Marginal Maximum Likelihood Regression Estimation

Developed by Paul Bailey and Harold Doran*

March 27, 2020

This document describes Marginal Maximum Likelihood (MML) estimation for student test data in the `Dire` package. In these models, failing to account for the measurement variance can bias the regression or variance estimates.

The student test data are assumed to have been generated by an Item Response Theory (IRT) model, where students' responses are correct or incorrect (have an increasing score) when student i has higher ability (θ_i) and decreasing when the item is more difficult (d_j), so the probability of a correct response increases as the quantity $\theta_i - d_j$ increases. See the appendix for the likelihood functions for various response models.

This package considers a regression model of the following form (one case in Cohen & Jiang 1999):

$$\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of student abilities, $\boldsymbol{X}$ is a matrix of covariates with unknown parameters $\boldsymbol{\beta}$ and residual variance $\boldsymbol{\epsilon}$. For estimation, we assume that the residual variance is normally distributed without covariance across observations (students) sharing variance of unknown level σ^2 so that

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{2}$$

where $N(\mathbf{0}, \sigma^2\boldsymbol{I})$ is the normal distribution with mean zero, and covariance $\sigma^2\boldsymbol{I}$, and $\boldsymbol{I}$ is the identity matrix. The variance estimation then allows for covariances between students (e.g., in a two-stage sample or clustered within schools).

The next section describes the estimation of $\boldsymbol{\beta}$ and σ^2. The final section describes five methods for variance estimation available in MML estimation, including the traditional (consistent) method, two heteroskedasticity robust methods, and two methods appropriate to a two-stage survey sample, such as the National Assessment of Educational Progress (NAEP).

Parameter Estimation

Student test data[1] consist of a series of items on which a student receives a score. The matrix $\boldsymbol{R}$ has row i regarding a student and column j regarding an item so that R_{ij} is student i's score on item j and takes on integer values from 0 to the maximum score on the item. Many possible models exist for the $\boldsymbol{R}$ matrix data, which are covered, briefly, in the appendix to this document. The rest of this document simply assumes that item parameters have been estimated with a consistent estimator and are treated as being estimated without error.

In an MML model for test data for N individuals, conditional on a set of parameters for a set of K test items, the likelihood of a regression equation is

$$\mathcal{L}(\boldsymbol{\beta}, \sigma | \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(\theta_i - \boldsymbol{X}_i\beta)^2}{2\sigma^2} \prod_{j=1}^{K} \Pr(\boldsymbol{R}_{ij} | \theta_i, \boldsymbol{P}_j) d\theta_i \right]^{\boldsymbol{w}_i} \tag{3}$$

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

[1]Note that these methods equivalently apply to survey construct data that are scored in the same way.

where $\mathcal{L}$ is the likelihood[2] of the regression parameters $\boldsymbol{\beta}$ with full sample weights $\boldsymbol{w}_i$ conditional on item score matrix $\boldsymbol{R}$, student covariate matrix $\boldsymbol{X}$, and item parameter data $\boldsymbol{P}$; σ^2 is the variance of the regression residual; θ_i is the ith student's latent ability measure that is being integrated out; $\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j)$ is the probability of individual i's score on test item j, conditional on the student's ability and item parameters $\boldsymbol{P}_j$—see the appendix for example forms of $\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j)$. Note that if the user is only interested in the population mean, it can be regarded as a special case; $\boldsymbol{X}$ is a vector of all ones, and the value of $\boldsymbol{\beta}$ has only one element that is the mean estimate.

The integral is evaluated using the trapezoid rule[3] at quadrature points t_q and quadrature weights δ so that

$$\mathcal{L}(\boldsymbol{\beta}, \sigma|\boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \prod_{i=1}^{N} \left[\sum_{q=1}^{Q} \delta \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(t_q - \boldsymbol{X}_i\boldsymbol{\beta})^2}{2\sigma^2} \prod_{j=1}^{K} \Pr(\boldsymbol{R}_{ij}|t_q, \boldsymbol{P}_j) \right]^{\boldsymbol{w}_i} \tag{4}$$

where δ is the distance between any two uniformly spaced quadrature points so that $\delta = t_{q+1} - t_q$ for any q that is at least one and less than Q. The range and value of Q parameterize the quadrature, and its accuracy and should be varied to ensure convergence. The advantage of the trapezoidal rule is that the fixed quadrature points allow the values of the probability to be calculated once per student.

The variance formulas use the log-likelihood, which is given by

$$\ell(\boldsymbol{\beta}, \sigma|\boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \sum_{i=1}^{N} \boldsymbol{w}_i \log \left[\delta \sum_{q=1}^{Q} \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(t_q - \boldsymbol{X}_i\boldsymbol{\beta})^2}{2\sigma^2} \prod_{j=1}^{K} \Pr(\boldsymbol{R}_{ij}|t_q, \boldsymbol{P}_j) \right] \tag{5}$$

Note that δ can be removed for optimization, and its presence adds $\log(\delta) \sum \boldsymbol{w}_i$ to the log-likelihood.

Composite Scores

When the outcome of interest is composite scores, the parameters are estimated by separately estimating the coefficients for each subscale ($\boldsymbol{\beta}_s$ for subscale s) and then calculating the composite scores ($\boldsymbol{\beta}_c$) using subscale weights (ω_s).[4]

$$\boldsymbol{\beta}_c = \sum_{s=1}^{S} \omega_s \boldsymbol{\beta}_s \tag{6}$$

where there are S subscales.

For variance estimation, the covariance matrix ($\boldsymbol{\Sigma}$) between subscales is of interest. The covariance terms are estimated one at a time using the submatrix

$$\boldsymbol{\Sigma}_{ij} = \begin{bmatrix} s_i & s_{ij} \\ s_{ij} & s_j \end{bmatrix} \tag{7}$$

so that the two are jointly bivariate normally distributed

$$\begin{pmatrix} \beta_i \\ \beta_j \end{pmatrix} | \boldsymbol{\Sigma}_{ij}, \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P} \sim \text{MVN}\left(\begin{pmatrix} \beta_i \\ \beta_j \end{pmatrix}, \boldsymbol{\Sigma}_{ij} \middle| \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P} \right) \tag{8}$$

where $\text{MVN}(u, S|\cdot)$ is the multivariate normal density function with mean u and covariance S, conditional on $\cdot$, which are additional parameters.

[2] When survey weights are applied, the likelihoods in this document are all pseudo-likelihoods.

[3] Using Big-O notation (Black, 2019), the trapezoid rule's convergence is in $O(\delta^2)$, meaning that the convergence is proportional to δ^2. If the bounds are set wide enough such that every student's likelihood is essentially zero at the edges, the convergence rate is faster than polynomial because the function is periodic and analytic (Johnson, 2010).

[4] We use the term *composite score* to mean those scores that are weighted sums of subscale scores, as in Eq. 6. Overall scores that use a unidimensional model are calculated according to the methods already described by simply pooling items into a single construct.

The likelihood is then

$$\ell\left(s_{ij}\,|\beta_i,\beta_j,s_i,s_j;\boldsymbol{w},\boldsymbol{R},\boldsymbol{X},\boldsymbol{P}\right)=\sum_{n=1}^{N}\boldsymbol{w}_n\log\left\{\delta^2\sum_{q_i=1}^{Q}\sum_{q_j=1}^{Q}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}\exp\left(\boldsymbol{r}_{q_1q_2}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_{q_1q_2}\right)\right. \tag{9}$$

$$\left.\times\left[\prod_{k=1}^{K}\Pr(\boldsymbol{R}_{nk}|t_{q_1},\boldsymbol{P}_k)\right]\left[\prod_{k=1}^{K}\Pr(\boldsymbol{R}_{nk}|t_{q_2},\boldsymbol{P}_k)\right]\right\} \tag{10}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and the residual term is defined as

$$\boldsymbol{r}_{q_1q_2}=\begin{pmatrix} t_{q_1}-\boldsymbol{X}_n\boldsymbol{\beta}_i \\ t_{q_2}-\boldsymbol{X}_n\boldsymbol{\beta}_j \end{pmatrix} \tag{11}$$

Notice that the parameters $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_j$, s_i, and s_j are used from the by-subscale estimation and optimization of the density function is exclusively over the covariance term s_{ij}.

The joint distribution of the vector

$$\boldsymbol{\beta}_{\cdot}=\begin{pmatrix}\boldsymbol{\beta}_1\\ \vdots \\ \boldsymbol{\beta}_S\end{pmatrix} \tag{12}$$

is then

$$\boldsymbol{\beta}_{\cdot},\boldsymbol{\Sigma}|\boldsymbol{w},\boldsymbol{R},\boldsymbol{X},\boldsymbol{P}=\mathrm{MVN}(\,\boldsymbol{\beta}_{\cdot},\boldsymbol{\Sigma}|\,\boldsymbol{w},\boldsymbol{R},\boldsymbol{X},\boldsymbol{P}) \tag{13}$$

which has an intractably high dimensional log-likelihood because it involves S sums inside the log-likelihood.

Variance Estimation

Estimating variance of the parameters $\boldsymbol{\beta}$ can be done in one of several ways.[5]

The inverse Hessian matrix is a consistent estimator when the estimator of $\boldsymbol{\beta}$ is consistent (Green, 2003, p. 520):

$$\mathrm{Var}(\boldsymbol{\beta})=-\boldsymbol{H}(\boldsymbol{\beta})^{-1}=-\left[\frac{\partial^2\ell(\boldsymbol{\beta},\sigma|\boldsymbol{w},\boldsymbol{R},\boldsymbol{X})}{\partial\boldsymbol{\beta}^2}\right]^{-1} \tag{14}$$

This variance is returned when the variance method is set to `consistent` or left as the default.

A class of variance estimators typically called "sandwich" or "robust" variance estimators allow for variation in the residual and are of the form

$$\mathrm{Var}(\boldsymbol{\beta})=H(\boldsymbol{\beta})^{-1}\boldsymbol{V}H(\boldsymbol{\beta})^{-1} \tag{15}$$

where V is an estimate of the variance of the summed score function (Binder, 1983).

For a convenience sample, we provide two robust estimators. First, the so-called `robust` (Huber or Huber-White) variance estimator uses

$$\boldsymbol{V}=\sum_{i=1}^{N}\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta}\right]\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta}\right]' \tag{16}$$

[5]Strictly speaking, σ^2 also is a parameter, but we are rarely interested in the variance of the variance. Nevertheless, the package generates an estimate of σ^2 along with the coefficients themselves. For notational simplicity, all formulas ignore this.

Second, for the `cluster robust` case, the partial derivatives are summed within the cluster so that

$$\boldsymbol{V} = \sum_{c=1}^{n'} \left[\frac{\partial \ell(\beta, \sigma | \boldsymbol{w}_c, \boldsymbol{R}_c, \boldsymbol{X}_c)}{\partial \beta} \right] \left[\frac{\partial \ell(\beta, \sigma | \boldsymbol{w}_c, \boldsymbol{R}_c, \boldsymbol{X}_c)}{\partial \beta} \right]' \tag{17}$$

where there are n' clusters, indexed by c, and the partial derivatives are summed within the group of which there are n_c members:

$$\frac{\partial \ell(\beta, \sigma | \boldsymbol{w}_c, \boldsymbol{R}_c, \boldsymbol{X}_c)}{\partial \beta} = \sum_{i=1}^{n_c} \frac{\partial \ell(\beta, \sigma | \boldsymbol{w}_i, \boldsymbol{R}_i, \boldsymbol{X}_i)}{\partial \beta} \tag{18}$$

We also provide two survey sampling variance estimation techniques. The first one uses replicate weights, either from the jackknife, including Fay's method for the jackknife, or from balanced repeated replication. In this approach, the typical method of estimating sampling variance still works, and the sampling covariance matrix can be calculated as

$$\text{Var}(\boldsymbol{\beta}) = \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0) (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)' \tag{19}$$

where there are J replicate weights and the result of applying direct estimation under the set of weights j is $\boldsymbol{\beta}_j$, whereas $\boldsymbol{\beta}_0$ is the estimate of $\boldsymbol{\beta}$ under the full sample weights. We recomend using this method when `replicate` variance estimation is requested.

The second survey sampling method is called the `Taylor series` method and uses the same formula as Eq. 15, but $\boldsymbol{V}$ is the estimate of the variance of the score vector (Binder, 1983). Our implementation assumes a two-stage design with n_a primary sampling units (PSUs) in stratum a and summed across the A strata according to

$$\boldsymbol{V} = \sum_{a=1}^{A} \boldsymbol{V}_a \tag{20}$$

where $\boldsymbol{V}_a$ is a variance estimate for stratum a and is defined by

$$\boldsymbol{V}_a = \frac{n_a}{n_a - 1} \sum_{p=1}^{n_a} (\boldsymbol{s}_p - \bar{\boldsymbol{s}}_a) (\boldsymbol{s}_p - \bar{\boldsymbol{s}}_a)' \tag{21}$$

where s_p is the sum of the weighted (or pseudo-) score vector that includes all units in PSU p in stratum a and $\bar{\boldsymbol{s}}_a$ is the (unweighted) mean of the $\boldsymbol{s}_p$ terms in stratum a so that

$$s_p = \sum_{i \in \text{PSU } p} \frac{\partial \ell(\beta, \sigma | \boldsymbol{w}_i, \boldsymbol{R}_i, \boldsymbol{X}_i)}{\partial \beta} \qquad \bar{\boldsymbol{s}}_a = \frac{1}{n_a} \sum_{p \in \text{stratum } a} s_p \tag{22}$$

When a stratum has only one PSU, $\boldsymbol{V}_a$ is undefined. The best approach is for the analyst to adjust the strata and PSU identifiers, in a manner consistent with the sampling approach, to avoid singleton strata. Two simpler but less defensible options are available. First, the strata with single PSUs can be dropped from the variance estimation, yielding an underestimate of the variance.

The second option is for the singleton stratum to use the overall mean of s_p in place of $\bar{s}_a$. So,

$$\bar{\boldsymbol{s}} = \frac{1}{n'} \sum s_p \tag{23}$$

where the sum is across all PSUs, and n' is the number of PSUs across all strata. Then, for each singleton stratum, Eq. 21 becomes

$$\boldsymbol{V}_a = 2 (\boldsymbol{s}_p - \bar{\boldsymbol{s}}) (\boldsymbol{s}_p - \bar{\boldsymbol{s}})' \tag{24}$$

where the value 2 is used in place of $\frac{n_a}{n_a - 1}$, which is undefined when $n_a = 1$. This option can underestimate the variance but is thought to more likely overestimate it.

Composite Scores

The likelihood of composite scores (Eq. 13) is additively separable, the covariances (including the variances) can be calculated in two steps using Eq. 13. First, the covariance matrix of $\boldsymbol{\xi}$ is formed, and then the composite covariance terms are estimated as the variance of a linear combination of the elements of $\boldsymbol{\xi}$.

In the first step, any of the methods in the section "Variance Estimation" are applied to Eq. 13, treating $\boldsymbol{\xi}$ in the same fashion Eq. 13 treats $\boldsymbol{\beta}$. This step results in a block diagonal inverse Hessian matrix, with a block for each subscale, and a potentially dense matrix for $\boldsymbol{V}$. Each matrix is square and has $S \cdot (\zeta + 1)$ rows and columns, where ζ is the number of elements in the regression formula (each subscale), to which one is added for the σ terms.

This step results in the following matrix:

$$\text{Var}(\boldsymbol{\xi}) = H(\boldsymbol{\xi})^{-1} \boldsymbol{V} H(\boldsymbol{\xi})^{-1} \tag{25}$$

For the second step, the composite coefficient then has an ith variance term of

$$\text{Var}(\boldsymbol{\xi}_{ci}) = \boldsymbol{e}_i H(\boldsymbol{\xi})^{-1} \boldsymbol{V} H(\boldsymbol{\xi})^{-1} \boldsymbol{e}_i \tag{26}$$

where $\boldsymbol{\xi}_{ci}$ is the composite coefficient for the ith coefficient, and $\boldsymbol{e}_i$ is the vector of weights arranged such that

$$\boldsymbol{\xi}_{ci} = \boldsymbol{e}_i^T \boldsymbol{\xi} \tag{27}$$

The covariance between two terms, i and j, is a simple extension

$$\text{Cov}(\boldsymbol{\beta}_{ci}, \boldsymbol{\beta}_{cj}) = \boldsymbol{e}_i H(\boldsymbol{\beta})^{-1} \boldsymbol{V} H(\boldsymbol{\beta})^{-1} \boldsymbol{e}_j \tag{28}$$

which uses the definition,

$$\boldsymbol{\xi}_{cj} = \boldsymbol{e}_j^T \boldsymbol{\xi} \tag{29}$$

A simple example may help clarify. Imagine a composite score composed of two subscales, 1 and 2, with weights $\omega_1 = 0.4$ and $\omega_2 = 0.6$. Supposed a user is interested in a regression of the form

$$\theta = a + x_1 \cdot b + \epsilon \tag{30}$$

$$\epsilon \sim N(0, \sigma) \tag{31}$$

Then the regression in Eq. 30 would be fit once for subscale 1 and once for subscale 2; the first fit would yield estimated values $\{a_1, b_1, \sigma_1\}$, and the second fit would yield $\{a_2, b_2, \sigma_2\}$. The estimated value, for example, a_c, would be $a_c = 0.4 \cdot \alpha_1 + 0.6 \cdot \alpha_2$. By stacking the estimates together,

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 \\ b_1 \\ \sigma_1 \\ a_2 \\ b_2 \\ \sigma_2 \end{bmatrix} \tag{32}$$

the covariance matrix can then be estimated and will result in a matrix $\boldsymbol{\Omega} \equiv \text{Var}(\boldsymbol{\beta})$ from Eq. 14 that has six rows and six columns. Using the vector

$$\boldsymbol{e}_1 = \begin{bmatrix} 0.4 \\ 0 \\ 0 \\ 0.6 \\ 0 \\ 0 \end{bmatrix} \tag{33}$$

it can easily be confirmed that $a_c = \boldsymbol{e}_1^T \boldsymbol{\xi}$, so $\text{Var}(a_c) = \boldsymbol{e}_1^T \boldsymbol{\Omega} \boldsymbol{e}_1$.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(3), 279–292.

Black, P. E. (2019). Big-O notation. In P. E. Black (Ed.), *Dictionary of algorithms and data structures*. Washington, DC: National Institute of Standards and Technology. Retrieved from https://www.nist.gov/dads/HTML/bigOnotation.html

Cohen, J. D., & Jiang, T. (1999). Comparison of partially measured latent traits across nominal subgroups. *Journal of the American Statistical Association*, 94(*448*), 1035–1044.

Green, W. H. (2003). *Econometric analysis* Upper Saddle River, NJ: Prentice Hall.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. I: *Statistics* (pp. 221–233). Berkeley, CA: University of California Press.

Johnson, S. G. (2010). *Notes on the convergence of trapezoidal-rule quadrature*. Retrieved from https://math.mit.edu/~stevenj/trapezoidal.pdf

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. (2nd ed.). London, UK: Chapman & Hall/CRC.

NAEP. (2008). The generalized partial credit model [NAEP Technical Documentation Website]. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_models_gen.aspx.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(*4*), 817–838.

Appendix. Test Probability Density Functions

For all cases scored as either correct or incorrect, we use the *three parameter logit* (3PL) model:

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = g_j + \frac{1 - g_j}{1 + \exp\left[-D\, a_j\, (\theta_i - d_j)\right]} \tag{34}$$

where g_j is the guessing parameter, a_j is the discrimination factor, d_j is the item difficulty, and D is a constant, usually set to 1.7, to map the θ_i and d_j terms to a probit-like space; this term is applied by tradition.

When a *two parameter logit* (2PL) is used, Eq. 34 is modified to omit g_j (effectively setting it to zero):

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_j)\right]} \tag{35}$$

When a *Rasch model* is used, Eq. 35 is further modified to set all a_j to a single a, and D is set to one.

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-a\, (\theta_i - d_j)\right]} \tag{36}$$

The *Graded Response Model* (GRM) has a probability density that generalizes an ordered logit (McCullagh & Nelder, 1989):

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_{R_{ij},j})\right]} - \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_{1+R_{ij},j})\right]} \tag{37}$$

Here the parameters $\boldsymbol{P}_j$ are the cut points d_{cj}, where $d_{0j} = -\infty$ and $d_{C+1,j} = \infty$. In the first term on the right side of Eq. 37, the subscript R_{ij} on $d_{R_{ij},j}$ indicates it is the cut point associated with the response

level to item j for person i, whereas the last subscript (j) indicates that it is the d term for item j. In the second term, the cut point above that cut point is used.

The *Generalized Partial Credit Model* (GPCM) has a probability density that generalizes a multinomial logit (McCullagh & Nelder, 1989)

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{\exp\left[\sum_{c=0}^{R_{ij}} Da_j(\theta_i - d_{cj})\right]}{\sum_{r=0}^{C} \exp\left[\sum_{c=0}^{r} Da_j(\theta_i - d_{cj})\right]} \tag{38}$$

where c indexes cut points, of which there are C, and j indexes the item.

The GPCM equation has an indeterminancy because all d_j terms could increase and make the values of the probability the same. We can solve the indeterminacy in several ways.

NAEP (2008) uses a mean difficulty (b_j), and the d_j values are then given by

$$d_{0j} = 0 \qquad\qquad d_{cj} = b_j - \delta_{jc}\,;\, 1 \leq c \leq C \tag{39}$$

where the δ_{jc} values are estimated so that $0 = \sum_{c=1}^{C} \delta_{jc}$. In this package, when the `polyParamTab` has an `itemLocation`, it serves as `b`. When there is no `itemLocation`, the package uses the δ values directly

$$d_{0j} = 0 \qquad\qquad d_{cj} = \delta_{jc}\,;\, 1 \leq c \leq C \tag{40}$$

When a *Partial Credit Model* (PCM) is used, and the value of D is set to one, whereas a_j is again shared across all items. So

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{\exp\left[\sum_{c=0}^{R_{ij}} a(\theta_i - d_{cj})\right]}{\sum_{r=0}^{C} \exp\left[\sum_{c=0}^{r} a(\theta_i - d_{cj})\right]} \tag{41}$$