Sensorveiledning MAE4120 - Item Response Theory

Björn Andersson

Exam 25/3 2019

Task 9

From the graph we have that:

$$P(\text{Item 1 correct}|\theta = 0) \approx 0.9$$
$$P(\text{Item 2 correct}|\theta = 0) \approx 0.5$$

Assuming that item responses are independent conditional on θ, we then obtain:

$$P(\text{Item 1 and item 2 correct}|\theta = 0) = P(\text{Item 1 correct}|\theta = 0) \times P(\text{Item 2 correct}|\theta = 0)$$
$$\approx 0.9 \times 0.5 \approx 0.45$$

- Up to 1p for correct calculation. Full points are possible for an approximately correct answer, since the numbers are read off of a graph.
- Up to 1p for correct statement of assumption.

Task 10

We can not say beforehand which item is generally more difficult since we do not know the distribution of the latent variable in the range for which it is defined. We can say that item 1 is more difficult for individuals below $\theta \approx 0.8$ and item 2 more difficult for individuals above $\theta \approx 0.8$.

- Full 2p should include identifying that it depends on the latent distribution and a comment on the difficulty of the items in relation to the latent variable.
- Up to 1p if stating only that the items are differently difficult for different values of the latent variable.
- Identifying the ICCs 2-PL models with the same location parameter gives up to 1p. Hence, a consideration of the distribution is needed in order to obtain full credits.

Task 11

a)

Example answer: Since the items are scored based on the completion of the item, a polytomous model for partial credit is suitable. We could consider a Rasch model or a generalized partial credit model, for example.

- Any polytomous IRT model can give up to 1p depending on the motivation.
- 0.5p for an OK model without proper motivation.

b)

Example answer: We can view the probabilities of category responses as a function of a latent variable and fit the graded response model.

- Any polytomous IRT model can give up to 1p depending on the motivation.
- 0.5p for an OK model without proper motivation.

Task 12

With a large amount of items, the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ has variance $1/I(\theta)$ and thus standard error equal to $1/\sqrt{I(\theta)}$. Hence, the standard error at $\theta = 0.4$ is $1/\sqrt{I(\theta = 0.4)} = 1/\sqrt{4} = 1/2$. To obtain the confidence interval we assume a large number of items and use normal theory to obtain the confidence interval:

$$\hat{\theta}^{obs}_{\text{MLE}} + / - 1.96 \times se(\hat{\theta}^{obs}_{\text{MLE}})$$

We observe $\hat{\theta}^{obs}_{\text{MLE}} = 0.4$ and hence obtain

$$0.4 + / - 1.96 \times \sqrt{1/4} => 0.4 + / - 1.96 \times 0.5 => 0.4 + / - 0.98$$

Thus, the estimated confidence interval for the individual θ is

$$(0.4 - 0.98, 0.4 + 0.98) = (-0.58, 1.38).$$

- Up to 1p for identifying that the variance/standard error is based on the information function. It is not necessary to state the expression entirely correctly to obtain 1p.
- Up to 1p for calculating the confidence interval correctly. Small errors in calculation can still give full marks if the method used is correct.

Task 13

Using the linking parameters β_1 and β_2 which transform the parameters from grade 8 to group 9 we then obtain, for the discrimination parameters

$$\alpha_{9j} = \alpha_{8j}/\beta_1$$

and for the location parameters

$$\delta_{9j} = \beta_1\delta_{8j} + \beta_2.$$

Thus, we also have the inverse relationships

$$\alpha_{8j} = \beta_1\alpha_{9j} \text{ and } \delta_{8j} = \frac{\delta_{9j} - \beta_2}{\beta_1}.$$

Thus, the linking coefficients translate the parameters from one grade to the next, with a linear transformation $\alpha_{G8}/\beta_1 = \alpha_{G9}$. Hence, we obtain the estimated discrimination parameter on the grade 9 metric by:

$$\begin{aligned}\hat{\alpha}_{G9} =& \hat{\alpha}_{G8}/\beta_1 \\ =& 1.2/1.2 \\ =& 1\end{aligned}$$

- Up to 1p for demonstrating knowledge of how to convert parameters from one metric to another via a linear transformation. Hence, it is not required to calculate anything for 1p. It is also acceptable to specify the opposite direction to the transformation.
- Up to 1p for correct calculation of the equated discrimination parameter, or a calculation that is consistent with an acceptable description of the transformation.

Task 14

Examples are:
1) Item selection among scale items.
2) Estimation of scale reliability.
3) Evaluation of further psychometric properties concerning for example dimensionality.
4) Evaluating the items with respect to the contribution to the precision of measurement

- Up to 1p for each depending on how much it makes sense, up to a maximum of 2p.
- If listing more than two things, only full 2p if all things are correct. Otherwise, maximum 1p.

Task 15

Examples are:
1) Tailored precision by estimating the standard error of measurement conditional on the latent variable
2) Realistic model for categorical data which accounts for the fact that we have observed item responses that are non-linear functions of an underlying latent variable

- Up to 1p for each depending on how much it makes sense, up to a maximum of 2p.
- If listing more than two things, only full 2p if all things are correct. Otherwise, maximum 1p.
- Examples of incorrect answers (yielding no points):
  - IRT is a modern test theory
  - IRT always improves measurement accuracy

Task 16

The listed probabilities are not consistent with a unidimensional model since the joint probabilities are not equal to the product of the marginal probabilities, e.g.

$$\begin{aligned} P(Y_1 = 1|\theta = -0.6) \times P(Y_2 = 1|\theta = -0.6) &= 0.5 \times 0.7 \\ &= 0.35 \\ &\neq P(Y_1 = 1, Y_2 = 1|\theta = -0.6) \end{aligned}$$

- Up to 2p for correctly identifying and motivating that the stated probabilities are not consistent with the unidimensional assumption.

Task 17

a)

One example of response is:
1) Start with individual model fit.
2) Fix item parameters of 10 first items to be equal and conduct LR tests for each item individually, with some significance level α.
3) Relax items that show significant results.

- There are many possible types of answers that can yield full 2p. What is needed for full marks is a statement regarding sufficient individual model fit in the two groups and a way of identifying DIF that makes sense. For example, in step 2 other hypothesis tests can be used or an information criterion like AIC or BIC can be used.

- If only describing the procedure to detect DIF, a maximum of 1.5p is obtained.
- Partial credit can be obtained for answers that show understanding of the concepts involved.

b)

There are several possible correct answers here:
1) Use MLE or use EAP or MAP with the same prior, with DIF analysis results incorporated.
2) Use the model without the DIF items, again with MLE or EAP/MAP with the same prior for all individuals.
3) Use the sum score without the DIF items.

- To obtain full marks it must be clear that DIF is accounted for by incorporating it in the model or by removing the items. It is also required to make clear that the effect of using different priors is removed when using a Bayesian method.
- Partial credit can be obtained for answers that show understanding of the concepts involved.

Task 18

a)

Example answer: Specify the 1-PL/2-PL models for all dichotomous items and the PCM/GPCM/GRM for all polytomous items, yielding a total of 6 models. Select the best model based on the BIC.

- Up to 1p for appropriate models for the two sets of items.
- Up to 1p for appropriate selection of these models, which can then be based on hypothesis testing or information criteria depending on the type of model. Hypothesis testing is appropriate for models that are nested, while information criteria can be used more generally.

b)

Examples are:
1) S-chisq hypothesis tests - test if the proportions correct conditional on the sum score are equal between the observed proportions and the model-implied proportions.
2) Graphical illustration of ICC/ICRF and proportion correct conditional on the ability estimates, checks if the observations are approximately consistent with the estimated model.
3) Item RMSEA, gives a measure of approximate item fit taking into account model complexity.

- Up to 1p for each depending on how much it makes sense, up to a maximum of 2p.
- If listing more than two things, only full 2p if all things are correct. Otherwise, maximum 1p.
- If describing a model fit procedure instead of an item fit procedure, partial credit but not full credit can be obtained, depending on the correctness of the reasoning.

Task 19

1) RMSEA ≤ 0.05 indicates close but not excellent approximate fit with binary models, taking model complexity into account.
2) SRMSR ≤ 0.05 indicates adequate but not excellent approximate fit with binary models, without considering model complexity.
3) The M2-test is rejected which indicates non-perfect fit with respect to a limited-information statistic.
4) M2 tests if the marginal one-way and two-way joint probabilities between all item pairs are equal between observations and the model implications.

- Up to 1p each for correctly assessing each of the results. It is fine to say that the model fit is "good" instead of "close" or "adequate", with the same cutoff values. Marks are not given for saying "excellent fit".
- Up to 1p for correctly specifying what M2 tests.

Task 20

a)

The estimates indicate how the intercept parameters for each item type are related to the item domain. β_0 indicates the intercept parameter for the algebra domain, while β_1 indicates how much higher or lower the intercept parameter is for the calculus domain in relation to the algebra domain. Lastly, β_2 indicates how much higher or lower the intercept parameter is for the statistics domain in relation to the algebra domain.

- Up to 2p depending on the correctness of the argument.

b)

We have the predicted value of the item intercept as $\hat{\gamma}_5 = 0.5 - 1.5 = -1$. Furthermore, the location parameter is equal to $-\gamma_5/\alpha_5$, providing the predicted value of the location parameter equal to $\hat{\delta}_5 = -(-1)/2 = 0.5$.

- Up to 1p for predicted value of γ_5 correct.
- Up to 1p for predicted value of δ_5 correct.

- Partial credit can be obtained for partly correct calculations that demonstrate understanding of the model.

c)

A higher SES by one unit means that on average the mathematics proficiency is 0.4 units higher, when all other covariates remain fixed. This result is statistically significantly different from 0 judging from the standard errors. A female individual has 0.01 higher mean mathematics proficiency, when all other covariates remain fixed. This is not statistically significant. In summary, there is a difference in the mean mathematics proficiency with respect to the SES but not with respect to gender. The model implies that female individuals with a SES of 0.5, has a mean mathematics proficiency of 0.21.

- Up to 1.5p for correct interpretation of the relationship.
- Additional 0.5p for correctly interpreting the results for the case given. Also acceptable to disregard the non-significant result for gender, yielding the mean 0.2.