax files for these analyses are available at www.uic.edu/∼hedeker/long.html. Additionally, the LSAT-6 dataset can be downloaded from this website.

```
/* Rasch logistic model in mixed regression formulation */
proc nlmixed ;
parms c1=0 c2=0 c3=0 c4=0 c5=0 a=1 ;
z = c1*item1 + c2*item2 + c3*item3 + c4*item4 + c5*item5 + a*theta;
if (lsat6=0) then p = 1 - (1 / (1 + exp(-z)));
else p = 1 / (1 + exp(-z));
if (p > 1e-8) then ll = log(p);
else ll = -1e20;
model lsat6 ~ general(ll);
random theta ~ normal(0,1) subject=id;
run;


/* 2 parameter probit model in mixed regression formulation */
proc nlmixed ;
parms a1=1 a2=1 a3=1 a4=1 a5=1 c1=0 c2=0 c3=0 c4=0 c5=0;
bounds a1>0, a2>0, a3>0, a4>0, a5>0;
z = (c1*item1 + c2*item2 + c3*item3 + c4*item4 + c5*item5) +
```

```
(a1*item1 + a2*item2 + a3*item3 + a4*item4 + a5*item5) * theta;

if (lsat6=0) then p = probnorm(0-z);

else p = probnorm(z) ;

if (p > 1e-8) then ll = log(p);

else ll = -1e20;

model lsat6 ~ general(ll);

random theta ~ normal(0,1) subject=id;

run;
```

Acknowledgements

Thanks are due to Siu Chi Wong for statistical analysis. This work was supported by National Institutes of Mental Health grant MH56146, National Cancer Institute grant CA80266, and by a grant from the Tobacco Etiology Research Network, funded by the Robert Wood Johnson Foundation. Correspondence to Donald Hedeker, Division of Epidemiology & Biostatistics (M/C 923), School of Public Health, University of Illinois at Chicago, 1603 West Taylor Street, Room 955, Chicago, IL, 60612-4336. e-mail: hedeker@uic.edu

References

R. J. Adams, M. Wilson, and W. Wang. The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21:1–23, 1997a.

R. J. Adams, M. Wilson, and M. Wu. Multilevel item response models: an approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, 22:47–76, 1997b.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Academiai Kiado, Budapest, 1973.

E. B. Anderson. Estimating latent correlations between repeated testings. *Psychometrika*, 50:3–16, 1985.

D. J. Bartholomew and M. Knott. *Latent variable models and factor analysis (2nd edition)*. Oxford University Press, New York, 1999.

A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, editors, *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA, 1968.

R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51, 1972.

R. D. Bock. A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16:21–32, 1997.

R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: An application of the em algorithm. *Psychometrika*, 46:443–459, 1981.

R. D. Bock, R. D. Gibbons, and E. Muraki. Full-information item factor analysis. *Applied Psychological Measurement*, 12:261–280, 1988.

R. D. Bock and M. Lieberman. Fitting a response model for n dichotomously scored items. *Psychometrika*, 35:179–197, 1970.

N. E. Breslow and X. Lin. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.

M. H. M. Bretelera, S. R. Hilberinkb, G. Zeemanc, and S. M. M. Lammersa. Compulsive smoking: the development of a Rasch homogeneous scale of nicotine dependence. *Addictive Behaviors*, 29:199–205, 2004.

G. Camilli and L. A. Shepard. *Methods for identifying biased test items.* Sage Publications, Thousand Oaks, CA, 1994.

J. B. Carlin, R. Wolfe, C. H. Brown, and A. Gelman. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*, 2:397–416, 2001.

D. Clayton. Generalized linear mixed models. In W. R. Gilks, S. Richardson, and

D. J. Spiegelhalter, editors, *Markov chain Monte Carlo methods in practice*, pages 275–303. Chapman and Hall, New York, 1996.

C. Corcoran, B. Coull, and A. Patel. *EGRET for Windows user manual*. CYTEL Software Corporation, Cambridge, MA, 1999.

S. E. Embretson. A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56:495–516, 1991.

S. E. Embretson. Multidimensional measurement from dynamic tests: Abstract reasoning under stress. *Multivariate Behavioral Research*, 35:505–542, 2000.

S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. Erlbaum, Mahwah, NJ, 2000.

L. Fahrmeir and G. T. Tutz. *Multivariate statistical modelling based on generalized linear models, 2nd edition*. Springer-Verlag, New York, 2001.

G.H. Fischer and I. Pononcy. An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59:177–192, 1994.

C. Fraser. *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. University of New England, Centre for Behavioral Studies, Armidale, N.S.W., 1988.

R. D. Gibbons and D. Hedeker. Full-information item bi-factor analysis. *Psychometrika*, 57:423–436, 1992.

W. Gilks, S. Richardson, and D.J. Speigelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, New York, 1997.

H. Goldstein. *Multilevel statistical models, 2nd edition*. Halstead Press, New York, 1995.

H. Goldstein and J. Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series B*, 159:505–513, 1996.

W. H. Greene. *LIMDEP version 7.0 user's manual (revised edition)*. Econometric Software, Inc., Plainview, NY, 1998.

R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of item response theory*. Sage, Newbury Park, CA, 1991.

T. F. Heatherton, L. T. Kozlowski, R. C. Frecker, and K. O. Fagerstrom. The Fagerstrom test for nicotine dependence: A revision of the Fagerstrom tolerance questionnaire. *British Journal of Addictions*, 86:1119–1127, 1991.

D. Hedeker. Generalized linear mixed models. In B. Everitt and D. Howell, editors, *Encyclopedia of Statistics for the Behavioral Sciences*. Wiley, London, 2005.

D. Hedeker and R. D. Gibbons. MIXOR: a computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.

T. Heinen. *Latent class and discrete latent trait models.* Sage, Thousand Oaks, CA, 1996.

C. L. Hulin, F. Drasgow, and C. K. Parsons. *Item response theory.* Dow Jones-Irwin, Homewood, IL, 1983.

S.-H. Kim. An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, 25:163–176, 2001.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

D. N. Lawley. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 61:273–287, 1943.

F. M. Lord. *Applications of item response theory to practical testing problems.* Erlbaum, Hillside, NJ, 1980.

F. M. Lord and M. R. Novick. *Statistical theories of mental health scores.* Addison-Wesley, Reading, MA, 1968.

C. E. McCulloch and S. R. Searle. *Generalized, linear, and mixed models.* Wiley, New York, 2001.

G. J. Mellenbergh. Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19:91–100, 1995.

R. Mermelstein, D. Hedeker, B. Flay, and S. Shiffman. Real-time data capture and adolescent cigarette smoking. In A. Stone, S. Shiffman, and Atienza A., editors, *The science of real-time data capture: self-report in health research*. Oxford University Press, New York, in press.

R. J. Mislevy. Estimating latent distributions. *Psychometrika*, 49:359–381, 1984.

R. J. Mislevy and R. D. Bock. *BILOG-3: Item analysis and test scoring with binary logistic models*. Scientific Software International, Inc., Chicago, 1998.

B. Muthén and L. Muthén. *Mplus user's guide*. Muthén & Muthén, Los Angeles, 2001.

M. Nichter, M. Nichter, P. J. Thompson, S. Shiffman, and A. B. Moscicki. Using qualitative research to inform survey development on nicotine dependence among adolescents. *Drug and Alcohol Dependency, Supplement 1*, 68:S41–S56, 2002.

M. R. Novick. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3:1–18, 1966.

R. J. Patz and B. W. Junker. A straightforward approach to Markov chain Monte Carlo methods for item response theory models. *Journal of Educational and Behavioral Statistics*, 24:146–178, 1999.

S. Rabe-Hesketh, A. Pickles, and A. Skrondal. *GLLAMM Manual*. Institute of Psychiatry, King's College, University of London, Technical Report 2001/01, Department of Biostatistics and Computing, 2001.

S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21, 2002.

J. Rasbash, F. Steele, W. Browne, and B. Prosser. *A user's guide to MLwiN version 2.0*. University of London, London: Institute of Education, 2004.

G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen, Denmark, 1960.

S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models in social and behavioral research: Applications and data-analysis methods (2nd edition)*. Sage, Thousand Oaks, CA, 2002.

S. W. Raudenbush, A. S. Bryk, Y. F. Cheong, and R. Congdon. *HLM 6: Hierarchical linear and nonlinear modeling*. Scientific Software International, Inc., Chicago, 2004.

S. W. Raudenbush, M.-L. Yang, and M. Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics*, 9:141–157, 2000.

T. R. C. Read and N. A. C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York, 1988.

S. P. Reise. Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35:543–568, 2000.

M. W. Richardson. The relationship between difficulty and the differential validity of a test. *Psychometrika*, 1:33–49, 1936.

F. Rijmen, F. Tuerlinckx, P. De Boeck, and P. Kuppens. A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8:185–205, 2003.

D. Rodriguez and J. Audrain-McGovern. Construct validity analysis of the early smoking experience questionnaire for adolescents. *Addictive Behaviors*, 29:1053–1057, 2004.

G. Rodríguez and N. Goldman. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 158:73–89, 1995.

J. Rost. Logistic mixture models. In W. J. van der Linden and R. K. Hambleton, editors, *Handbook of modern item response theory*, pages 449–463. Springer, New York, 1997.

F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, 1969.

N. C. Schaeffer. An application of item response theory to the measurement of depression. In C. Clogg, editor, *Sociological Methodology 1988*, pages 271–307. American Sociological Association, Washington, D.C., 1988.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel,*

Longitudinal and Structural Equation Models. Chapman and Hall/CRC, Boca Raton, FL, 2004.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

StataCorp. *Stata: release 6.0*. Stata Corporation, College Station, TXJ, 1999.

D. Thissen. Marginal maximum liklihood estimation for the one-parameter logistic model. *Psychometrika*, 47:175–186, 1982.

D. Thissen. *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Scientific Software International, Inc., Chicago, 1991.

D. Thissen and L. Steinberg. A taxonomy of item response models. *Psychometrika*, 51:567–577, 1986.

D. Thissen and L. Steinberg. Data analysis using item response theory. *Psychological Bulletin*, 104:385–395, 1988.

D. Thissen and H. Wainer. *Test scoring*. Erlbaum, Mahwah, NJ, 2001.

L. L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16:433–451, 1925.

L. L. Thurstone. The scoring of individual performance. *Journal of Educational Psychology*, 17:446–457, 1926.

L. L. Thurstone. Psychophysical analysis. *American Journal of Psychology*, 38: 368–389, 1927.

W. J. van der Linden and R. K. Hambleton, editors. *Handbook of modern item response theory*. Springer, New York, 1997.

B. D. Wright. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14:97–116, 1977.

M. L. Wu, R. J. Adams, and M. Wilson. *ACER ConQuest: Generalized item response modelling software*. Australian Council for Educational Research, Melbourne, Australia, 1998.

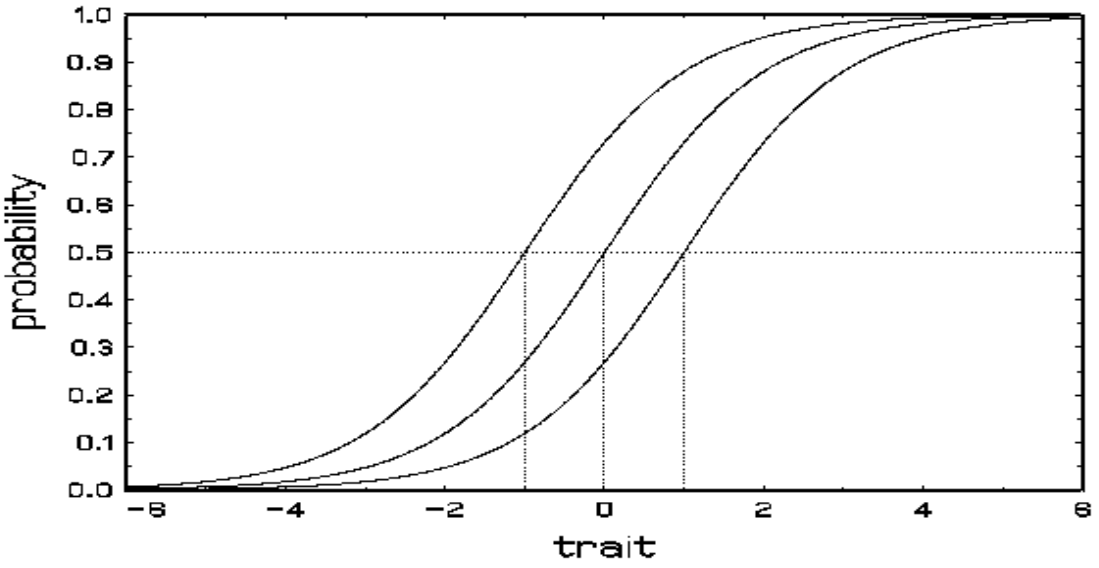

Figure 1. Rasch model with three items.

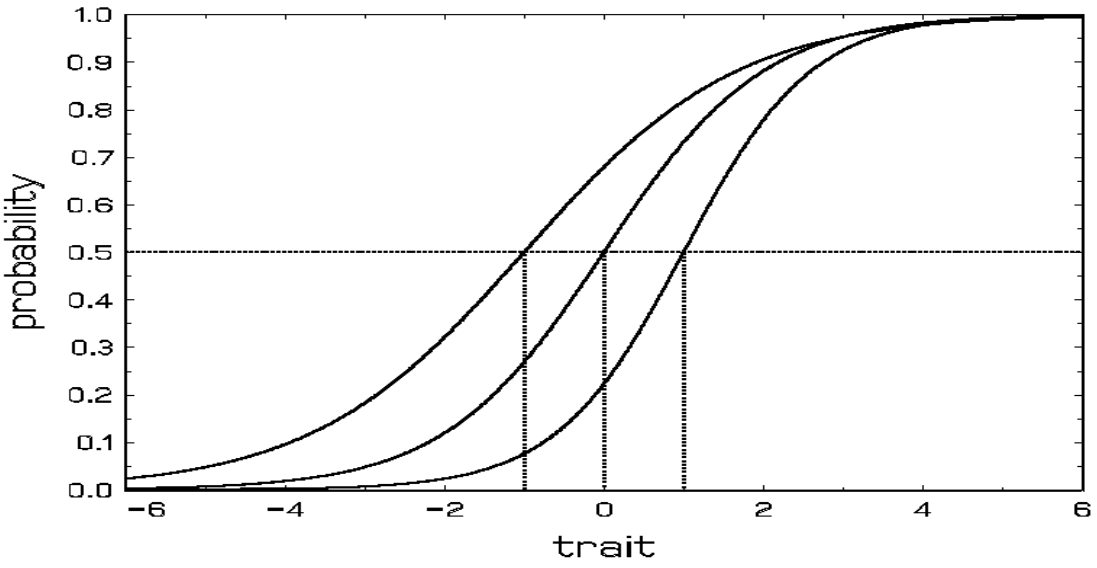

Figure 2. Two-parameter model with three items.

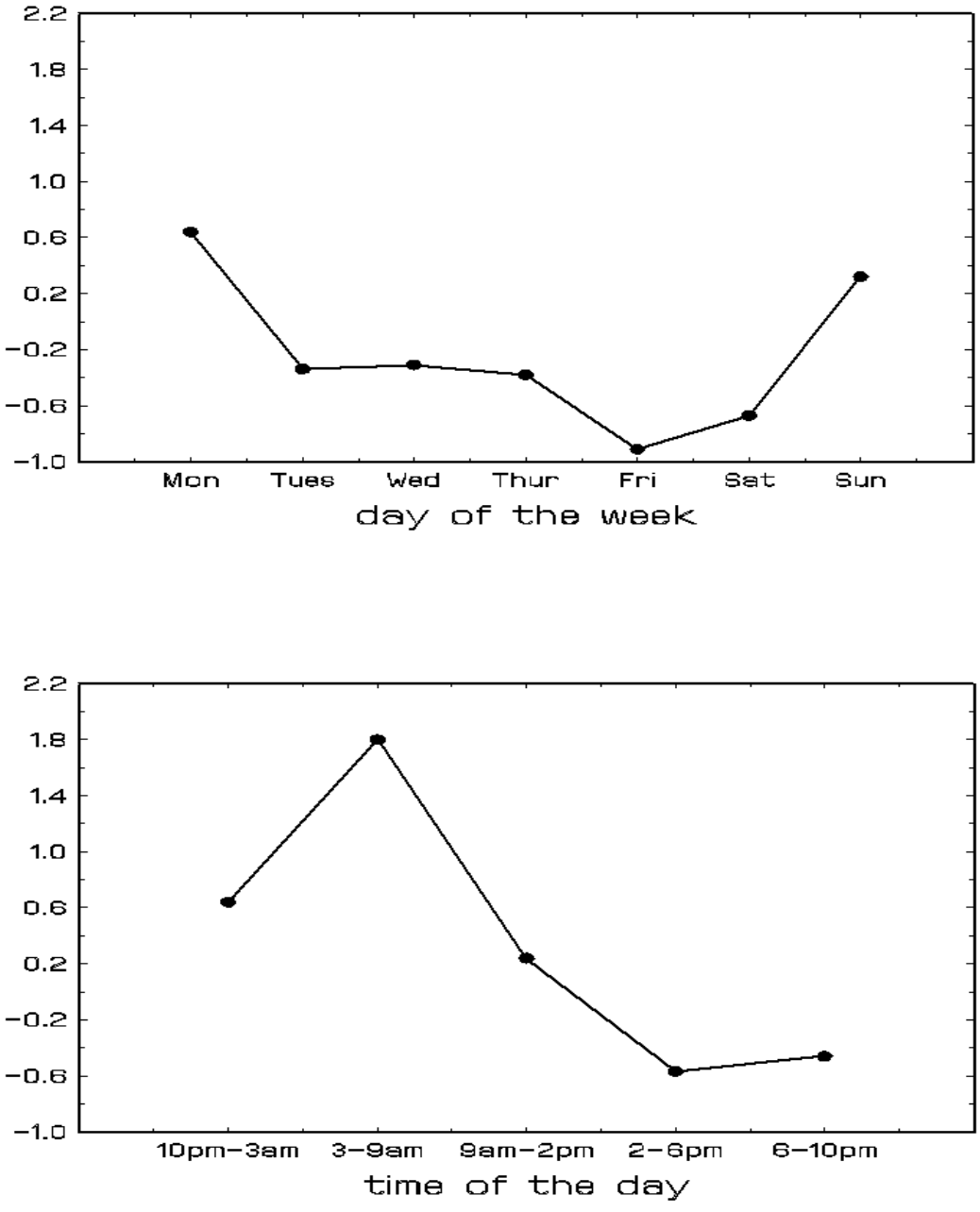

Figure 3. Difficulty estimates based on Day and Time varying slopes model.

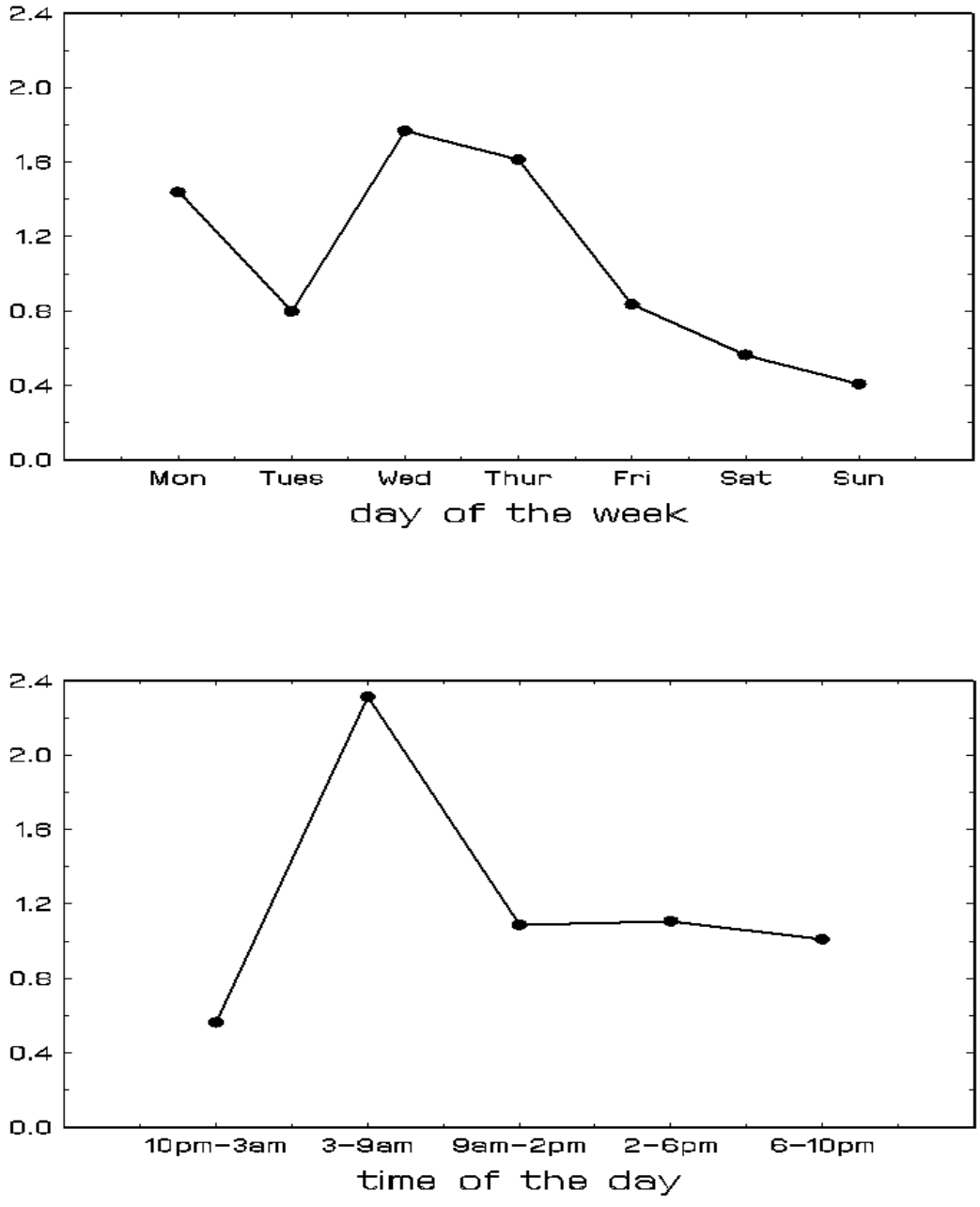

Figure 4. Discrimination estimates based on Day and Time varying slopes model.

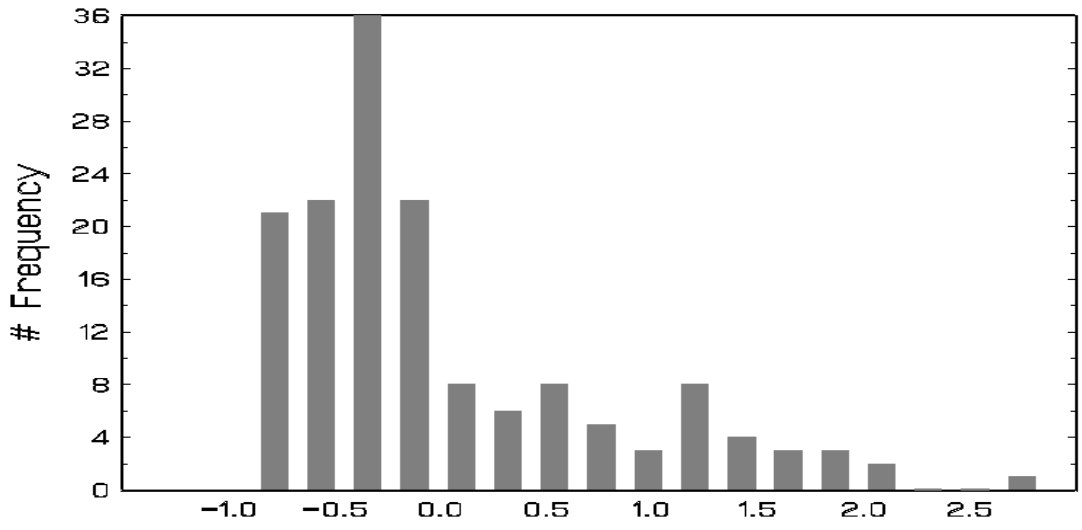

Figure 5. Histogram of empirical Bayes ability estimates from two-parameter model.

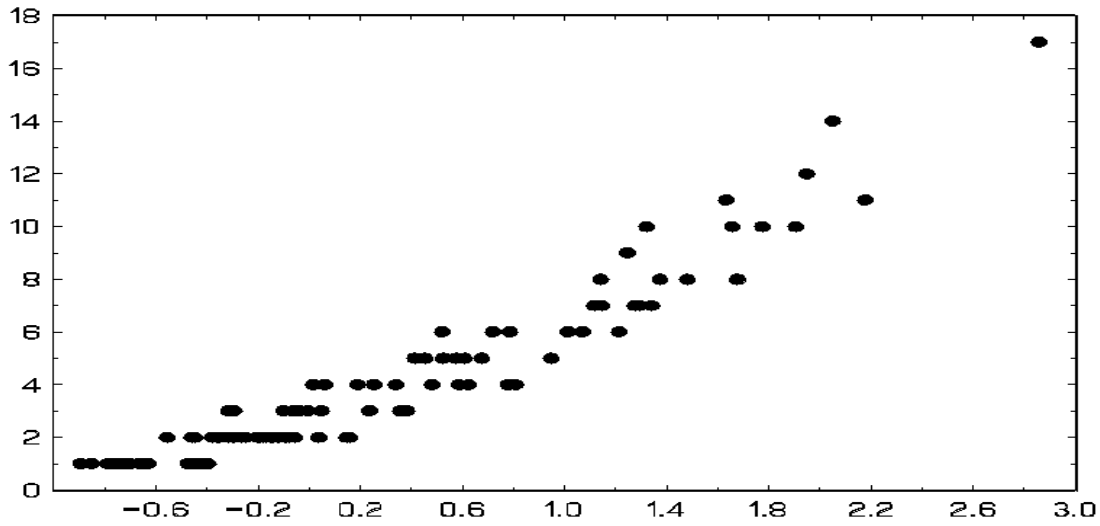

Figure 6. Number of smoking reports versus empirical Bayes estimates.

Table 1

Rasch model estimates for the LSAT-6 data

		NLMIXED	
item	IRT ($\hat{b}_j$)	raw ($\hat{\beta}_j$)	transformed ($\hat{b}_j$)
1	-1.255	2.730	-1.255
2	0.476	0.999	0.476
3	1.234	0.240	1.235
4	0.168	1.306	0.168
5	-0.624	2.099	0.625
	$\hat{a} = 0.755$	$\hat{\sigma} = \hat{a} = 0.755$	

Table 2

Two-parameter probit model estimates for the LSAT-6 data

	IRT		MIXOR raw		MIXOR transformed	
	difficulty	discrimination	difficulty	discrimination	difficulty	discrimination
item	$\hat{b}_j$	$\hat{a}_j$	$\hat{\beta}_j$	$\hat{\sigma}_j$	$\hat{b}_j$	$\hat{a}_j$
1	-.6787	.9788	1.5520	.4169	-.6804	.9779
2	.3161	1.0149	.5999	.4333	.3165	1.0164
3	.7878	1.2652	.1512	.5373	.7867	1.2603
4	.0923	.9476	.7723	.4044	.0926	.9487
5	-.5174	.8397	1.1966	.3587	-.5154	.8415

Table 3

Proportion (and n) of smoking reports by day of week and time of day

Day of Week	3am to 8:59am	9am to 1:59pm	2pm to 5:59pm	6pm to 9:59pm	10pm to 2:59am
Monday	1.3 (2)	4.6 (7)	4.0 (6)	9.2 (14)	2.6 (4)
Tuesday	1.3 (2)	6.6 (10)	15.8 (24)	14.5 (22)	3.3 (5)
Wednesday	5.9 (9)	9.2 (14)	21.7 (33)	11.8 (18)	3.3 (5)
Thursday	5.3 (8)	9.2 (14)	19.7 (30)	17.1 (26)	1.3 (2)
Friday	9.2 (14)	9.9 (15)	21.7 (33)	19.1 (29)	6.6 (10)
Saturday	0.0 (0)	10.5 (16)	16.5 (25)	11.8 (18)	13.2 (20)
Sunday	0.7 (1)	4.0 (6)	4.0 (6)	9.9 (15)	2.6 (4)

Note: proportion calculated as $n/152$, where 152 is the sample size.

Table 4

IRT difficulty estimates (standard errors)

variable	Common slopes	Time varying slopes	Day varying slopes	Day and Time varying slopes
Intercept	-4.16	-4.02	-4.35	-4.15
	(.247)	(.250)	(.324)	(.329)
Tuesday	.748	.747	1.01	.981
	(.227)	(.232)	(.322)	(.321)
Wednesday	1.03	1.03	.988	.945
	(.206)	(.211)	(.334)	(.332)
Thursday	1.04	1.04	1.07	1.02
	(.217)	(.221)	(.321)	(.324)
Friday	1.34	1.34	1.58	1.55
	(.190)	(.194)	(.309)	(.307)
Saturday	1.03	1.03	1.31	1.31
	(.208)	(.215)	(.319)	(.321)
Sunday	-.034	-.034	.359	.318
	(.248)	(.255)	(.374)	(.375)
3am to 9am	-.360	-1.09	-.363	-1.16
	(.221)	(.388)	(.227)	(.385)
9am to 2pm	.564	.455	.569	.400
	(.191)	(.233)	(.199)	(.230)
2pm to 6pm	1.37	1.25	1.38	1.21
	(.181)	(.210)	(.186)	(.212)
6pm to 10pm	1.24	1.11	1.25	1.10
	(.175)	(.220)	(.184)	(.221)

Table 5

IRT discrimination estimates (standard errors)

variable	Common slopes	Time varying slopes	Day varying slopes	Day and Time varying slopes
Intercept	.828			
	(.105)			
Monday			1.05	.740
			(.264)	(.343)
Tuesday			.694	.408
			(.215)	(.303)
Wednesday			1.17	.910
			(.243)	(.316)
Thursday			1.08	.834
			(.225)	(.304)
Friday			.711	.427
			(.180)	(.278)
Saturday			.625	.286
			(.141)	(.194)
Sunday			.472	.211
			(.188)	(.273)
3am to 9am		1.42		1.190
		(.239)		(.290)
9am to 2pm		.785		.564
		(.180)		(.226)
2pm to 6pm		.802		.573
		(.156)		(.199)
6pm to 10pm		.805		.518
		(.128)		(.161)
10pm to 3am		.621		.286
		(.199)		(.194)

Table 6

IRT model fit statistics

	Common slopes	Time varying slopes	Day varying slopes	Day and Time varying slopes
$-2\log L$	2811.79	2803.20	2798.37	2788.27
AIC	2835.79	2835.20	2834.37	2832.27
BIC	2872.08	2883.58	2888.80	2898.80
parameters q	12	16	18	22
Likelihood ratio tests				
comparisons to common slopes model				
χ^2		8.59	13.42	23.52
df		4	6	10
$p <$		.072	.037	.009
comparisons to time varying slopes model				
χ^2			4.83	14.93
df			2	6
$p <$			.089	.021
comparison to day varying slopes model				
χ^2				10.10
df				4
$p <$				.039

$\text{AIC} = -2\log L + 2q; \quad \text{BIC} = -2\log L + q\log N$

Table 7

Minimum and maximum EAP estimate stratified by number of smoking reports: Day of week and time of day of smoking reports (1=report, 0=no report)

Number of reports	EAP estimate	Day	3am to 8:59am	9am to 1:59pm	2pm to 5:59pm	6pm to 9:59pm	10pm to 2:59am
2	-.555	Sat	0	0	1	0	1
	.158	Mon	0	0	0	0	1
		Wed	1	0	0	0	0
3	-.315	Fri	0	0	0	1	1
		Sat	0	0	0	0	1
	.383	Tue	0	0	1	0	0
		Wed	0	0	1	0	0
		Fri	1	0	0	0	0
4	.017	Fri	0	0	1	1	0
		Sat	0	0	0	1	1
	.808	Tue	0	0	0	1	0
		Wed	1	0	0	1	0
		Thu	0	1	0	0	0
5	.414	Wed	0	0	0	1	0
		Fri	0	0	1	1	0
		Sat	0	0	0	1	1
	.947	Thu	1	0	1	0	0
		Fri	1	0	0	0	0
		Sat	0	1	0	1	0