Measurement and Research Department Reports 2003-5

The Nedelsky Model For Multiple Choice Items

Timo Bechger
Gunter Maris
Huub Verstralen
Norman Verhelst

Cito groep

Measurement and Research Department Reports 2003-5

THE NEDELSKY MODEL FOR MULTIPLE CHOICE ITEMS

Timo Bechger Gunter Maris

Huub Verstralen Norman Verhelst

CITO, NATIONAL INSTITUTE FOR EDUCATIONAL MEASUREMENT

ARNHEM

Citogroep
Arnhem, March 3, 2003

Abstract

This chapter is about a psychometric model for multiple choice items based upon the idea that the test-taker responds to a MC question by first eliminating the answers he recognizes as wrong and then guesses at random from the remaining answers. We focus on theoretical properties of the model. Estimation and testing are described briefly.

1. Introduction

Traditionally, multiple choice (MC) items are scored binary; one point is earned when the correct answer is chosen and none when any of the incorrect options (called "distractors") is chosen. This facilitates data analysis but it also entails loss of information (e.g., Levine and Drasgow, 1983). There have been various attempts to include the incorrect options into an IRT model (e.g., Bock, 1972; Thissen and Steinberg, 1984). Here, we discuss a novel model called *the Nedelsky model (NM)*. We focus on the theoretical properties of the NM. Estimation and testing are described briefly. An application to real data can be found in Verstralen (1997) and Verstralen and Verhelst (1998), who invented the model.

The model derives its name from a method for standard setting suggested by Leo Nedelsky in 1954. Nedelsky's method is based upon the idea that the borderline test-taker responds to a MC question by first eliminating the answers he recognizes as wrong and then guesses at random from the remaining answers. The NM generalizes this idea in the sense that the selection of the answers is probabilistic and applies to all levels of ability. It is further assumed that the correct alternative is never rejected, that is, respondents will never think that the correct answer is wrong.

The present discussion is focussed on the theoretical aspects of the NM (esp. Theorem (1) and (2)) and structured as follows. Section 2 provides a brief description of the NM. Some of the psychometric properties of the model are discussed in detail in Section 3. In Section 4, the NM is related to the two- (2PL), the three parameter (3PL) logistic models, and the DECIDE model proposed by Revuelta (2000). Section 5 describes how different scoring rules lead to different amounts of information about the latent ability. Section 6 describes marginal maximum likelihood estimation of the model using an EM-algorithm. Section 7 describes a model test. Section 8 describes the NM as a signal detection model. This topic will explored in more detail in the

ensuing chapter. The chapter is concluded in Section 9.

2. The Nedelsky Model

Consider a MC item i with $J_i + 1$ options arbitrarily indexed $0, 1, ..., J_i$. For convenience, 0 indexes the correct alternative. Let the random variable S_{ij} indicate whether alternative j is recognized as wrong, and define $\mathbf{S}_i$ by the vector $(0, S_{i1}, ..., S_{iJ_i})$. The first entry is fixed at 0 because it is assumed that the correct alternative is never rejected. We refer to $\mathbf{S}_i$ as *a latent subset*. The random variable $S_i^+ \equiv \sum_{j=1}^{J_i} S_{ij}$ denotes the number of distractors that are exposed. Realizations of random variables are denoted with lower case letters.

The probability that alternative answer j is recognized as wrong by a respondent with ability θ is modelled as

$$\Pr(S_{ij} = 1|\theta) = \frac{\exp(\theta - \zeta_{ij})}{1 + \exp(\theta - \zeta_{ij})}, \tag{1}$$

where ζ_{ij} represents the difficulty to recognize that option j of item i is wrong; $S_{i0} = 0$ implies that $\zeta_{i0} = \infty$. One may think of each distractor as a dichotomous Rasch item, where a correct answer is produced if the distractor is recognized to be wrong. This specification implies that $E[S_i^+|\theta] = \sum_{j=1}^{J_i} \Pr(S_{ij} = 1|\theta)$ is increasing in θ. The assumptions that give rise to the Rasch model are discussed by Fischer (1995a).

As explained in the introduction, the process that generates the response is assumed to consist of two stages. In the first stage, a respondent eliminates the answers he recognizes to be wrong. Formally, this means that he draws a latent subset from the set of possible subsets $\Omega_{\mathbf{S}_i}$. Assuming independence among the options given θ, the probability that a subject with ability θ chooses any latent

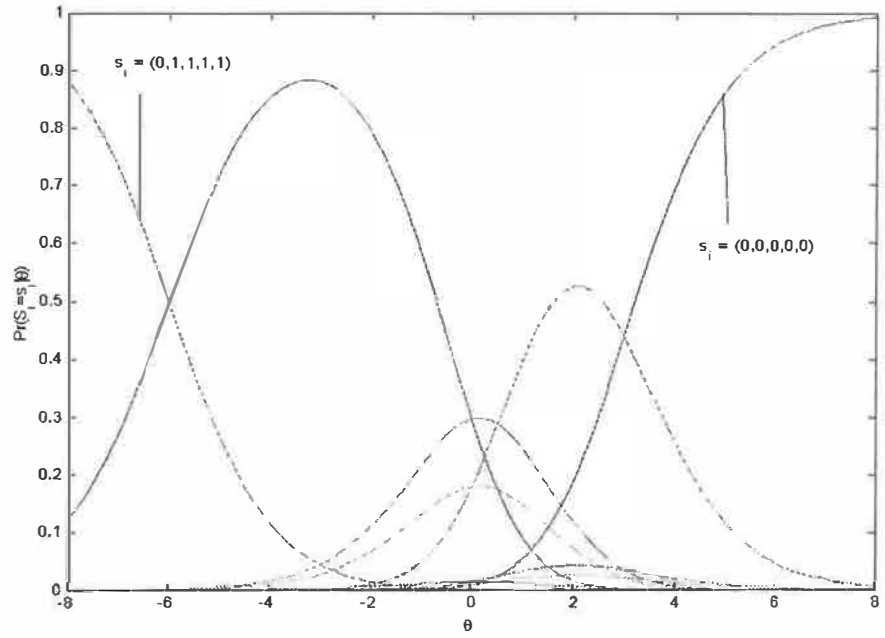

FIGURE 1.
$\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)$ against θ.

subset $\mathbf{s}_i \in \Omega_{\mathbf{S}_i}$ is given by the likelihood of J_i independent Rasch items; i.e.,

$$\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) = \prod_{j=1}^{J_i} \frac{\exp(\theta - \zeta_{ij})^{s_{ij}}}{1 + \exp(\theta - \zeta_{ij})} \tag{2}$$

$$= \frac{\exp\left[\theta s_i^+ - \sum_{j=1}^{J_i} s_{ij}\zeta_{ij}\right]}{\prod_{j=1}^{J_i}\left[1 + \exp(\theta - \zeta_{ij})\right]}, \tag{3}$$

where the sum $\sum_{j=1}^{J_i} s_{ij}\zeta_{ij}$ could be interpreted as a location parameter for the subset $\mathbf{s}_i$ (see Figure (1)).

Once a latent subset is chosen, a respondent guesses at random from the remaining answers. Thus, the conditional probability of responding with option j to item i, given latent subset $\mathbf{s}_i$, is given by:

$$\Pr(X_i = j|\mathbf{S}_i = \mathbf{s}_i) = \frac{1 - s_{ij}}{v(s_i^+)}, \tag{4}$$

where $X_i = j$ denotes the event that the respondent chooses alternative j, and $v(s_i^+) \equiv \sum_{h=0}^{J_i}(1 - s_{ih}) = J_i + 1 - s_i^+$ the number of alternatives to choose from.

This second stage involves a randomization not involving θ, and hence can carry no information about θ. For later reference, $\Pr(X_i = j|\mathbf{S}_i = \mathbf{s}_i)$ is called *the response mapping.* Note that Equation (4) implies that

$$\Pr(X_i = 0|\mathbf{S}_i = \mathbf{s}_i) \geq \Pr(X_i = j|\mathbf{S}_i = \mathbf{s}_i), \text{ and} \tag{5}$$

$$(J_i + 1)^{-1} \leq \Pr(X_i = 0|\mathbf{S}_i = \mathbf{s}_i) \leq 1.$$

Note further that, once a subset is chosen, each alternative in the subset is equally likely to be chosen. This assumption can be relaxed by changing the response mapping as in Equation (16), below.

Combining the two stages of the answer process, we find that the conditional probability of choosing option j with item i is equal to

$$\Pr(X_i = j|\theta) = \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \tag{6}$$

$$= \sum_{s_i^+} \frac{\Pr(S_{ij} = 0|S_i^+ = s_i^+)}{v(s_i^+)} \Pr(S_i^+ = s_i^+|\theta). \tag{7}$$

The second equality is ascertained using the fact that S_i^+ is a sufficient statistic for θ and $\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)$ may be written as $\Pr(\mathbf{S}_i = \mathbf{s}_i|S_i^+ = s_i^+)\Pr(S_i^+ = s_i^+|\theta)$.

Remark 1. *Since* $\Pr(S_{ij} = 1|\theta)$ *is modelled by the Rasch model, expressions for* $\Pr(S_{ij} = 0|S_i^+ = s_i^+)$ *and* $\Pr(S_i^+ = s_i^+|\theta)$ *are well known. Specifically, if we define* $\varepsilon_{ij} = \exp(-\zeta_{ij})$,

$$\Pr(S_i^+ = s_i^+|\theta) = \frac{\gamma_{s_i^+}(\varepsilon_i)\exp(s_i^+\theta)}{\prod_{j=1}^{J_i}[1 + \varepsilon_{ij}\exp(\theta)]}, \tag{8}$$

where $\gamma_{s_i^+}(\varepsilon_i)$ *denotes an elementary symmetric function of order* s_i^+ *with argument* $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ_i})$. *This is a special case of Equation (7) and ways to calculate* $\Pr(S_i^+ = s_i^+|\theta)$ *were discussed in the first chapter. Further,*

$$\Pr(S_{ij} = 0|S_i^+ = s_i^+) = 1 - \frac{\varepsilon_{ij}\gamma_{s_i^+-1}^{(ij)}(\varepsilon_i)}{\gamma_{s_i^+}(\varepsilon_i)}, \tag{9}$$

where the subscript (ij) *in* $\gamma^{(ij)}_{s_i^+-1}(\varepsilon_i)$ *denotes that* ε_{ij} *is ignored in the argument. If we make use of the result that* $\gamma_{s_i^+}(\varepsilon_i) - \varepsilon_{ij}\gamma^{(ij)}_{s_i^+-1}(\varepsilon_i) = \gamma^{(ij)}_{s_i^+}(\varepsilon_i)$, *it follows that*

$$\begin{aligned}
\Pr(X_i = j|\theta) &= \sum_{s_i^+} \left(\gamma_{s_i^+}(\varepsilon_i) - \varepsilon_{ij}\gamma^{(ij)}_{s_i^+-1(\varepsilon_i)}\right) \frac{\exp(s_i^+\theta)}{v(s_i^+)\prod_{j=1}^{J_i}[1+\varepsilon_{ij}\exp(\theta)]} \\
&= \left(\prod_{j=1}^{J_i}[1+\varepsilon_{ij}\exp(\theta)]\right)^{-1} \sum_{s_i^+} \frac{\gamma^{(ij)}_{s_i^+}(\varepsilon_i)\exp(s_i^+\theta)}{v(s_i^+)\prod_{j=1}^{J_i}[1+\varepsilon_{ij}\exp(\theta)]} \\
&= \left(\sum_{s_i^+} \gamma^{(ij)}_{s_i^+(\varepsilon_i)}\exp(s_i^+\theta)\right)^{-1} \sum_{s_i^+} \frac{\gamma^{(ij)}_{s_i^+}(\varepsilon_i)\exp(s_i^+\theta)}{v(s_i^+)}.
\end{aligned}$$

This formulation is chosen to indicate that $\prod_{j=1}^{J_i}[1+\varepsilon_{ij}\exp(\theta)]$ *is build up summing over values of* s_i^+ *to calculate the second factor.*

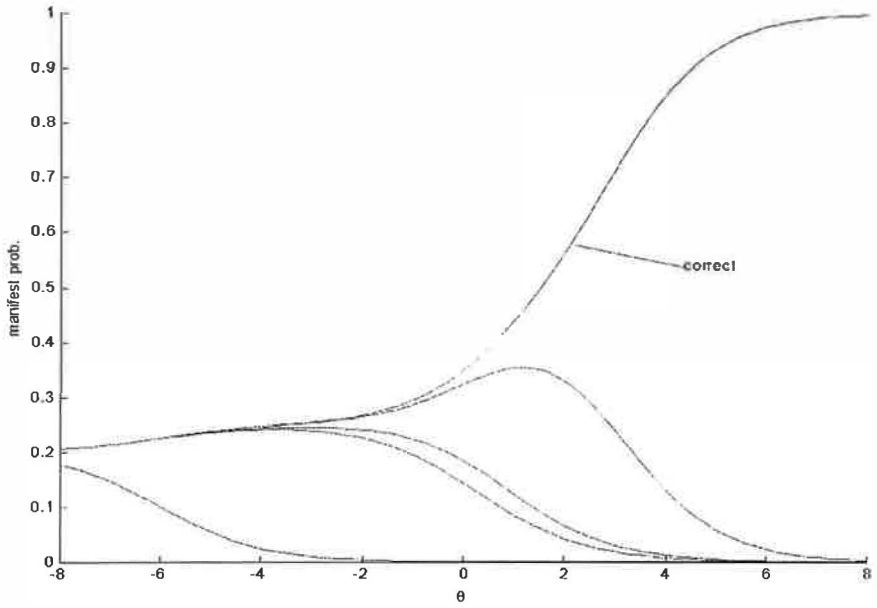

FIGURE 2.
$\Pr(X_i = j|\theta)$ against θ.

There are four properties of the model that are readily seen in Figure (2). First,

$\Pr(S_i^+ = 0|\theta) \to 1$ if $\theta \to -\infty$ which implies that, for $j = 0, ..., J_i$,

$$\lim_{\theta \to -\infty} \Pr(X_i = j|\theta) = \frac{1}{J_i + 1}. \tag{10}$$

Second, if $\theta \to \infty$, $\Pr(S_i^+ = J_i|\theta) \to 1$ and

$$\lim_{\theta \to \infty} \Pr(X_i = 0|\theta) = 1. \tag{11}$$

Third,

$$\Pr(X_i = 0|\theta) - \Pr(X_i = j|\theta) = \sum_{\mathbf{s}_i} \frac{s_{ij}}{v(s_i^+)} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) > 0 \tag{12}$$

and the probability of a correct response is always larger than the probability to choose a distractor. Finally, Figure (2) suggests that $\Pr(X_i = 0|\theta)$ is an increasing function of θ. This is proven in Corollary (2), below.

3. Psychometric Properties of the Nedelsky Model

3.1. Monotone Option Ratios

The *option ratios* are defined as:

$$\psi_{itj}(\theta) \equiv \frac{\Pr(X_i = t|\theta)}{\Pr(X_i = j|\theta)}. \tag{13}$$

The model has *monotone option ratios (MOR)* if all option ratios are monotone in θ. MOR means that

$$\begin{aligned} \Pr(X_i = t|X_i \in \{t, j\}, \theta) &= \frac{\Pr(X_i = t|\theta)}{\Pr(X_i = t|\theta) + \Pr(X_i = j|\theta)} \\ &= \frac{\psi_{itj}(\theta)}{1 + \psi_{itj}(\theta)} \end{aligned} \tag{14}$$

is monotone in θ. Hence, there exists an ordering of the alternatives so that the model classifies as a nonparametric partial credit model (Hemker, Sijtsma, Molenaar, and Junker, 1997). For the NM it can be shown that this ordering is given by the ordering of the location parameters; that is, $\psi_{itj}(\theta)$ (and $\Pr(X_i = t|X_i \in \{t, j\}, \theta)$) is increasing in θ if $\zeta_{it} > \zeta_{ij}$. To proof MOR, we use the following lemma:

lemma 1. *Assume that $J_i > 1$ and let $\Omega_{\mathbf{S}_i^{(j,t)}}$ denote the set of latent subsets without alternative j and t.*

$$\frac{\Pr(X_i = j|\theta)}{\Pr(S_{ij} = 0|\theta)} = \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t,j)}}} \left(\frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+) + 1} + \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 2} \right) \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta).$$

When summing over $\Omega_{\mathbf{S}_i^{(t,j)}}$, all quantities are calculated as if alternatives are missing. Note that $\Pr(S_{i0} = 0|\theta) = 1$. Note further that the ratio on the left side equals the probability that $X_i = j$ conditional upon θ and $S_{ij} = 0$; the alternative is taken into consideration.

Proof. *Let* $\gamma_t(\theta; \mathbf{s}_i) \equiv \prod_{h \neq t}^{J_i} \Pr(S_{ih} = 0|\theta)^{1-s_{ih}} \Pr(S_{ih} = 1|\theta)^{s_{ih}}$, *and let* $\gamma_{t,j}(\theta; \mathbf{s}_i) \equiv \prod_{h \neq t,j}^{J_i} \Pr(S_{ih} = 0|\theta)^{1-s_{ih}} \Pr(S_{ih} = 1|\theta)^{s_{ih}}$.

$$\Pr(X_i = j|\theta) = \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \prod_{h=1}^{J_i} \Pr(S_{ih} = 0|\theta)^{1-s_{ih}} \Pr(S_{ih} = 1|\theta)^{s_{ih}}$$

$$= \sum_{\mathbf{s}_i; s_{it}=1} (1 - s_{ij}) \frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+)} \gamma_t(\theta; \mathbf{s}_i) + \sum_{\mathbf{s}_i; s_{it}=0} (1 - s_{ij}) \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+)} \gamma_t(\theta; \mathbf{s}_i)$$

$$= \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t)}}} (1 - s_{ij}) \frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+)} \gamma_t(\theta; \mathbf{s}_i) + \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t)}}} (1 - s_{ij}) \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 1} \gamma_t(\theta; \mathbf{s}_i)$$

$$= \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t)}}} (1 - s_{ij}) \left(\frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+)} + \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 1} \right) \gamma_t(\theta; \mathbf{s}_i)$$

$$= \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t)}}; s_{ij}=0} \Pr(S_{ij} = 0|\theta) \left(\frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+)} + \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 1} \right) \gamma_{t,j}(\theta; \mathbf{s}_i)$$

$$= \Pr(S_{ij} = 0|\theta) \sum_{\mathbf{s}_i \in \Omega_{\mathbf{S}_i^{(t,j)}}} \left(\frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+) + 1} + \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 2} \right) \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta).$$

□

Theorem 1. *The NM has MOR.*

Proof. Some algebra (see Appendix) shows that

$$\frac{\partial}{\partial \theta} \psi_{itj}(\theta) = \frac{\Pr(X_i = t|\theta) \Pr(S_{ij} = 0|\theta) - \Pr(X_i = j|\theta) \Pr(S_{it} = 0|\theta)}{\Pr(X_i = j|\theta)^2}.$$

This function has the same sign as

$$\frac{\Pr(X_i = t|\theta)}{\Pr(S_{it} = 0|\theta)} - \frac{\Pr(X_i = j|\theta)}{\Pr(S_{ij} = 0|\theta)}. \qquad (*)$$

Now, Lemma (1) implies that (*) equals

$$\sum_{\mathbf{s}_i \in \Omega_{\mathbf{s}_i^{(j,t)}}} c_{tj}(\theta, \mathbf{s}_i) \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta),$$

where

$$\begin{aligned} c_{tj}(\theta, \mathbf{s}_i) &= \frac{\Pr(S_{ij} = 1|\theta)}{v(s_i^+) + 1} + \frac{\Pr(S_{ij} = 0|\theta)}{v(s_i^+) + 2} - \frac{\Pr(S_{it} = 1|\theta)}{v(s_i^+) + 1} - \frac{\Pr(S_{it} = 0|\theta)}{v(s_i^+) + 2} \\ &= (\Pr(S_{ij} = 1|\theta) - \Pr(S_{it} = 1|\theta)) \left[\frac{1}{v(s_i^+) + 1} - \frac{1}{v(s_i^+) + 2}\right]. \end{aligned}$$

If $\zeta_{it} > \zeta_{ij}$, $\Pr(S_{it} = 1|\theta) < \Pr(S_{ij} = 1|\theta)$, and it follows that $\frac{\partial}{\partial\theta}\psi_{itj}(\theta) > 0$ for all θ. Using (15) the reader may verify that, if $J_i = 1$, $\psi_{i0j}(\theta) = \frac{1}{2}\Pr(S_{i1} = 1|\theta)$, which is increasing in θ. □

Corollary 1. *If* $\zeta_{it} > \zeta_{ij}$, $\Pr(X_i = t|\theta) > \Pr(X_i = j|\theta)$.

Proof. It follows from (10) that $\lim_{\theta\to-\infty} \varphi_{itj}(\theta) = 1$. *Hence, MOR implies that* $\varphi_{itj}(\theta) > 1$ *for all* θ *and the result follows.* □

Corollary (1) implies that the ordering among the option parameters can be inferred from the marginal probabilities.

Corollary 2. $\Pr(X_i = 0|\theta)$ *is an increasing function of* θ.

Proof. MOR implies that

$$\sum_{j=1}^{J_i} \varphi_{ij0} = \Pr(X_i = 0|\theta)^{-1}(1 - \Pr(X_i = 0|\theta))$$

is decreasing in θ.

$$\frac{\partial}{\partial\theta}\Pr(X_i = 0|\theta)^{-1}(1 - \Pr(X_i = 0|\theta)) = -\frac{\frac{\partial}{\partial\theta}\Pr(X_i = 0|\theta)}{\Pr(X_i = 0|\theta)^2}.$$

Hence, $\frac{\partial}{\partial\theta}\Pr(X_i = 0|\theta) > 0$ *and* $\Pr(X_i = 0|\theta)$ *is be increasing in* θ. □

Further consequences of MOR are discussed in the next paragraph.

3.2. Monotone Likelihood Ratio in the Item Score

An item score is a discrete random variable $C_i : \{0, 1, ..., J_i\} \rightarrow \mathbb{R}$ such that the value $C_i(j)$ represents the credit given to the event $X_i = j$. If C_i is such that $C_i(t) > C_i(j)$ if $\zeta_{it} > \zeta_{ij}$ and $C_i(t) = C_i(j)$ if $\zeta_{it} = \zeta_{ij}$, MOR is equivalent to *monotone likelihood ratio* (MLR) in C_i (Lehmann, 1959, p. 68). MLR in C_i implies two useful properties: stochastic ordering of the latent trait by C_i (SOL in C_i), and stochastic ordering of the item score by the latent trait (SOM in C_i) (Lehmann, 1959, p. 74; Junker, 1993, Proposition 4.1).

Corollary 3. *SOL in* C_i*: If* $\zeta_{it} > \zeta_{ij}$, $\Pr(\theta > a|X_i = t) > \Pr(\theta > a|X_i = j)$ *for any constant value* a *of* θ.

Proof. It follows from Bayes' theorem that

$$\frac{\Pr(\theta|X_i = t)}{\Pr(\theta|X_i = j)} = \varphi_{itj}(\theta) \Leftrightarrow$$

$$\Pr(\theta|X_i = t) = \varphi_{itj}(\theta)\Pr(\theta|X_i = j) \Rightarrow$$

$$\int_a^\infty \Pr(\theta|X_i = t)d\theta = \int_a^\infty \varphi_{itj}(\theta)\Pr(\theta|X_i = j)d\theta.$$

Since $\varphi_{itj}(\theta) > 1$ *it may be concluded that*

$$\int_a^\infty \Pr(\theta|X_i = t)d\theta > \int_a^\infty \Pr(\theta|X_i = j)d\theta.$$

□

Corollary 4. *SOM in* C_i*: If* $\theta^{(2)} > \theta^{(1)}$, $\Pr(C_i \geq y|\theta^{(2)}) > \Pr(C_i \geq y|\theta^{(1)})$, *where* y *is a value in the range of* C_i.

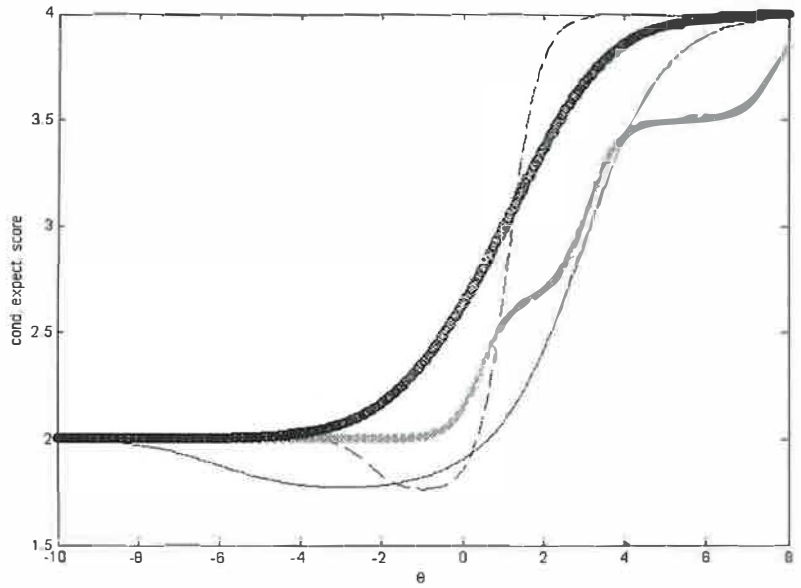

FIGURE 3.

Plot of $E[C_i|\theta]$ for different scoring functions.

Proof. *Let* $M \equiv \max(C_i)$. *MOR implies that*

$$\sum_{j=y}^{M} \varphi_{ijM}(\theta^{(2)}) > \sum_{j=y}^{M} \varphi_{ijM}(\theta^{(1)}) \Leftrightarrow$$

$$\frac{\Pr(C_i \geq y|\theta^{(2)})}{\Pr(C_i = M|\theta^{(2)})} > \frac{\Pr(C_i \geq y|\theta^{(1)})}{\Pr(C_i = M|\theta^{(1)})} \Leftrightarrow$$

$$\frac{\Pr(C_i \geq y|\theta^{(2)})}{\Pr(C_i \geq y|\theta^{(1)})} > \frac{\Pr(C_i = M|\theta^{(2)})}{\Pr(C_i = M|\theta^{(1)})} > 1.$$

$\Pr(C_i = M|\theta)$ *equals the probability of a correct response and the last inequality follows from Theorem (2).* □

SOL in C_i implies that θ is stochastically increasing in C_i so that $E[\theta|C_i]$ is an increasing function of C_i (Ross, 1996, Lemma 9.1.2), and $Corr(\theta, C_i) \geq 0$. SOM in C_i implies that C_i is stochastically ordered by θ so that $E[C_i|\theta]$ is monotone non-decreasing in θ; a desirable property in practical applications. Figure (3) illustrates that $E[\theta|C_i]$ need not be increasing in θ for just any scoring rule.

Let the test score C be defined as $f(C_1, C_2, ..., C_I)$, where f is an increasing function.

Proposition 1. *SOM in C_i implies SOM in C.*

Proof. Let $\theta^{(2)} > \theta^{(1)}$. Consider $C_1, C_2, ..., C_I$ given $\theta^{(2)}$ and $C_1, C_2, ..., C_I$ given $\theta^{(1)}$ which are both independent. Example 9.2.(a) in Ross (1996) shows that SOM in C_i implies that

$$\Pr(f(C_1, C_2, ..., C_I) > a|\theta^{(2)}) > \Pr(f(C_1, C_2, ..., C_I) > a|\theta^{(1)}),$$

i.e., SOM in C. □

SOM in C implies that $E[C|\theta]$ is increasing in θ. Unfortunately, SOL in C_i is not a sufficient condition for SOL in C (Hemker, et. al., 1997). Lemma (2), below, implies SOL in the number of correct responses but it remains a topic for future study to describe the class of functions $f(\cdot)$ that give SOL in C.

4. Relations between the Nedelsky Model and Other IRT Models

Relations to the Two- and Three Parameter Logistic Models If the item is dichotomous (i.e., $J_i = 1$)

$$\Pr(X_i = 0|\theta) = \frac{1}{2} + \left(1 - \frac{1}{2}\right) \Pr(S_{i1} = 1|\theta), \tag{15}$$

where $\frac{1}{2} = \Pr(X_i = 0|S_{i1} = 0)$; the probability to find the correct answer by guessing. $\Pr(S_{i1} = 1|\theta)$ is the probability that the respondent knows the correct answer. The probability to find the correct answer by guessing is fixed at $\frac{1}{2}$ because respondents who failed to eliminate the wrong answer find both alternatives equally attractive. To relax this assumption, we include parameters $\tau_{ij} > 0$ to represent the attractiveness of distractors relative to the correct alternative. In general, the response mapping

would then become

$$\Pr(X_i = j|\mathbf{S}_i = \mathbf{s}_i; \tau_{i1}, ..., \tau_{iJ_i}) = \frac{(1 - s_{ij})\,\tau_{ij}}{\sum_{k=0}^{J_i}(1 - s_{ik})\tau_{ik}}, \tag{16}$$

with $\tau_{i0} = 1$. It follows that, in the dichotomous case,

$$\Pr(X_i = 0|\theta) = \frac{1}{1 + \tau_{i1}} + \left(1 - \frac{1}{1 + \tau_{i1}}\right) \Pr(S_{i1} = 1|\theta). \tag{17}$$

Note that this model is not identifiable. Specifically, for any constant c, the transformations

$$\begin{aligned} \theta^* &= \ln(\exp(\theta) + c) \\ \zeta_{i1}^* &= \ln(\exp(\zeta_{i1}) - c) \\ \tau_{i1}^* &= \tau_{i1}\frac{\exp(\zeta_{i1}) + c}{\exp(\zeta_{i1}) - c(1 + \tau_{i1})} \end{aligned} \tag{18}$$

leave the probability $\Pr(X_i = 0|\theta)$ unchanged. This problem remains if the value of ζ_{i1} is fixed (Maris, 2002).

Remark 2. *To see where these transformations come from it is useful to reparameterize the model as follows:*

$$\begin{aligned} \Pr(X_i = 0|\theta) &= \lambda_i + (1 - \lambda_i)\frac{\exp(\theta - \zeta_{ij})}{1 + \exp(\theta - \zeta_{ij})} \\ &= \lambda_i + (1 - \lambda_i)\frac{\exp(\theta)\exp(-\zeta_{ij})}{1 + \exp(\theta)\exp(-\zeta_{ij})} \\ &= \lambda_i + (1 - \lambda_i)\frac{\exp(\theta)}{\exp(\zeta_{ij}) + \exp(\theta)} \\ &= \frac{\lambda_i \exp(\zeta_{ij}) + \lambda_i \exp(\theta) + \exp(\theta) - \lambda_i \exp(\theta)}{\exp(\zeta_{ij}) + \exp(\theta)} \\ &= \frac{\exp(\theta) + \lambda_i \exp(\zeta_{ij})}{\exp(\zeta_{ij}) + \exp(\theta)} \\ &= \frac{\exp(\theta) + \lambda_i \exp(\zeta_{ij})}{\exp(\theta) + \lambda_i \exp(\zeta_{ij}) + (1 - \lambda_i)\exp(\zeta_{ij})} \end{aligned}$$

Then reparameterize using the following definitions:

$$t = \exp(\theta),\ t \in \mathbb{R}^+$$

$$a_i = \lambda_i \exp(\zeta_{ij}),\ a_i \in \mathbb{R}^+$$

$$b_i = (1 - \lambda_i) \exp(\zeta_{ij}),\ b_i \in \mathbb{R}^+$$

such that; $t = \exp(\theta)$,$\zeta_{ij} = \ln(a_i + b_i)$, $\lambda_i = \frac{a_i}{a_i + b_i}$, *and we may write:*

$$\Pr(X_i = 0|\theta) = \frac{t + a_i}{t + a_i + b_i}.$$

With this parameterization it is relatively easy to see, that e.g., simultaneous translations of t *and* a_i *would not change the probability* $\Pr(X_i = 0|\theta)$*; the details are in Maris (2002). This parameterization turns out to be quite useful in some situations and it will be used in the final chapter of this booklet.*

The 3PL (Birnbaum, 1968) is obtained if we further include an item-specific discrimination parameter $a_i > 0$ so that

$$\Pr(S_{ij} = 1|\theta) = \frac{\exp(a_i\theta - \zeta_{ij})}{1 + \exp(a_i\theta - \zeta_{ij})}. \tag{19}$$

Thus, the NM is a special case of the 3PL where $\tau_{i1} = a_i = 1$, for all items. The 2PL is obtained when $\tau_{i1} \dashrightarrow \infty$; meaning that respondents who did not exclude the incorrect alternative will never choose the correct alternative by guessing. Including an option-specific discrimination parameter would establish the identifiability of the NM with attractiveness parameters but we will not consider this possibility.

4.1. Relations Between the NM and the DECIDE model

Revuelta (2000) has developed the DECIDE model for MC items which appears to behave very similar to the NM. The DECIDE model assumes that

$$\varphi_{i0_j}(\theta) = \exp(\theta - \zeta_{ij}) + 1 \tag{20}$$

Let $\Omega_{\mathbf{S}_i^{(j)}}$ denote the set of latent subsets not involving alternative j. Using Equations (12), and (6), it is readily seen that

$$\varphi_{i0j}(\theta) = 1 + \exp(\theta - \zeta_{ij}) \frac{\sum_{s_i \in \Omega_{\mathbf{S}_i^{(j)}}} \frac{1}{1+\sum_{h \neq j}(1-s_{ih})} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)}{\sum_{s_i \in \Omega_{\mathbf{S}_i^{(j)}}} \frac{1}{\sum_{h \neq j}(1-s_{ih})} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)} \tag{21}$$

under the Nedelsky model. We see that the NM is never exactly equal to the DECIDE model but the selection ratio under the NM is well approximated by $1+\exp(\theta - \zeta_{ij})$ (see Figure (4)). This means that, in practice, the two models will hardly be distinguishable.

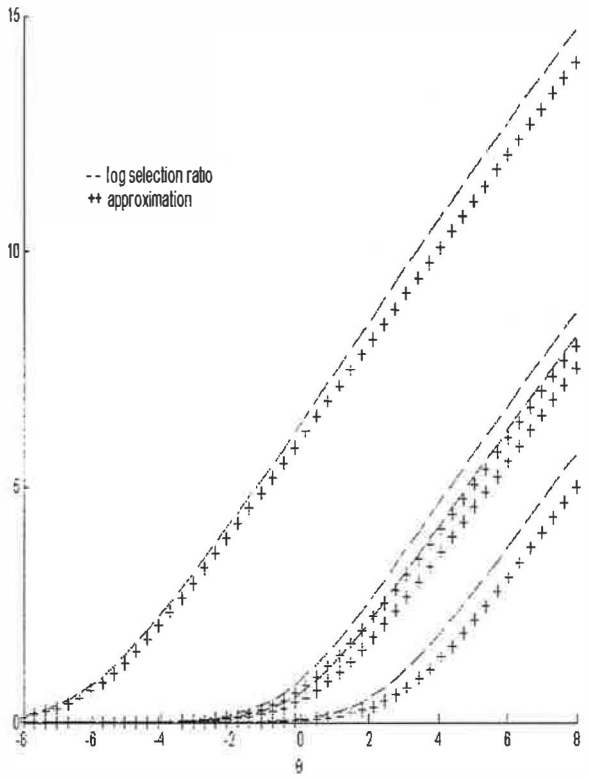

FIGURE 4.

Two plots for each of four incorrect options. One of $\ln \varphi_{i0j}(\theta)$ and one of $\ln(\exp(\theta - \zeta_{ij}) + 1)$ using the same parameters as the previous figures.

The close resemblance between the Nedelsky and the DECIDE model is useful

for two reasons: First, Revuelta has established many relations between the DECIDE model and other models. Second, it allows us to generalize the results of Revuelta's simulation studies to the Nedelsky model. These results support for instance, the conclusion that estimating a NM with option-specific discrimination parameters is unfeasible unless one has a huge sample.

5. Information About θ Provided by Different Types of Data

In this section we consider item information functions for: option scoring, where we register the answer that the respondent has chosen, binary scoring, where it is registered whether the answer was correct or not, and subset scoring, where we have somehow been able to observe the respondents latent subset. It will be shown that subset scoring will provide more information than option scoring and option scoring will provide more information than binary scoring unless there are only two alternative answers.

In general, the item information function is defined by

$$Inf_X(\theta) \equiv E_\theta \left[\left(\frac{\partial}{\partial \theta} \ln L(\theta|X) \right)^2 \right] \tag{22}$$

$$= \sum_x \frac{\left[\frac{\partial}{\partial \theta} \Pr(X = x|\theta) \right]^2}{\Pr(X = x|\theta)}, \tag{23}$$

where subscript X refers to the type of data used. This equation will be used to derive the information.

First, the expected information that would be obtained about θ if the latent

subset would be observed is

$$
\begin{aligned}
Inf_{\mathbf{S}_i}(\theta) &= \sum_{\mathbf{s}_i} \frac{\left(\frac{\partial}{\partial\theta}\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)\right)^2}{\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)} \\
&= \sum_{\mathbf{s}_i} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)(s_i^+ - E[S_i^+|\theta])^2 \\
&= Var(S_i^+|\theta),
\end{aligned}
\tag{24}
$$

which is what one expects since $\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta)$ belongs to the exponential family. Note that the cardinality of a respondent 's subset contains all information about θ. Thus, if we observe the latent subsets we have complete data.

The information function for option scoring is equal to

$$
Inf_{X_i}(\theta) = a_i^2 \sum_{j=0}^{J_i} \frac{\left(E[v(S_i^+)|\theta]\Pr(X_i = j|\theta) - \Pr(S_{ij} = 1|\theta)\right)^2}{\Pr(X_i = j|\theta)},
\tag{25}
$$

where $\Pr(S_{i0} = 1|\theta) = 1$. With option scoring we no longer know the latent subset. We do know that the subset must have contained the alternative that was chosen. The missing information principle implies that $Inf_{X_i}(\theta) < Inf_{\mathbf{S}_i}(\theta)$.

The information function for option scoring can be compared to the information with *binary scoring*; when only correct and incorrect are registered. Let Z_i denote a random variable that is 0 is $X_i = 0$ and 1 otherwise. With binary scoring the information function is equal to

$$
\begin{aligned}
Inf_{Z_i}(\theta) &= \frac{\left[\frac{\partial}{\partial\theta}\Pr(X_i = 0|\theta)\right]^2}{\Pr(X_i = 0|\theta)} + \frac{\left[\frac{\partial}{\partial\theta}\left(1 - \Pr(X_i = 0|\theta)\right)\right]^2}{1 - \Pr(X_i = 0|\theta)} \\
&= \frac{\left[\frac{\partial}{\partial\theta}\Pr(X_i = 0|\theta)\right]^2}{\Pr(X_i = 0|\theta)(1 - \Pr(X_i = 0|\theta))} \\
&= \frac{\left(E[v(S_i^+)|\theta]\Pr(X_i = 0|\theta) - 1\right)^2}{\Pr(Z_i = 0|\theta)(1 - \Pr(Z_i = 0|\theta))},
\end{aligned}
\tag{26}
$$

where $v(S_i^+) = J_i - S_i^+ + 1$. With binary scoring, we only know that a respondent with a wrong answer, included one or more wrong answers in his subset. Thus, more data is lost compared to option scoring, and it can be shown that in general,

$Inf_{\mathbf{S}_i}(\theta) > Inf_{X_i}(\theta) \geq Inf_{Z_i}(\theta)$ (Maris & Bechger, 2003). It is easy to see that the information functions for binary- and option scoring coincide if $J_i = 2$, or when all options are equally difficult.

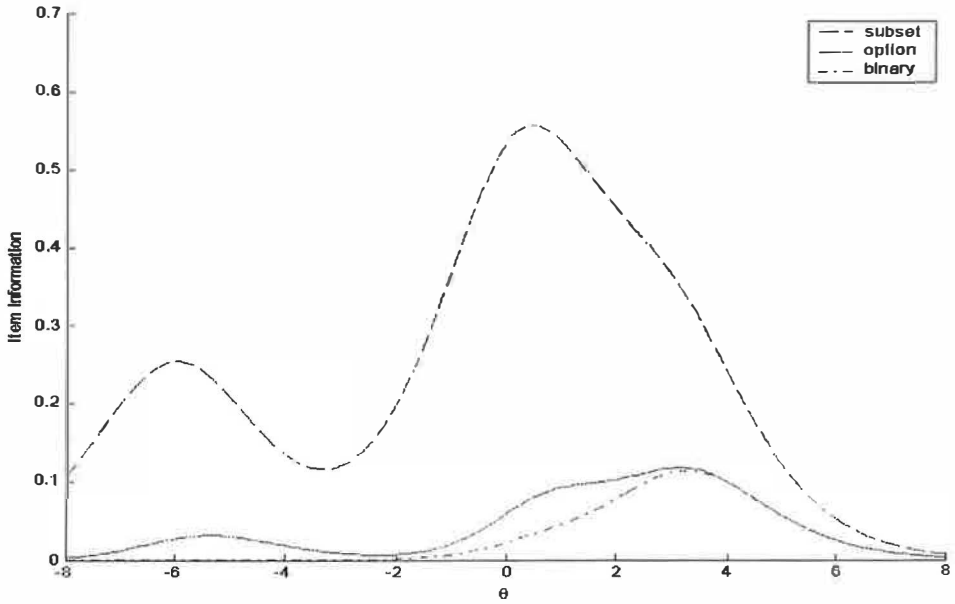

FIGURE 5.

Plots of $Inf_{\mathbf{S}_i}(\theta)$, $Inf_{X_i}(\theta)$ and $Inf_{Z_i}(\theta)$ for an item.

Figure (5) shows item information functions for option-, binary- and subset scoring. It clearly suggest that subset scoring provides much more information than the other approaches while option scoring does slightly better than binary scoring. Noteworthy is also that the information functions are not single peaked as in the Rasch model. Verstralen and Verhelst (1998) report a very similar figure using a real data set.

6. Marginal Maximum Likelihood Estimation By Means of a Generalized EM-Algorithm

6.1. Introduction

Consider a random sample of N respondent from a normal population with zero mean and variance σ_θ^2. Let v denote an arbitrary respondent. Each respondent is administered I items resulting in a data matrix denoted by $\mathbf{x}$ which constitutes *observed data.* The rows of $\mathbf{x}$ are denoted by $\mathbf{x}_v$. The latent subsets and the abilities constitute *latent data.* Abilities are in a vector $\boldsymbol{\theta}$ and the latent subsets are gathered in a latent subset matrix $\mathbf{s}$. The rows of $\mathbf{s}$ are latent subset vectors $\mathbf{s}_v = (\mathbf{s}_{v1}, ..., \mathbf{s}_{vI})$, where $\mathbf{s}_{vi} = (s_{vi1}, ..., s_{viJ_i})$ is a latent subset. Further, let $s_{vi}^+ = \sum_{j=1}^{J_i} s_{vij}$, $s_v^+ = \sum_{i=1}^{I} s_{vi}^+$, and $\mathbf{s}^+ = (s_1^+, ..., s_N^+)$. The entry corresponding to the correct alternative is now ignored and an index for the respondent is added. When possible, $\mathbf{s}_{vi}$ is used in place of $\mathbf{S}_{vi} = \mathbf{s}_{vi}$, x_{vi} in place of $X_{vi} = x_{vi}$, etc. to avoid excessive notation.

Complete data is defined as observed data plus latent data. Ignoring the probability $\Pr(\mathbf{x}_v|\mathbf{s}_v)$ that is independent of the parameters, the logarithm of the complete data likelihood for a respondent is given by:

$$\begin{aligned} Cl_v &\equiv \ln \phi(\theta_v; \sigma_\theta^2) + \ln \Pr(\mathbf{s}_v|\theta_v; \lambda_m) \qquad (27) \\ &= \ln \phi(\theta_v; \sigma_\theta^2) + \theta_v s_v^+ - \sum_{i=1}^{I}\sum_{j=1}^{J_i} \zeta_{ij} s_{vij} - \sum_{i=1}^{I}\sum_{j=1}^{J_i} \ln\left[1 + \exp(\theta_v - \zeta_{ij})\right], \end{aligned}$$

where $\phi(\theta_v; \sigma_\theta^2)$ denotes the normal p.d.f. with mean zero and variance σ_θ^2, and $\lambda_m \equiv (\lambda_1, ..., \lambda_I)$, where $\lambda_i \equiv (\zeta_{i1}, ..., \zeta_{iJ_i})$ contains the location parameters of item i. Note that Cl_v is independent of $\mathbf{x}_v$. Respondents are assumed to be independent (i.e., cribbing is not allowed) so that the complete-data loglikelihood $Cl_N = \sum_{v=1}^{N} Cl_v$.

The EM-algorithm (Dempster, et al., 1977) entails the maximization of the conditional expectation of the complete-data likelihood over the distribution of the latent data, given the observed data and a preliminary estimate of the parameters. This conditional expectation is called the Q-function. Determining the Q-function is called the E (expectation) step. Finding parameter values that maximize the Q-function is called the M (maximization) step. Iteration of these steps then yields the EM-algorithm. The EM-algorithm continues until the change between the parameter values from the previous and the present iteration is considered small enough. It can be shown that the resulting estimates are marginal maximum likelihood estimates and we assume that these are consistent. Estimation of latent abilities is discussed in Verstralen (1997). Note that the extension to incomplete designs and multiple populations is straightforward, but we have yet to decide which formulation is most efficient for presentation and/or programming.

6.2. The E-step

Let $h(\mathbf{s}, \theta|\mathbf{x}; \lambda_0) = h(\theta|\mathbf{s}; \lambda_0)h(\mathbf{s}|\mathbf{x}; \lambda_0)$ denote the distribution of the latent data given the observed data, where $\theta = (\theta_1, \ldots, \theta_N)$.[1] Let $Q(\lambda; \lambda_0)$ denote the

[1] Verstralen (1997), and Verstralen and Verhelst (1998) factor $h(\mathbf{s}, \theta|\mathbf{x}; \lambda_0)$ as $h(\mathbf{s}|\theta, \mathbf{x}; \lambda_0)h(\theta|\mathbf{x}; \lambda_0)$. This leads to a different EM-algorithm.

Q-function. By definition,

$$Q(\lambda;\lambda_0) \equiv E\{Cl_N|\mathbf{x};\lambda_0\} \tag{28}$$

$$= \sum_{\mathbf{s}} \int_{\theta_1} \cdots \int_{\theta_N} Cl_N h(\mathbf{s},\theta|\mathbf{x}_v;\lambda_0) d\theta_1 \cdots d\theta_N$$

$$= \sum_{\mathbf{s}} \left[\int_{\theta_1} \cdots \int_{\theta_N} Cl_N h(\theta|\mathbf{s};\lambda_0) d\theta_1 \cdots d\theta_N \right] h(\mathbf{s}|\mathbf{x};\lambda_0)$$

$$= E\{E[Cl_N|\,\mathbf{s};\lambda_0]|\,\mathbf{x};\lambda_0\}. \tag{29}$$

where λ_0 denotes a preliminary estimate of $\lambda \equiv (\lambda_m, \sigma_\theta^2)$. Note that $E[Cl_N|\,\mathbf{s};\lambda_0]$ is the Q-function in the EM-algorithm for the marginal Rasch model. It is known (e.g., Glas, 1989) that

$$E[Cl_N|\,\mathbf{s};\lambda_0] = \sum_{v=1}^{N} \left[E[\vartheta(\theta)|s_v^+;\lambda_0] - \sum_{i=1}^{I} \sum_{j=1}^{J_i} \zeta_{ij} s_{vij} \right], \tag{30}$$

where $\vartheta(\theta) \equiv Cl_v + \sum_i \sum_j \zeta_{ij} s_{vij} = \ln \phi(\theta;\sigma_\theta^2) + \theta s_v^+ - \sum_{i=1}^{I} \sum_{j=1}^{J_i} \ln[1+\exp(\theta-\zeta_{ij})]$, and $E[\vartheta(\theta)|s_v^+;\lambda_0]$ is the expectation of $\vartheta(\theta)$ taken over the posterior of θ given s_v^+: that is,

$$h(\theta|s_v^+;\lambda_0) = \frac{\Pr(s_v^+|\theta;\lambda_{m,0})\phi(\theta;\sigma_{\theta,0}^2)}{\int \Pr(s_v^+|\theta;\lambda_{m,0})\phi(\theta;\sigma_{\theta,0}^2)d\theta}. \tag{31}$$

Since $\Pr(S_{ij}=1|\theta)$ is modelled as a Rasch model, the probability $\Pr(s_v^+|\theta;\lambda_{m,0})$ has a well-known expression.

Remark 3. *Equation (30) is well-known but it may be helpful to know how it can be derived. For simplicity, assume that there are only two respondents. Hence,*

$$E[Cl_N|\,\mathbf{s};\lambda_0] = \int_{\theta_1}\int_{\theta_2} Cl_N h(\theta_1,\theta_2|\mathbf{s};\lambda_0) d\theta_2 d\theta_1$$

$$= \int_{\theta_1}\int_{\theta_2} (Cl_1 + Cl_2) h(\theta_1,\theta_2|\mathbf{s};\lambda_0) d\theta_2 d\theta_1$$

$$= \int_{\theta_1}\int_{\theta_2} (Cl_1 h(\theta_1,\theta_2|\mathbf{s};\lambda_0) + Cl_2 h(\theta_1,\theta_2|\mathbf{s};\lambda_0)) d\theta_2 d\theta_1$$

$$= \int_{\theta_1}\int_{\theta_2} Cl_1 h(\theta_1,\theta_2|\mathbf{s};\lambda_0) d\theta_2 d\theta_1 + \int_{\theta_1}\int_{\theta_2} Cl_2 h(\theta_1,\theta_2|\mathbf{s};\lambda_0) d\theta_2 d\theta_1$$

Since Cl_v is a function of θ_v only we may write

$$\begin{aligned}
&\int_{\theta_1} Cl_1 \int_{\theta_2} h(\theta_1,\theta_2|\mathbf{s};\lambda_0)d\theta_2 d\theta_1 + \int_{\theta_2} Cl_2 \int_{\theta_1} h(\theta_1,\theta_2|\mathbf{s};\lambda_0)d\theta_1 d\theta_2 \\
&= \int_{\theta_1} Cl_1 h(\theta_1|\mathbf{s};\lambda_0)d\theta_1 + \int_{\theta_2} Cl_2 h(\theta_2|\mathbf{s};\lambda_0)d\theta_2 \\
&= \sum_v \int_\theta Cl_v h(\theta|\mathbf{s};\lambda_0)d\theta \\
&= \sum_v \left[\int_\theta \vartheta(\theta)h(\theta|\mathbf{s};\lambda_0)d\theta - \sum_{i=1}^{I}\sum_{j=1}^{J_i}\zeta_{ij}s_{vij}\right]
\end{aligned}$$

Note that $\vartheta(\theta)$ only depends on s_v^+. Since respondents are independent, $h(\theta|\mathbf{s};\lambda_0)$ may be replaced by $h(\theta|\mathbf{s}_v;\lambda_0)$. Further, since s_v^+ is sufficient for θ, $h(\theta|\mathbf{s}_v;\lambda_0)$ can be replaced by $h(\theta|s_v^+;\lambda_0)$. The result is Equation (30).

Using (30) the Q-function may be written as:

$$\sum_{v=1}^{N}\sum_{s_v^+}\left[E[\vartheta(\theta)|s_v^+;\lambda_0] - \sum_{i=1}^{I}\sum_{j=1}^{J_i}\zeta_{ij}\Pr(S_{vij}=1|s_v^+,\mathbf{x}_v)\right]\Pr(s_v^+|\mathbf{x}_v;\lambda_0), \tag{32}$$

where

$$\Pr(S_{vij}=1|s_v^+,\mathbf{x}_v) = \left(1+\frac{\Pr(\mathbf{x}_v|S_{vij}=0,s_v^+)\Pr(S_{vij}=0|s_v^+)}{\Pr(\mathbf{x}_v|S_{vij}=1,s_v^+)\Pr(S_{vij}=1|s_v^+)}\right)^{-1}, \tag{33}$$

$$\Pr(s_v^+|\mathbf{x}_v;\lambda_0) = \frac{\Pr(\mathbf{x}_v|s_v^+)\int\Pr(s_v^+|\theta;\lambda_{m,0})\phi(\theta;\sigma^2_{\theta,0})d\theta}{\sum_h \Pr(\mathbf{x}_v|S_v^+=h)\int\Pr(S_v^+=h|\theta;\lambda_{m,0})\phi(\theta;\sigma^2_{\theta,0})d\theta}, \tag{34}$$

$$\Pr(\mathbf{x}_v|S_{vij}=u,s_v^+) = \sum_{\mathbf{s}_v|s_v^+} s_{vij}^u(1-s_{vij})^{1-u}\Pr(\mathbf{x}_v|\mathbf{s}_v), \text{ and} \tag{35}$$

$$\Pr(\mathbf{x}_v|s_v^+) = \sum_{\mathbf{s}_v|s_v^+}\Pr(\mathbf{x}_v|\mathbf{s}_v). \tag{36}$$

The symbol $\sum_{\mathbf{s}_v|s_v^+}$ denotes summation over all possible subset vectors $\mathbf{s}_v$ whose entries sum to s_v^+ and $\sum_{s_v^+}$ summation over all values of s_v^+; i.e., from 0 to $\sum_{i=1}^{I} J_i$. Details of the derivations are presented in the Appendix. Note that

$$\Pr(\mathbf{x}_v|\mathbf{s}_v) = \prod_i \Pr(x_{vi}|\mathbf{s}_{vi}) = \prod_i \frac{1-s_{vi(x_{vi})}}{v(s_{vi}^+)}. \tag{37}$$

Hence, $\Pr(\mathbf{x}_v|s_v^+)$ is an elementary symmetric function of order s_v^+ with argument $\mathbf{s}_{vi}$. The probability $\Pr(S_{vij}=1|s_v^+;\lambda_{m,0})$ equals the conditional probability of a correct response on item (i,j) given a sum score of s_v^+ under the Rasch model, which has a well-known expression that was given earlier.

6.3. The M-step

Let ∂_{ih} and $\partial_{\sigma_\theta^2}$ denote differentiation with respect to ζ_{ij} and σ_θ^2, respectively. The symbol ∂ is used for ∂_{ih} or $\partial_{\sigma_\theta^2}$. In the M-step, the values of the parameters are changed in such a way that the value of the Q-function increases (Tanner, 1993, p.43). For the item parameters this is done with a single Newton-Raphson (NR) step. That is,

$$\lambda_{ih} = \lambda_{ih,0} - \frac{\partial_{ih}Q(\lambda_0;\lambda_0)}{\partial_{ih}^2Q(\lambda_0;\lambda_0)}, \tag{38}$$

where subscript 0 now denotes a parameter value from the previous M-step. Using (32) it is straightforward to derive that $\partial_{ih}Q(\lambda;\lambda_0)$ equals

$$\sum_{v=1}^{N}\sum_{s_v^+}\left[E[\Pr(S_{ih}=1|\theta;\zeta_{ij})|s_v^+;\lambda_0]-\Pr(S_{vih}=1|s_v^+,\mathbf{x}_v)\right]\Pr(s_v^+|\mathbf{x}_v;\lambda_0). \tag{39}$$

The second order derivative is

$$\partial_{ih}^2Q(\lambda;\lambda_0) = -\sum_{v=1}^{N}\sum_{s_v^+}E[Var(S_{ih}|\theta;\zeta_{ij})|s_v^+;\lambda_0)]\Pr(s_v^+|\mathbf{x}_v;\lambda_0). \tag{40}$$

Second order derivatives with respect to different item parameters are zero and this enables us to do a separate M-step for each item parameter.

Remark 4. *The derivatives should look familiar to those who use the Rasch model. First, it is clear that*

$$\partial_{ih}Q(\lambda;\lambda_0) =$$

$$=\sum_{v=1}^{N}\sum_{s_v^+}\left[\partial_{ih}E[\vartheta(\theta)|s_v^+;\lambda_0]-\Pr(S_{vih}=1|s_v^+,\mathbf{x}_v)\right]\Pr(s_v^+|\mathbf{x}_v;\lambda_0),$$

where $\partial_{ih} E[\vartheta(\theta)|s_v^+;\lambda_0]$ *equals*

$$\begin{aligned}
&\partial_{ih}\int\left(\ln\phi(\theta;\sigma_\theta^2)+\theta s_v^+-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\ln\left[1+\exp(\theta-\zeta_{ij})\right]\right)h(\theta|s_v^+;\lambda_0)d\theta\\
&=-\int h(\theta|s_v^+;\lambda_0)\sum_{i=1}^{I}\sum_{j=1}^{J_i}\partial_{ih}\ln\left[1+\exp(\theta-\zeta_{ij})\right]d\theta\\
&=-\int\partial_{ih}\ln\left[1+\exp(\theta-\zeta_{ih})\right]h(\theta|s_v^+;\lambda_0)d\theta\\
&=\int\Pr(S_{ih}=1|\theta;\zeta_{ih})h(\theta|s_v^+;\lambda_0)d\theta\\
&=E[\Pr(S_{ih}=1|\theta;\zeta_{ij})|s_v^+;\lambda_0]
\end{aligned}$$

The same derivations must also be made when the marginal Rasch model is estimated by means of an EM-algorithm.

A NR step is not required to update σ_θ^2. Straightforward differentiation shows that

$$\partial_{\sigma_\theta^2}Q(\lambda;\lambda_0)=\frac{1}{2\sigma_\theta^4}\sum_{v=1}^{N}\sum_{s_v^+}\Pr(s_v^+|\mathbf{x}_v;\lambda_0)\left[E[\theta^2|s_v^+;\lambda_0]-\sigma_\theta^2\right] \tag{41}$$

$$=\frac{1}{2\sigma_\theta^4}\left[\sum_{v=1}^{N}\sum_{s_v^+}\Pr(s_v^+|\mathbf{x}_v;\lambda_0)E[\theta^2|s_v^+;\lambda_0]-\sigma_\theta^2 N\right]. \tag{42}$$

It follows from (42) that $\partial_{\sigma_\theta^2}Q(\lambda;\lambda_0)=0$ if

$$\sigma_\theta^2=\frac{1}{N}\sum_{v=1}^{N}\sum_{s_v^+}\Pr(s_v^+|\mathbf{x}_v;\lambda_0)E[\theta^2|s_v^+;\lambda_0]. \tag{43}$$

As expected from (29), the estimation equations in the M-step have the form $E\{\partial E[Cl_N|\mathbf{s};\lambda_0]|\mathbf{x};\lambda_0\}=0$, where $\partial E[Cl_N|\mathbf{s};\lambda_0]=0$ are the corresponding equation for the marginal Rasch model. We hope that the reader who knows the Rasch model has recognized the form of the derivatives. Routines for the marginal Rasch model may be used to calculate $\partial E[Cl_N|\mathbf{s};\lambda_0]$. The main technical problem here is

the numerical approximation to the integrals in expressions of the form

$$E[g(\theta)|s_v^+] = \int g(\theta)h(\theta|s_v^+;\lambda_0)d\theta \qquad (44)$$
$$= \frac{\int g(\theta)\Pr(s_v^+|\theta)\phi(\theta)d\theta}{\int \Pr(s_v^+|\theta)\phi(\theta)d\theta},$$

where $g(\theta)$ is some smooth function of θ. It is customary to use Gauβian quadrature to this aim. An alternative approximation, called the Laplace method, is briefly discussed in the appendix. We are currently investigating whether the normal distribution can be replaced by a similar distribution that would give closed form expressions for the integrals (Maris, *in preparation*).

6.4. Approximating the Observed Data Information Matrix

From (39) and (41) it can be seen that $\partial_{ih}Q(\lambda;\lambda_0)$ and $\partial_{\sigma_\theta^2}Q(\lambda;\lambda_0)$ can be written as $\sum_{v=1}^N \partial_{ih}Q_v(\lambda;\lambda_0)$ and $\sum_{v=1}^N \partial_{\sigma_\theta^2}Q_v(\lambda;\lambda_0)$, respectively. Let $Inf(\hat{\lambda},\hat{\lambda})$ denote an approximation to the information matrix with elements:

$$Inf(\hat{\zeta}_{ih},\hat{\zeta}_{jk}) = \frac{1}{N}\sum_{v=1}^{N}\partial_{ih}Q_v(\hat{\lambda};\hat{\lambda})\partial_{jk}Q_v(\hat{\lambda};\hat{\lambda}), \qquad (45)$$

$$Inf(\hat{\zeta}_{ih},\widehat{\sigma_\theta^2}) = \frac{1}{N}\sum_{v=1}^{N}\partial_{ih}Q_v(\hat{\lambda};\hat{\lambda})\partial_{\sigma_\theta^2}Q_v(\hat{\lambda};\hat{\lambda}), \text{ and} \qquad (46)$$

$$Inf(\widehat{\sigma_\theta^2},\widehat{\sigma_\theta^2}) = \frac{1}{N}\sum_{v=1}^{N}\left(\partial_{\sigma_\theta^2}Q_v(\hat{\lambda};\hat{\lambda})\right)^2. \qquad (47)$$

The hats indicate that the quantities are evaluated at the final estimates. The hats indicate that the quantities are evaluated at the final estimates (see also Redner and Walker, 1984; Meilijson, 1989; Friedl and Gauermann, 2000). To see why this approximation works, one should first note that $\partial Q_v(\hat{\lambda};\hat{\lambda})$ equals $\partial \ln \Pr(\mathbf{x}_v;\hat{\lambda})$ which is the first-order derivative of the marginal log-likelihood. Second, $Inf(\hat{\zeta}_{ih},\hat{\zeta}_{jk})$, for instance, is calculated as $\frac{1}{N}\sum_{v=1}^N\left(\partial_{ih}\ln\Pr(\mathbf{x}_v;\hat{\lambda})\right)^2$ and this is a consistent estimate of $E[(\partial_{ih}\ln\Pr(\mathbf{x}_v))^2;\lambda]$; by definition the information with respect to $\hat{\zeta}_{ih}$. It can be shown that this approximation becomes equivalent to the one proposed by Louis

(1982) if N becomes large.

The diagonal elements of the inverse of $Inf(\hat{\lambda},\hat{\lambda})$ can be used to calculate approximate confidence intervals of the parameters. Specifically, if $Inf(\hat{\lambda},\hat{\lambda})^{-1}_{ih,ih}$ denotes a diagonal entry of the inverse of $Inf(\hat{\lambda},\hat{\lambda})$ corresponding to ζ_{ih}, an approximate 95% confidence interval is

$$\hat{\zeta}_{ih} - 1.96\sqrt{\frac{Inf(\hat{\lambda},\hat{\lambda})^{-1}_{ih,ih}}{N}} < \zeta_{ih} < \hat{\zeta}_{ih} + 1.96\sqrt{\frac{Inf(\hat{\lambda},\hat{\lambda})^{-1}_{ih,ih}}{N}}. \tag{48}$$

The approximation may also be used in Lagrange multiplier (LM) tests (Rao,1947; Aitchison and Silvey, 1958) for the NM against more general alternatives. Note that a LM test may not be used to test for unity of attractiveness parameters since the model with attractiveness parameters is not identifiable.

7. A Test Based Upon MOR

Consider a test with I items. Let $\mathbf{x}_{-i}$ denote a vector of item responses except for the ith item and $\#0_{-i}$ denote the number of correct responses in $\mathbf{x}_{-i}$.

lemma 2. *Let $s_2 > s_1$. Under the NM*

$$E[f(\theta)|\#0 = s_2] \geq E[f(\theta)|\#0 = s_1] \tag{49}$$

for all increasing functions f

Proof. With binary scoring, the NM is monotone, unidimensional and responses to different items are independent given θ. It then follows from Theorem 2 in Grayson (1988) that $\#0_{-i}$ has MLR; that is,

$$\frac{\Pr(\#0_{-i} = s_2|\theta)}{\Pr(\#0_{-i} = s_1|\theta)}$$

is non-decreasing in θ if $s_2 > s_1$. MLR implies SOL in $\#0_{-i}$ which is equivalent to

$$E[f(\theta)|\#0 = s_2] \geq E[f(\theta)|\#0 = s_1]$$

for all increasing functions f (Ross, 1996, Proposition 9.1.2). □

Let $X_i \in \{t, j\}$ denote the event that a respondent chooses either option t or option j. Note that

$$\begin{aligned} &\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i} = s) \\ &\quad = \int \Pr(X_i = t | X_i \in \{t,j\}, \theta) f(\theta | \#0_{-i} = s) d\theta \\ &\quad = E[\Pr(X_i = t | X_i \in \{t,j\}, \theta) | \#0_{-i} = s]. \end{aligned} \tag{50}$$

If $\zeta_{it} > \zeta_{ij}$, $\Pr(X_i = t | X_i \in \{t,j\}, \theta)$ is increasing in θ and Lemma (2) implies that $\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i})$ is increasing in $\#0_{-i}$. Similarly, $\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i})$ is decreasing in $\#0_{-i}$ when $\zeta_{it} < \zeta_{ij}$, and constant if $\zeta_{it} = \zeta_{ij}$. Thus, the NM is violated if $\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i})$ is a non-monotonic function of $\#0_{-i}$. For later reference this is stated as a theorem.

Theorem 2. *The NM is violated if, for some t and j, $\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i})$ is not a monotonic function of $\#0_{-i}$.*

How may we employ this result to obtain a statistical test for model fit? Consider a table of the following form:

$\#0_{-i}:$	0	1	$\cdots$	I
$X_i = t$	$n(t,0)$	$n(t,1)$	$\cdots$	$n(t,I)$
$X_i = j$	$n(j,0)$	$n(j,1)$	$\cdots$	$n(j,I)$

In this table, $n(t, \#0_{-i})$ denotes the number of respondents that have chosen response t to item i and $\#0_{-i}$ correct responses to the items in $\mathbf{x}_{-i}$. The percentage $\Pr(X_i = t | X_i \in \{t,j\}, \#0_{-i} = s)$ can be consistently estimated by the statistic

$$\frac{n(t,s)}{n(t,s) + n(j,s)} \tag{51}$$

The rank correlation between the ranks of these estimates and $\#0_{-i}$ provides an appropriate statistic to test whether the relation is monotone. Specifically, the rank correlation may be 1, -1 or 0. To limit the number of test we may consider only pairs where $t = 0$ and $j \neq 0$. Then, the rank correlation should be 1 and we obtain a consistent test for rising selection ratios. There are J_i such tests for each item. To avoid chance capitalization it is recommended to look only at items and alternatives that are *a priory* considered suspicious. Further research is directed at finding ways to thicken the information in the tables and obtain a test for the model as a whole.

8. The Nedelsky Model as a Signal Detection Model

Consider a MC item that consists of a "signal" (or stimulus), an instruction to respond, and a number of alternative answers that give interpretations of the signal one of which is the correct interpretation. The signal can take many forms; for example, a picture, a speech fragment, a newspaper article, etc. It is characterized by a number of key properties or "content elements" (CEs) that respondents must recognize in order interpret the signal correctly and be certain to choose the correct answer. Each alternative answer provides an interpretation of the signal. The correct alternative has all relevant properties of the signal while the item writers have been careful to ensure that distractors lack one or more.

Consider, for example, an item showing a picture of a car on a traffic circle with his right blinker on; this is the signal. This item is administered to people who apply for a drivers licence. Respondents are instructed to choose the alternative that provides a correct interpretation of the situation. Response alternatives are, for instance, a) the driver may turn off the traffic circle, b) the driver should give priority to the cyclist, etc. A similar situation arises in examinations for vine tasters, where the stimulus would be say a white wine produced in the hills around the Musel

river and examinees are supposed to recognized its particular taste.

Using the basic notion of the NM, it is assumed that an alternative is left out of consideration if a respondent recognizes any CE that is missing. That is, we interpret $\Pr(S_{ij} = 1|\theta)$ as the probability that the respondent recognizes at least one of the CEs that are missing in alternative answer j. A respondent who recognizes all CEs will consequently reject all distractors and choose the correct alternative with probability 1.

We ignore the item index for a moment and introduce some notation. Assume that there are c CEs and each CE is indexed with f, where $f = 1, ..., c$. Let $(E_f = 1)$ denote the event that CE f is recognized. Further, define T_{jf} as an indicator of missing CEs; that is, T_{jf} equals 1 if CE f is missing in alternative j, and 0 otherwise. Since the correct alternative has all relevant properties, $T_{0f} = 0$ for all f. The T_{jf} may be considered as the entries of a matrix which represents the *content structure of the item.*

It follows that

$$\Pr(S_j = 1|\theta) = \Pr(\oplus T_{jf}E_f = 1|\theta). \tag{52}$$

where $\oplus$ denotes the Boolean sum (OR) over f, and $\oplus T_{jf}E_f = 1$ the event that one or more of the missing CEs are recognized. Since there are no CEs missing in the correct alternative, it cannot be excluded and $\Pr(S_0 = 1|\theta) = \Pr(0 = 1|\theta) = 0$.

Note that $S_j = 0$ if none of the missing CE is recognized. Since $S_j = 0$ is the negation of $S_j = 1$, it follows from De Morgan's Laws that

$$\Pr(S_j = 0|\theta) = \Pr(\otimes(1 - T_{jf}E_f) = 1|\theta) \tag{53}$$

where $\otimes$ is the Boolean product (AND) over f, and $\otimes(1 - T_{jf}E_f) = 1$ denotes the event that none of the missing CEs is recognized. This probability simplifies to $\Pr(S_j = 0|\theta) = \prod_f \Pr(E_f = 0|\theta)^{T_{jf}}$ if we assume that respondents evaluate each CE

independently. Given this assumption, it is readily seen that the NM equals a signal detection model when the content structure of the item is such that

$$\Pr(S_j = 0|\theta) = \Pr(E_f = 0|\theta), \tag{54}$$

for each $j = 1, \ldots, J_i$, and $\Pr(E_j = 0|\theta)$ is modelled as in (1). This means that there is only one CE missing in each distractor. Consider, for example, an item where respondents are assumed simply to "see" the correct answer or not. Hence, there is a single CE, $T_{j1} = 1$ for $j = 1, ..., J_i$, and $\Pr(S_j = 0|\theta) = \Pr(E_1 = 0|\theta)$. It is seen that all option location parameters are equal if we apply the NM to an item with this content structure.

The development of the signal detection version of the NM is a topic for future research. A first attempt is described in the next chapter. Among other things, it will be shown that Bock's (1972) nominal response model can be considered a signal detection NM model.

9. Conclusion

The NM is a restrictive model based upon a simple theory about the response process. The advantage of a theory-based model is that the theory guides the interpretation of the parameter estimates. Estimated abilities may, for example, be used to predict which alternatives respondents reject. The advantages of a simple model are clear. Simulation studies by Revuelta (2000) show quite dramatically that the inclusion of option-specific discrimination parameters, for instance, would require huge numbers of respondents. On the other hand, a parsimonious model, however beautiful, may be inadequate for many (if not all!) data and should be tested before it is accepted. We have discussed a possible test but it is clear that further tests must be developed before the NM could be brought in for our daily work. Note that, even when the model fits the data, the process interpretation need not be valid. Farr,

Pritchard and Smitten (1990), for instance, found no evidence in support of the theory underpinning the NM.

In the introduction it was noted that binary scoring entails loss of information. Using the missing information principle (Orchard and Woodbury, 1972; Louis, 1982), it has been shown that binary scoring will indeed provide less precision in estimated abilities than option scoring, provided the distractors differ in difficulty. It was also shown that precision could be increased further if we could somehow entice respondents to reveal their latent subsets. This observation led Verstralen and Verhelst (2000) to develop a model for subjective probabilities given by respondents to each of the alternative answers. Their work shows that careful construction of MC items may be a good way to enhance the precision of estimation. We believe that, speaking in general terms, "richer data" is key to estimate more complex models and make more specific statements about test-takers. The same purpose is served by the signal detection NM which makes use of the content structure of the items. To get the latent subsets we might consider, for instance, offering the alternatives one by one and ask respondents if they think that the alternative is incorrect. It is a small step from here to ask them to provide a subjective probability. In the NM the subjective probabilities are implicitly assumed to be either 0 or 1 divided by the number of alternatives minus the alternatives considered wrong.

Finally, although we have described an estimation procedure we have not formally proven that the NM is identifiable and the regularity conditions required for consistency are met. These are essential issues that must be addressed in future research.

10. Appendix:

10.1. Deriving $\frac{\partial}{\partial\theta}\psi_{itj}(\theta)$ in Theorem (1)

First,

$$\begin{aligned}
\frac{\partial}{\partial\theta}\Pr(X_i = j|\theta) &= \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \frac{\partial}{\partial\theta}\Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \\
&= \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \frac{\partial}{\partial\theta} \ln \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \\
&= \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \left(s_i^+ - E[S_i^+|\theta]\right) \\
&= \sum_{\mathbf{s}_i} \frac{1 - s_{ij}}{v(s_i^+)} \Pr(\mathbf{S}_i = \mathbf{s}_i|\theta) \left(E[v(S_i^+)|\theta] - v(s_i^+)\right) \\
&= E[v(S_i^+)|\theta] \Pr(X_i = 0|\theta) - 1 + \Pr(S_{ij} = 1|\theta).
\end{aligned}$$

The fourth equality holds since $E[v(S_i^+)|\theta] - v(s_i^+) = J_i - E[S_i^+|\theta] + 1 - J_i + s_i^+ - 1 = s_i^+ - E[S_i^+|\theta]$. Now,

$$\frac{\partial}{\partial\theta}\psi_{itj}(\theta) = \frac{\frac{\partial}{\partial\theta}\Pr(X_i = t|\theta)\Pr(X_i = j|\theta) - \Pr(X_i = t|\theta)\frac{\partial}{\partial\theta}\Pr(X_i = j|\theta)}{\Pr(X_i = j|\theta)^2}$$

which simplifies to the expression in the proof of Theorem (1).

10.2. Simplifying the Q-function

Let $\vartheta(\theta, s_v^+) \equiv \left[\ln\phi(\theta; \sigma_\theta^2) + \theta s_v^+ - \sum_i \sum_j \ln\left[1 + \exp(\theta - \zeta_{ij})\right]\right]$. Then, using (30), the Q-function can be written as:

$$\sum_{\mathbf{s}} h(\mathbf{s}|\mathbf{x}; \lambda_0) \sum_v \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+; \lambda_0) d\theta - \sum_{\mathbf{s}} h(\mathbf{s}|\mathbf{x}; \lambda_0) \sum_v \sum_i \sum_j \zeta_{ij} s_{vij}$$

Let $\mathbf{s}^{(-v)}$ denote the latent subsets without $\mathbf{s}_v$; the subset vector of respondent v.

$$
\begin{aligned}
&\sum_v \sum_{\mathbf{s}} h(\mathbf{s}|\mathbf{x};\lambda_0) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \\
&= \sum_v \sum_{\mathbf{s}_v} \sum_{\mathbf{s}^{(-v)}} \Pr(\mathbf{s}^{(-v)}|\mathbf{s}_v, \mathbf{x};\lambda_0) \Pr(\mathbf{s}_v|\mathbf{x}) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \\
&= \sum_v \sum_{\mathbf{s}_v} \Pr(\mathbf{s}_v|\mathbf{x}) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \left[\sum_{\mathbf{s}^{(-v)}} \Pr(\mathbf{s}^{(-v)}|\mathbf{s}_v, \mathbf{x};\lambda_0) \right] \\
&= \sum_v \sum_{\mathbf{s}_v} \Pr(\mathbf{s}_v|\mathbf{x}) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \\
&= \sum_v \sum_{s_v^+} \sum_{\mathbf{s}_v|s_v^+} \Pr(s_v^+|\mathbf{x}_v) \Pr(\mathbf{s}|s_v^+, \mathbf{x}_v) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \\
&= \sum_v \sum_{s_v^+} \Pr(s_v^+|\mathbf{x}_v) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta \left[\sum_{\mathbf{s}_v|s_v^+} \Pr(\mathbf{s}|s_v^+, \mathbf{x}_v) \right] \\
&= \sum_v \sum_{s_v^+} \Pr(s_v^+|\mathbf{x}_v) \int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta.
\end{aligned}
$$

Further,

$$\sum_v \sum_{\mathbf{s}} h(\mathbf{s}|\mathbf{x};\lambda_0) \sum_i \sum_j \zeta_{ij} s_{vij}$$

$$= \sum_v \sum_{\mathbf{s}_v} \Pr(\mathbf{s}_v|\mathbf{x}_v) \sum_i \sum_j \zeta_{ij} s_{vij}$$

$$= \sum_v \sum_{\mathbf{s}_v} \sum_i \sum_j \Pr(\mathbf{s}_v|\mathbf{x}_v) \zeta_{ij} s_{vij}$$

$$= \sum_v \sum_i \sum_j \zeta_{ij} \sum_{\mathbf{s}_v} \Pr(\mathbf{s}_v|\mathbf{x}_v) s_{vij}$$

$$= \sum_v \sum_i \sum_j \zeta_{ij} \sum_{s_v^+} \sum_{\mathbf{s}_v|s_v^+} \Pr(\mathbf{s}_v|s_v^+, \mathbf{x}_v) \Pr(s_v^+|\mathbf{x}_v) s_{vij}$$

$$= \sum_v \sum_i \sum_j \zeta_{ij} \sum_{s_v^+} \Pr(s_v^+|\mathbf{x}_v) \left[\sum_{\mathbf{s}_v|s_v^+} \Pr(\mathbf{s}_v|s_v^+, \mathbf{x}_v) s_{vij} \right]$$

$$= \sum_v \sum_i \sum_j \zeta_{ij} \sum_{s_v^+} \Pr(s_v^+|\mathbf{x}_v) \Pr(S_{vij} = 1|s_v^+, \mathbf{x}_v)$$

$$= \sum_v \sum_{s_v^+} \sum_i \sum_j \zeta_{ij} \Pr(S_{vij} = 1|s_v^+, \mathbf{x}_v) \Pr(s_v^+|\mathbf{x}_v)$$

It follows that the Q-function may thus be written as

$$\sum_v \sum_{s_v^+} \left[\int_\theta \vartheta(\theta, s_v^+) h(\theta|s_v^+;\lambda_0) d\theta - \sum_i \sum_j \zeta_{ij} \Pr(S_{vij} = 1|s_v^+, \mathbf{x}_v) \right] \Pr(s_v^+|\mathbf{x}_v)$$

It remains to work out the probabilities $\Pr(S_{vij} = 1|s_v^+, \mathbf{x}_v)$ and $\Pr(s_v^+|\mathbf{x}_v)$. First,

$$\begin{aligned}
\Pr(S_{vij}=1|s_v^+,\mathbf{x}_v) &= \frac{\Pr(S_{vij}=1,s_v^+,\mathbf{x}_v)}{\Pr(s_v^+,\mathbf{x}_v)} \\
&= \frac{\Pr(S_{vij}=1,s_v^+,\mathbf{x}_v)}{\Pr(\mathbf{x}_v,s_v^+,S_{vij}=1)+\Pr(\mathbf{x}_v,s_v^+,S_{vij}=0)} \\
&= \frac{1}{1+\frac{\Pr(\mathbf{x}_v,s_v^+,S_{vij}=0)}{\Pr(\mathbf{x}_v,s_v^+,S_{vij}=1)}} \\
&= \frac{1}{1+\frac{\Pr(\mathbf{x}_v|S_{vij}=0,s_v^+)\Pr(S_{vij}=0|s_v^+)\Pr(s_v^+)}{\Pr(\mathbf{x}_v|S_{vij}=1,s_v^+)\Pr(S_{vij}=1|s_v^+)\Pr(s_v^+)}} \\
&= \frac{1}{1+\frac{\Pr(\mathbf{x}_v|S_{vij}=0,s_v^+)\Pr(S_{vij}=0|s_v^+)}{\Pr(\mathbf{x}_v|S_{vij}=1,s_v^+)\Pr(S_{vij}=1|s_v^+)}},
\end{aligned}$$

where

$$\Pr(\mathbf{x}_v|S_{vij}=1,s_v^+)=\sum_{\mathbf{s}_v|s_v^+} s_{vij}\Pr(\mathbf{x}_v|\mathbf{s}_v), \text{ and}$$

$$\Pr(\mathbf{x}_v|S_{vij}=0,s_v^+)=\Pr(\mathbf{x}_v|s_v^+)-\Pr(\mathbf{x}_v|S_{vij}=1,s_v^+)$$

which leads to the expression in the text. Finally, $\Pr(s_v^+|\mathbf{x}_v)=\frac{\Pr(\mathbf{x}_v|s_v^+)\Pr(s_v^+)}{\Pr(\mathbf{x}_v)}$.

10.3. Laplace Approximation

An important technical issue involved in the EM-algorithm is the evaluation of a ratio of integrals of the form:

$$\begin{aligned}
E[g(\theta)|s_v^+] &= \int g(\theta)h(\theta|s_v^+;\lambda_0)d\theta \\
&= \frac{\int g(\theta)\Pr(s_v^+|\theta)\phi(\theta)d\theta}{\int \Pr(s_v^+|\theta)\phi(\theta)d\theta}
\end{aligned}$$

where $g(\theta)$ is a function of θ, and $\phi(\theta)$ the p.d.f. of the normal distribution with zero mean. We see that $E[g(\theta)|s_v^+]$ has the form of a ratio of two univariate integrals.

This ratio can be written as

$$\frac{\int g(\theta)\exp\{\ln[\Pr(s_v^+|\theta)\phi(\theta)]\}\,d\theta}{\int \exp\{\ln[\Pr(s_v^+|\theta)\phi(\theta)]\}\,d\theta} \tag{55}$$

$$= \frac{\int g(\theta)\exp\{-L(\theta)\}\,d\theta}{\int \exp\{-L(\theta)\}\,d\theta}. \tag{56}$$

where $\exp\{-L(\theta)\} = \Pr(s_v^+|\theta)\phi(\theta)$ is the posterior of θ given s_v^+ with a maximum at $\hat{\theta}$. The idea is to expand around $\hat{\theta}$ to obtain:

$$\begin{aligned}
\int g(\theta)\exp\{-L(\theta)\}\,d\theta &\approx \int g(\hat{\theta})\exp\left\{-\left[L(\hat{\theta}) + (\theta-\hat{\theta})L'(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}L''(\hat{\theta})\right]\right\}d\theta \\
&= g(\hat{\theta})\int \exp\left\{-\left[L(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}L''(\hat{\theta})\right]\right\}d\theta \\
&= g(\hat{\theta})\exp\left(-L(\hat{\theta})\right)\int \exp\left\{-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}\right\}d\theta \\
&= g(\hat{\theta})\exp\left(-L(\hat{\theta})\right)
\end{aligned}$$

where $\sigma^2 = \left(L''(\hat{\theta})\right)^{-1}$. Intuitively, if $\exp(-L(\theta))$ is very peaked around $\hat{\theta}$ the integral can be well approximated by the behavior of the integrand near $\hat{\theta}$. This is the case when there are many items. Formally, it was demonstrated by Tierney and Kadane (1986) the error of approximation is of order $O((\sum_i J_i)^{-2})$. It follows that

$$\begin{aligned}
\frac{\int g(\theta)\exp\{-L(\theta)\}\,d\theta}{\int \exp\{-L(\theta)\}\,d\theta} &\approx \frac{g(\hat{\theta})\exp\left(-L(\hat{\theta})\right)}{\exp\left(-L(\hat{\theta})\right)} \\
&= g(\hat{\theta})
\end{aligned}$$

Thus, the important part is to find $\hat{\theta}$ that maximizes $\ln[\Pr(s_v^+|\theta)\phi(\theta)]$. (See also Tanner, 1993, p. 24-25).

If we apply this scheme in the present situation, we find that Laplace's method

(Tanner, 1993) gives the following rough approximations:

$$\Pr(s_v^+|\mathbf{x}_v;\lambda_0) \approx \frac{\Pr(\mathbf{x}_v|s_v^+)\exp(-\hat{\theta}s_v^+)}{\sum_h \Pr(\mathbf{x}_v|S_v^+=h)\exp(-\hat{\theta}h)},$$

$$E[\Pr(S_{ih}=1|\theta)|s_v^+;\lambda_0] \approx \Pr(S_{ih}=1|\hat{\theta}), \text{ and}$$

$$E[Var(S_{ih}|\theta)|s_v^+;\lambda_0) \approx \Pr(S_{ih}=1|\hat{\theta})(1-\Pr(S_{ih}=1|\hat{\theta})),$$

where $\hat{\theta}$ is the unique maximum of the function $s_v^+\theta - \sum_i \sum_j \ln[1+\exp(\theta-\zeta_{ij})] - \frac{1}{2}\left(\frac{\theta}{\sigma_{\theta,0}}\right)^2$. This maximum can be found using Newtons method; that is

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{E[S_v^+|\theta] - s_v^+ + \frac{\hat{\theta}_i}{\sigma_{\theta,0}^2}}{Var(S_v^+|\hat{\theta}_i) + \sigma_{\theta,0}^{-2}},$$

where i denotes iteration and iterations continue until $|\hat{\theta}_{i+1} - \hat{\theta}_i|$ is sufficiently small. To start up the iterations at the first iteration of the EM-algorithm we need initial estimates. Here is a suggestion. For $0 < s_v^+ < \sum_i J_i$,

$$\ln\left(\frac{s_v^+}{\sum_i J_i - s_v^+}\right)\frac{1}{\sqrt{1+\left(\frac{\sigma_{\theta,0}}{1.7}\right)^2}}$$

provides a reasonable initial estimate of $\hat{\theta}$ (see Cohen, 1979). For $s_v^+ \in \{0, \sum_i J_i\}$ we use $\hat{\theta}$ from the adjacent sum-score. For further iterations of the EM-algorithm we simply use the estimated values from the previous one.

References

Aitchison, W., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, *28*, 813-828.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading: Addison-Wesley.

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

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *32*, 113-120.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM-algorithm. (with discussion). *Journal of the Royal Statistical Society*, Series B, *39*, 1-38.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple choice reading comprehension test. *Journal of Educational Measurement*, *27*, 209-226.

Fischer, G. H. (1995a). Derivations of the Rasch model. Chapter 2 in "*Rasch models: Foundations, recent developments, and applications* ", edited by G.H. Fischer and I.W. Molenaar. New-York: Springer.

Fischer, G. H. (1995b). The linear logistic test model. Chapter 8 in "*Rasch models: Foundations, recent developments, and applications* ", edited by G.H. Fischer and I.W. Molenaar. New-York: Springer.

Friedl, H. and Kauermann, G. (2000). Standard Errors for EM Estimates for Generalized Linear Models with Random Effects. *Biometrics*, *56*, 761-767.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished Dissertation.

Glas, C. A. W. , & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications.* New-York: Springer Verlag.

Grayson, D.A. (1988). Two-group classification in latent trait theory: scores with monotone likelihood ratio. *Psychometrika, 53*, 383-392.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 63*, 331-347.

Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.

Lehmann, E.L. (1959). *Testing statistical hypothesis.* New-York: Wiley.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated proficiency. *Educational and Psychological Measurement, 43*,675-685.

Louis, T.A. (1982). Finding the observed information when using the EM algorithm. *Journal of the Royal Statistical Society*, B, *44*, 226-233.

Maris, G., & Bechger, T.M. (2003). *A Note on the missing information principle.* R&D report 2003-6. Arnhem: Cito.

Maris, G. (2002). *Concerning the identification of the 3PL model.* R&D report 2002-3. Arnhem: Cito.

Meilijson, I. (1989). A fast improvement of the EM-algorithm on its own terms. *Journal of the Royal Statistical Society*, Series B, *51*, 127-138.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 16*, 159-176.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: the-

ory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability. 1*, 697-715.

Rao, C. R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to the problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44*, 50-57.

Redner, R.A., & Walker,H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev., 26*,195-239.

Revuelta, J. (2000). *A psychometric model for multiple choice items.* Doctoral Dissertation Universidad Autonoma de Madrid.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Ross, S.M. (1996). *Stochastic processes.* 2nd Editon, New-York:Wiley.

Tanner, M. A. (1993). *Tools for statistical inference.* New-York: Springer-Verlag.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*,501-519.

Tierney, L., & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *JASA, 81*, 82-86.

Verstralen, H.H.F.M. (1997). *A logistic latent class model for multiple choice items.* R&D report 97-1. Arnhem:Cito.

Verstralen, H.H.F.M., & Verhelst, N.D. (1998). *A latent class IRT model for options of multiple choice items.* Unpublished Manuscript. Arnhem:Cito.

Verstralen, H.H.F.M., & Verhelst, N.D. (2000). *IRT models for subjective weights of options of multiple choice items.* R&D report 2000-3. Arnhem:Cito.