A Hierarchical Mixed Membership Model for Multiple-Answer Multiple-Choice Items with Signal Detection Theory

*****

Total number of words: 1992

Abstract

In this study, we use the signal detection theory to measure decision-making behavior across alternatives. The mixed membership model is applied to capture the grouped data structure, and hierarchical framework is used to measure the item and examinee properties. It can provide rich information relevant to psychological behavior and measurement.

Keywords: Singal Detection Theory, Mixed Membership, Hierarchical Model, Multiple-answer Multiple-choice Item

Introduction

Multiple-choice tests are designed to allow the correctness of the response to reflect the intended ability of interest (Bolt, Wollack, & Suh, 2012). Multiple-answer multiple-choice (MAMC) is a generalization of SAMC item in a way that more than one alternative may be the key. An example of a MAMC item for use in a Graduate Record Examinations (GRE) quantitative reasoning test is the following:

Which of the following integers are multiples of both 2 and 3? (Indicate <u>all</u> such integers.)

A.8 B.9 C.12 D.18 E.21 F.36

The correct answer consists of alternatives C (12), D (18), and F (36).

Objectives

In this study, we illustrate the use of Hierarchical Mixed Membership Model with Signal Detection Theory (HMM-SDT) in MAMC items. The model will be introduced here is an extension of that proposed by DeCarlo (2019) for True–False Exams.

Model Theory

Single alternative Selection Behavior at Individual Level. Figure 1 illustrates the basic ideas of SDT for the situation where there are two cognitive conditions and the two observed response. "*" indicates the true answer of decision. First, examinee compares its *perception* to the decision criterion (c). In Figure 1, if an examinee perceives the *perception* above the criterion, then the decision is "select". In the example of the left plot, the probability for those who know to select is much bigger than who do not know. SDT ensure that examinee who knows is more likely to make the right decision than who do not know. The left side of the decision criterion always represents "not select", and the right side always represents "select". In the example of the plot on the left, signal distribution on the left represents the *plausibility* of decision (select or not select) for the examinee who does not know the alternative. Let δ denote the cognitive condition of the examinee as known ($\delta = 1$) and unknown ($\delta = 0$). SDT model can be viewed as a latent class model with categorical latent variable δ. The center of signal distribution for the examinee who does not know is fixed at zero to avoid identification problem. The distance between two signal distributions is d which reflects how well that alternative discriminating powers between examinee' cognitive conditions. Conceptually, d plays the same role as discriminating power parameters in IRT models. Larger d indicates those who know and unknown are more likely to behave differently. For the kth alternative, *perception* ϕ_k can be viewed as:

$$\phi_k = \delta_k d + \epsilon_k \tag{1}$$

, where ϵ are assume to be identically and independently distributed.

Next, let Z indicates the true answer as "select" ($z = 1$) or "not select " ($z = -1$). In the left plot of Figure 1 ($z = 1$), when the decision boundary moves to the right, the alternative becomes more difficult since the probability for all examinee to make the wrong decision is increasing. When the decision boundary moves to the left, the alternative is

easier. In the right plot of Figure 2 ($z = -1$), the results are opposite. Thus, the location of c decision criterion and true answer status z together represent the difficulty of the alternative. We can use cz denote the difficulty of the alternative. When the decision boundary is equal to the center of signal distribution for the examinee who does not know ($c = 0$), an examinee who does not know can only make the decision based on random "guessing". The difficulty level is zero. When c and z have different signals, the difficulty level is negative (easy). Even for the one who does not know, it has a better-than-even chance of making the right decision. And the difficulty level keeps decreasing if the absolute value of c is increasing. When c and z have the same signal, the results are opposite. *Guessing* and *Slipping* (see the right plot in Figure 1) can be easily defined in SDT models. The probability of who do not know to make the right decision is guessing, and the probability of who know to make the wrong decision is slipping.

Finally, let Y denote the observed response as "select" ($y = 1$) or "not select" ($y = 0$). Examinees' observed response Y for every single alternative in a MAMC item is based on cognitive condition δ, distance (or discriminating power) parameter d, decision criterion c, and true answer z. The probability of observed response in a single alternative is:

$$Pr(y|z, \delta, d, c) = F(c - \delta dz)^{1-y}(1 - F(c - \delta dz))^{y} \tag{2}$$

Different probability distributions can be used for signal distributions, such as normal, logistic, and extreme value distribution.

Multiple-Answer Multiple-Choice Behavior at Individual Level. Let Y_{jk} ($j = 1, 2, ..., J, n = 1, 2, ..., m_j$) denote the observed response of the kth alternative in the jth MAMC item. m_j represents the number of alternatives in the jth item (Usually $m_j = 4$ or 5). Similar to the observed response, parameters in last section can be extend in the same way. Similar to the Generalized Partial Credit Model, we define d_j as the item discriminating power parameter which common across all alternatives, but unique to each item. Since, the concept of the item discriminating power is closely related to the item reliability index in classical test theory, we want to keep the connection with the classical test theory.

MAMC data is grouped data since different alternatives share the same stem and test the same ability under one item. As we view the MAMC item Y_{jk} as the grouped data, the SDT model can be straightforwardly extended to a Mixed Membership or Grade of Membership Model (Davidson, Zisook, & Giller, 1989; Erosheva, Fienberg, & Lafferty, 2004). Mixed Membership model assumes a continuous distribution of latent variables over several categories which reflects the original idea that individuals can be partial members in more than one class (Davidson et al., 1989). If we force the latent variable to have exclusive membership in only one category and no membership in all the other category, this is the latent class model.

Let λ_j denote the partial membership score for the "known" latent class ($\delta_{jk} = 1$), and $1 - \lambda_j$ for the "unknown" latent class ($\delta_{jk} = 0$). Partial membership score is the *propensity* of examinee to know each alternative ($E(\delta_{jk}) = \lambda_n$) independently. Instead of setting each alternative a new partial membership score, we assume that every alternative under one item shares the same partial membership score. This *parallel alternative* design is analogous to the assumption of *parallel* item in classical test theory. We do not derive the model as a process model, although that might be possible. In this way, partial membership score

captures the grouped structure of MAMC data. The shared λ_j across different alternatives capture the grouped structure of MAMC data and avoid treating each alternative separately in a "True-False" sense. The probability of select for the kth alternative in j MAMC item is:

$$\begin{aligned} Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}, \lambda_j) &= Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}) Pr(\delta_{jk}|\lambda_j) \\ &= \lambda_j(1 - F(c_{jk} - d_j z_{jk})) + (1 - \lambda_j)(1 - F(c_{jk})) \end{aligned} \tag{3}$$

This result has a similar format as the IRT 3PL model. (DeCarlo, 2019) did the comparison of Mixed Membership SDT model with IRT model in True/False items.

Multiple-Answer Multiple-Choice Behavior at Group Level. One of the ultimate targets of psychometrics modeling in MAMC item is to improve estimates of examinee ability and item difficulty. Previous sections focus on grouped data Y_{jk} for a signal examinee. Let Y_{ijk} denote the observed response of the kth alternative in jth item for i examinee. Similarly, all the other parameters can all be extended for the kth examinee. We do not extend discriminating power d_{jk} and decision criteria c_{jk} since they are used for capturing the alternative properties. The probability of select for the kth alternative in j MAMC item for the ith examinee is:

$$\begin{aligned} Pr(y_{ijk} = 1|z_{jk}, \delta_{ijk}, d_j, c_{jk}, \lambda_{ij}) &= Pr(y_{ijk} = 1|z_{jk}, \delta_{ijk}, d_j, c_{jk}) Pr(\delta_{ijk}|\lambda_{ij}) \\ &= \lambda_{ij}(1 - F(c_{jk} - d_j z_{jk})) + (1 - \lambda_{ij})(1 - F(c_{jk})) \end{aligned} \tag{4}$$

This formula is the extension of formula (5) in multiple examinee situation.

As we noted above, practical membership score indicates how much the examinee knows each item in a probability scale. Larger λ_{ij} indicates examinee are more likely to know each alternative and select them correctly. We incorporate the idea from Rasch model (Rasch, 1960). The probability that an examinee getting a score (getting all alternative right) depends on the examinee's ability and item difficulty. Similarly, we denote examinee i's ability as θ_i and item difficulty as b_j for the jth item. The model can be extend as:

$$\lambda_{ij} = logit^{-1}(\theta_i - b_j) \tag{5}$$

We name this model as Hierarchical Mixed Membership model with Signal Detection Theory (HMM-SDT). The variables in this model and their interpretations can be summarized in Table 1.

We select the week prior information based on the suggestion from Gelman (2006). See appendix to get more detail about the Stan code. As long as:

$$JK + I + 2J < 2^{IJK} \tag{6}$$

satisfied, we will not have an identification problem. If we apply all independence scenarios we assume before, the log-likelihood function would be:

$$\begin{aligned} log\ p(\cdot) = \sum_i \sum_j \sum_n (&log\ p(Y_{ijk}|\rho_{ijk}) + log\ p(\delta_{ijk}|\lambda_{ij}) + log\ p(\rho_{ijk}|\delta_{ijk}, z_{jk}, c_{jk}, d_j)) \\ &+ \sum_j \sum_n log\ p(c_{jn}|d_j) + \sum_i log\ p(\theta_i) + \sum_j (log\ p(b_j) + log\ (d_j)) \end{aligned} \tag{7}$$

Simulation Study

Data are simulated for 500 examinees, 40 MAMC items, and five alternatives in each item. We need to note that: each item can have a different number of alternatives to use HMM-SDT. The signal distributions are assumed to follow the standard normal distribution. In total there are $500 \times 40 \times 5 = 100000$ decision-making behaviors.

Estimation Convergence & Parameter Recovery. The R-hat statistics for all parameters are equal to one, which indicates the convergence of estimation. As we can see from table one, the parameter would recover to the true value well. And there does not exist a systematical bias in estimation.

Model Implications. In Figure 2, the color of a point is the difficulty of each item (b_j). When examinees have high abilities for the whole exam, their abilities at every item is high as well. Examinees have higher abilities on the easy items and lower ability on the hard items.

The relative location of the decision boundary and two centers of signal distributions provides a hint about the quality of alternatives. If the decision boundary goes beyond the center of signal distribution for the examinee who does not know, the alternative is "too easy". In contrast, if the decision boundary goes beyond the center of signal distribution for the examinee who knows, the alternative is "too hard". Figure 3 shows the visualization of the difficulty analysis with our estimates from HMM-SDT. The dashed line separates the alternatives into different items (five alternatives in one item). We can quickly identify that the 21th item (alternative index from 106 to 110) need to be improved.

In HMM-SDT model, the alternative difficulty and item difficulty are not required to be linear related. In Figure 4, we pick the decision-making behavior of the examinee with the lowest ability ($\theta = -2.603$). In contrast, we pick the examinee with the highest ability in Figure 5 ($\theta = 2.056$). This time, the average slipping decrease dramatically and slipping becomes more unstable across different alternatives. While the average guessing increases a little and becomes more stable.

Model Comparison

As we can see from Figure 3, the distribution of score is more symmetric and dispersed using a Likert-type basis. The score distribution of all-or-none basis is right-skewed. Extreme lower score is more likely to appear on an all-or-none basis.

Instead, we will compare HMM-SDT with 3PL and GPCM based on the estimation performance on their common parameters: overall ability, item discriminating power, and item difficulty. As we can see from Table 3, the 3PL model does not perform well in terms of rank correlation with the true ranking. For average Posterior Standard Deviations (PSD), all three models have a similar level performance. All-or-none scoring basis wastes a lot of useful information. In particular, we cannot distinguish the ability of nine examinees who get zero scores. Meanwhile, extreme cases of one get the item correct are more likely to happen.

References

Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a Multidimensional Nested Logit Model to Multiple-Choice Test Items. *Psychometrika*. doi: 10.1007/s11336-012-9257-5

Davidson, J. R., Zisook, S., & Giller, E. L. (1989). Classification of depression by grade of membership: A confirmation study. *Psychological Medicine*. doi: 10.1017/S0033291700005717

DeCarlo, L. T. (2019). An Item Response Model for True–False Exams Based on Signal Detection Theory. *Applied Psychological Measurement*. doi: 10.1177/0146621619843823

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0307760101

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*. doi: 10.1214/06-BA117A

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*. doi: 10.1016/S0019-9958(61)80061-2

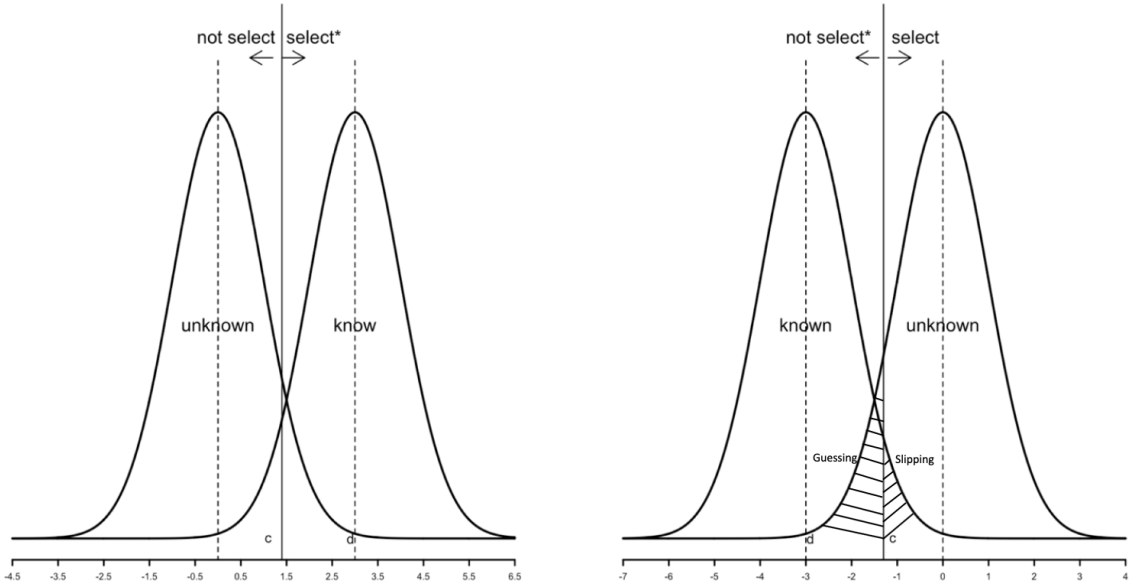

Figure 1. An illustration of signal detection theory with two latent classes and two response categories

Table 1: Parameter and interpretation

Level	Parameter	Interpretation	Scale
alternative Level	c_{jk}	Decision Boundary	$(-\infty, \infty)$
	δ_{ijk}	Membership Assignment (Cognitive Situation: Known & Unknown)	{0,1}
	z^*_{jk}	True Answer (Select & Not Select)	{-1,1}
	Y^*_{ijk}	Response (Select & Not Select)	{0,1}
Item Level	d_j	Item discriminating power	$(0, \infty)$
	b^{**}_j	Item difficulty	$(-\infty, \infty)$
	λ_{ij}	Partial Membership Score (Item Ability)	[0,1]
Examinee Level	θ^{**}_i	Overall ability	$(-\infty, \infty)$

[*] observed variable
[**] hierarchical parameters

Table 2: Parameter Recovery for HMM-SDT model

parameter	measurements	value
θ_i	RMSE	0.513
	Average Biase	-0.009
b_j	RMSE	0.167
	Average Biase	0.002
c_{jk}	RMSE	0.097
	Average Biase	0.000
λ_{ij}	RMSE	0.159
	Average Biase	0.000
d_j	RMSE	0.105
	Average Biase	-0.006

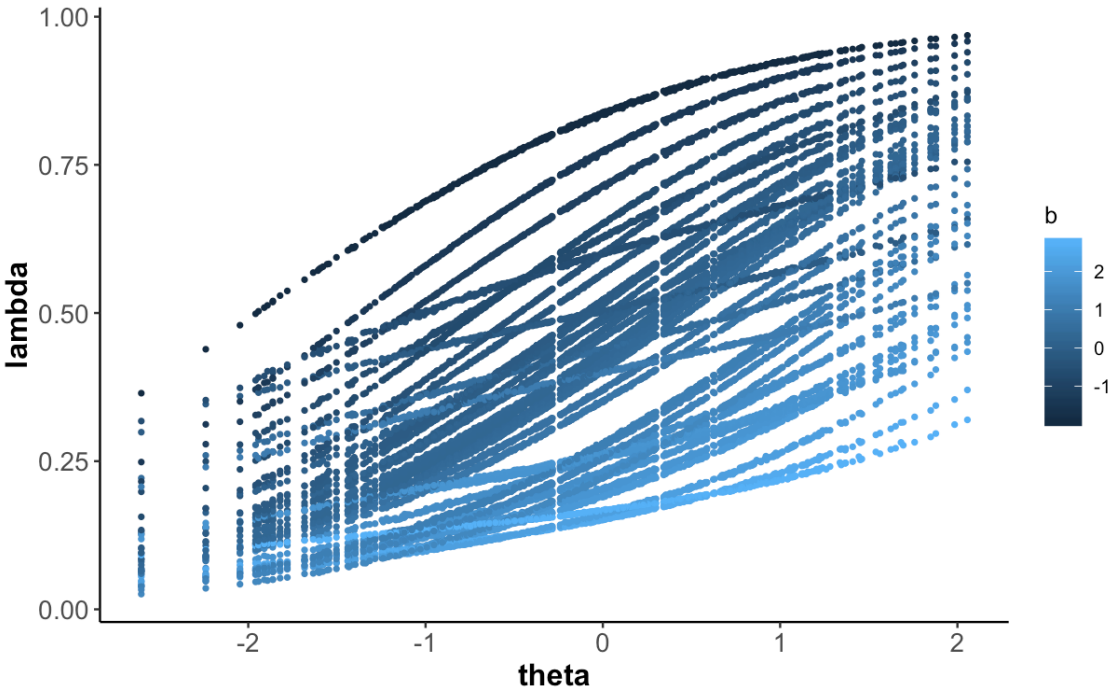

Figure 2. ICC in HMM-SDT

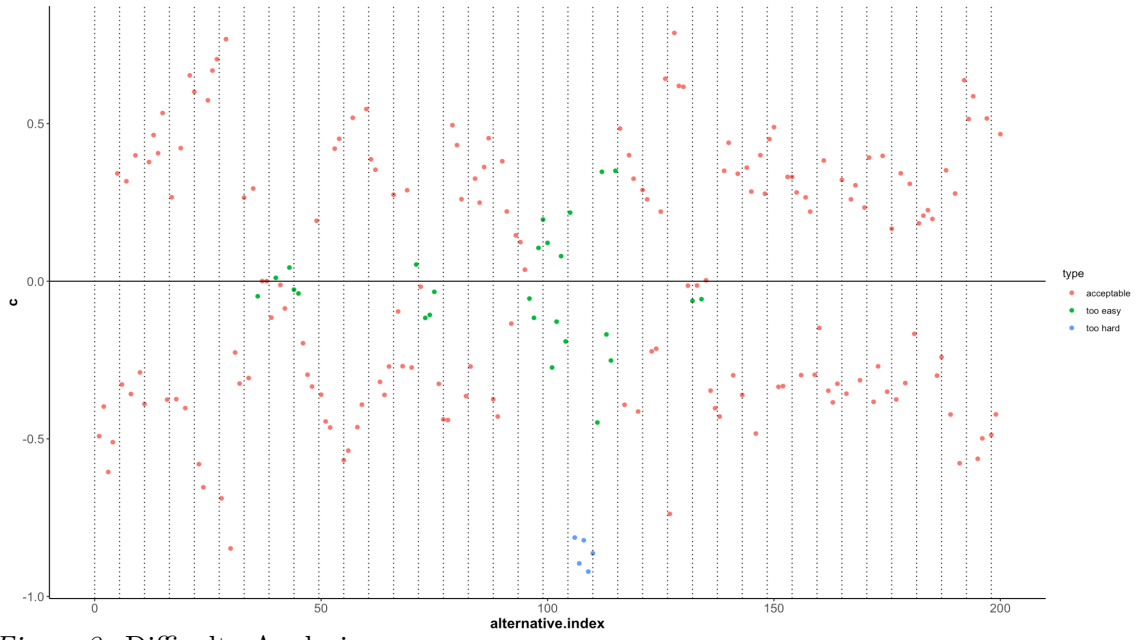

Figure 3. Difficulty Analysis

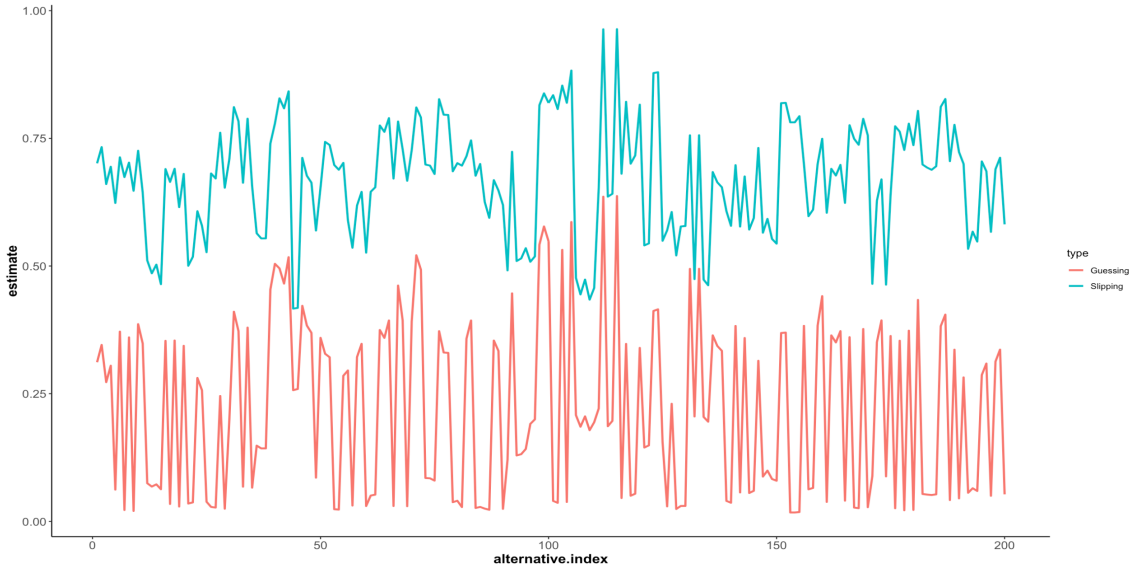

Figure 4. A Random Sample of Guessing & Slipping Estimates for the Examinee with Lowest Ability

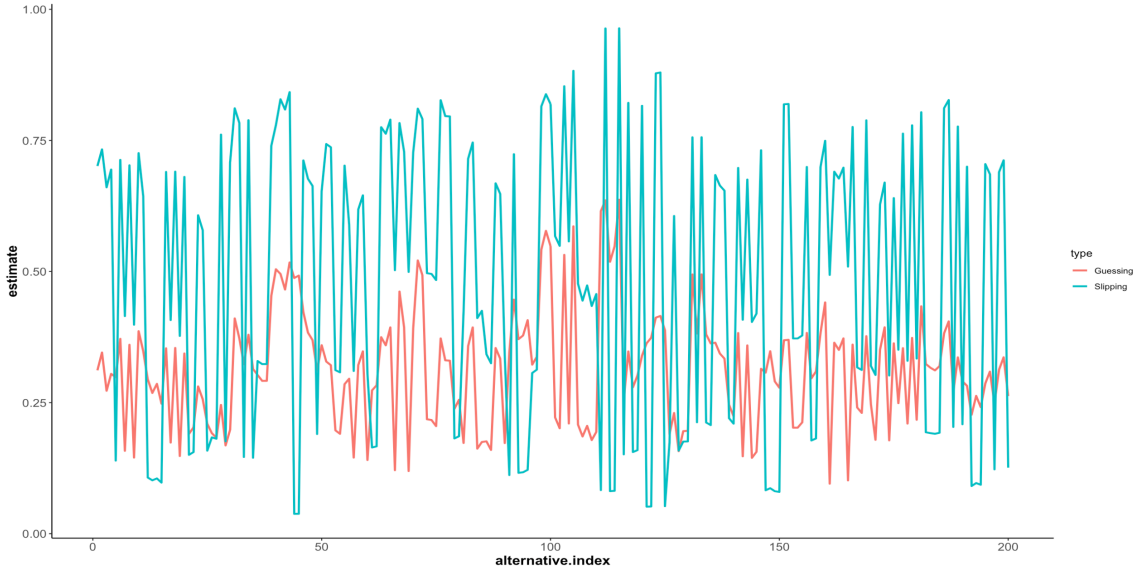

Figure 5. A Random Sample of Guessing & Slipping Estimates for the Examinee with Highest Ability

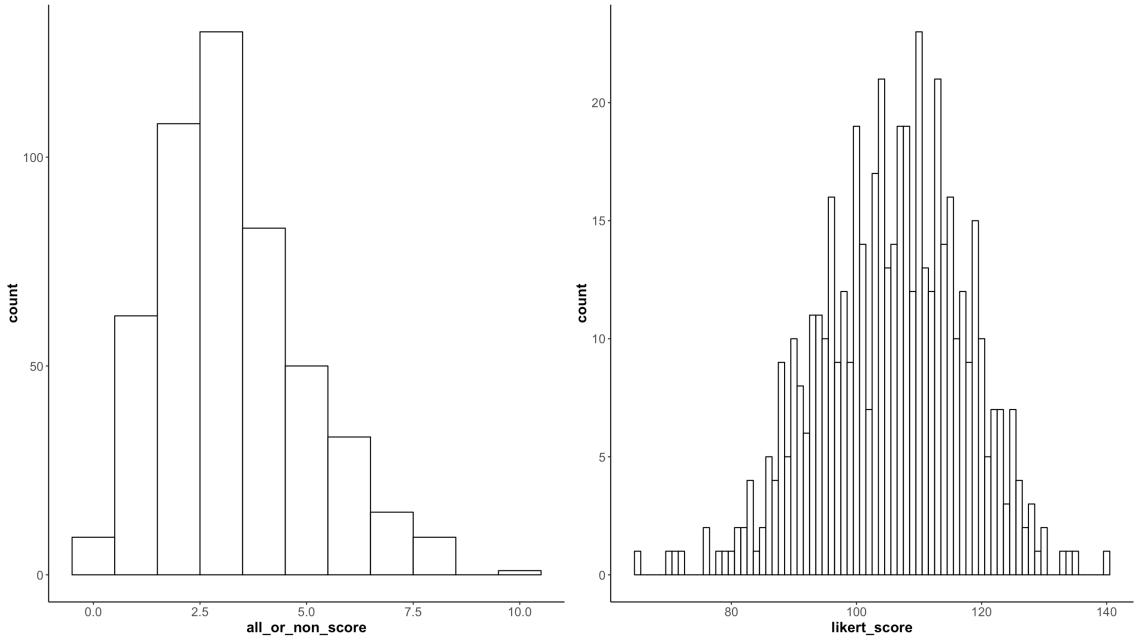

Figure 6. Histogram for all-or-none score and Likert-type score

Table 3: Model Comparison

parameter	model	τ^*	average PSD
Ability	MHH-SDT	71.85%	0.009
	3PL	4.6%	0.014
	GPCM	23.9%	0.003
Difficulty	MHH-SDT	58.46%	0.086
	3PL	9.5%	0.020
	GPCM	21.97%	0.004
Discriminating power	MHH-SDT	45.90%	0.011
	3PL	15.6%	0.014
	GPCM	22.8%	0.003

* Kendall rank correlation coefficient (τ) with the true ranking