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

Yi Chen, HuiSoo Chae, Gary Natriello
Teachers College, Columbia University

Introduction

Multiple-choice tests are designed to allow the correctness of the response to reflect the intended ability of interest (Bolt, Wollack, & Suh, 2012). One item consists of a stem and several alternatives. Single-answer multiple-choice (SAMC) item is the simplest and most commonly discussed type of multiple-choice items. A SAMC item has only one correct alternative (key) and at least one incorrect alternatives (distracters). 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). The checkbox is often used for designing a MAMC question.

In this study, we illustrate the use of Hierarchical Mixed Membership Model with Signal Detection Theory (HMM-SDT) in MAMC items.

Model Theory

Single alternative Selection Behavior at Individual Level. In the MAMC scenarios, examinee's task is to decide whether to select each alternative. The decisions are based on a continuous latent variable ϕ which represents examinee's *perception* of the presented event. Even applying the same stimulus (e.g., answering the same question) to the same observer the response may not be consistent. Thus, *perception* is represented by probability distributions to capture the uncertainty. DeCarlo (2019) provided an innovative interpretation of SDT in testing data. This study follows his interpretation of SDT.

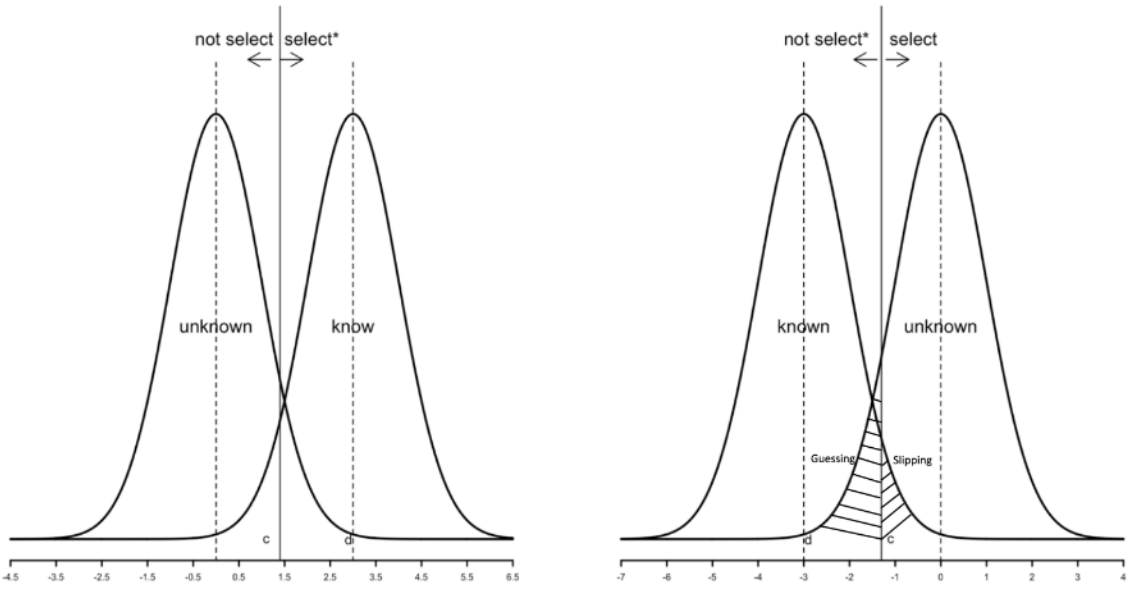

Figure 1. An illustration of signal detection theory with two latent classes and two response categories

Figure 1 illustrates the basic ideas of SDT for the situation where there are two cognitive conditions (known or unknown) and the two observed response (select or not select). The x-axis represents the level of perception ϕ, and the y-axis represents the probabilities. "*" indicates the true answer of decision. First, examinee compares its *perception* to the decision criterion (c, decision boundary). In Figure 1, if an examinee perceives the *perception* above the criterion, then the decision is "select". Otherwise, the decision is "not select". In the example of the left plot, the probability for those who know to select is much bigger than who do not know. And "select" is the right decision for this alternative. The right plot in Figure 1 shows the situation when the true answer is "not select". This time, the signal distribution for who know is on the left while for who unknown is on the right. 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. Signal distribution on the right is for the examinee who knows. 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.

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.

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:

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

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 (e.g., ϕ_{jk}, c_{jk}, z_{jk}, and δ_{jk}) 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 (e.g., ϕ_{jk}) over several categories (e.g., δ_{jk}) 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 probability of select for the kth alternative in j MAMC item is:

$$\begin{aligned} Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}, \lambda_j) &= Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}) Pr(\delta_{jk}|\lambda_j) \\ &= \lambda_j(1 - F(c_{jk} - d_j z_{jk})) + (1 - \lambda_j)(1 - F(c_{jk})) \end{aligned} \tag{2}$$

Multiple-Answer Multiple-Choice Behavior at Group Level. Let Y_{ijk} denote the observed response of the kth alternative in jth item for i examinee. Similarly, all the other parameters (e.g., ϕ_{ijk} , δ_{ijk}, and λ_{ij}) 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:

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

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:

$$\lambda_{ij} = logit^{-1}(\theta_i - b_j) \tag{4}$$

We name this model as Hierarchical Mixed Membership model with Signal Detection Theory (HMM-SDT).

Given distance parameters d_j, criteria parameters c_{jk} and true answer z_{jk}, λ_{ij} is a non-linear transformation of the probability examinee i will get score on the item j. For each alternative, let P_{ijk} the probability of making right decision for the ith examinee on the kth alternative of jth item:

$$P_{ijk} = \begin{cases} \lambda_{ij}F(c_{jk} + d_j) + (1 - \lambda_{ij})F(c_{jk}) & , \quad z_{jk} = -1 \\ \lambda_{ij}(1 - F(c_{jk} - d_j)) + (1 - \lambda_{ij})(1 - F(c_{jk})) & , \quad z_{jk} = 1 \end{cases} \tag{5}$$

Given c_{jk}, d_j, and z_{jk}, P_{ijk} is a linear transformation of λ_{ij}.

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

Simulation Study

Data are simulated for 500 examinees, 40 MAMC items, and five alternatives in each item.

Estimation Convergence & Parameter Recovery. R-hat substantially above one indicates a lack of convergence. 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.

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

Model Implications. In HMM-SDT model, two levels of ability have been examined, J item level ability and one overall ability for each examinee. Item level ability is captured by partial membership score λ_{ij}, and overall ability is captured by the hierarchical parameter θ_i. In Figure 2, the x-axis represents the overall ability and the y-axis represents the item level ability. This plot has a similar idea of ICC of the Rasch model in IRT models. The color of a point is the difficulty of each item (b_j). The darker color represent the easier item. 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. As we noted, the best location for decision boundary is the intersection of two signal distributions ($\frac{1}{2}d_j z_{jn}$). When the decision boundary is between two centers of signal distributions, the alternative difficult is at a reasonable range. In this range, guessing and slipping are smaller than 50%. If the decision boundary goes beyond the center of signal distribution for the examinee who does not know, the alternative is "too easy". Slipping is converge to zero and guessing is converge to one. In contrast, if the decision boundary goes beyond the center of signal distribution for the examinee who knows, the alternative is "too hard". Slipping is converge to one and guessing is converge to zero. 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. Since there are four alternatives to hard and one

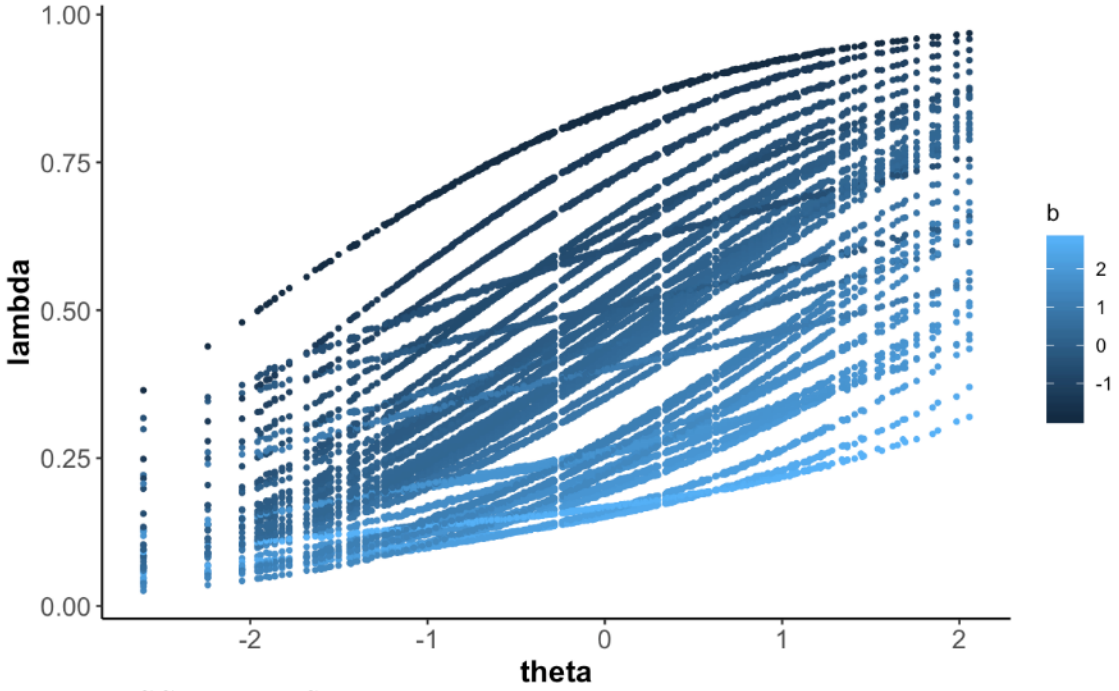

Figure 2. ICC in HMM-SDT

alternative "too easy". Similar the 20th item has five "too easy" items. Again, we can easily find that when c is close to zero, the examinees who do not know have about 50% chance to guess correctly.

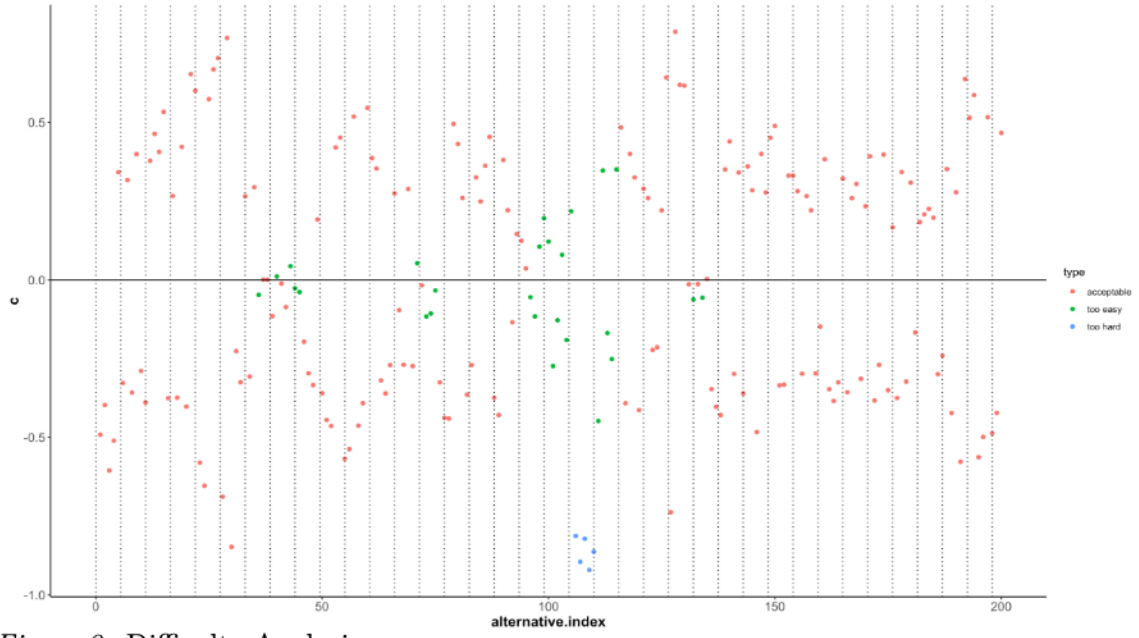

Figure 3. Difficulty Analysis

In HMM-SDT model, the alternative difficulty and item difficulty are not required to

be linear related. Two levels of uncertainties are captured in HMM-SDT. Firstly, alternatives in the same MAMC items are unique. For example, the main skill that one item is testing: "divide fractions". The first alternative may also require the skill of "add fractions", and the second alternative may need "multiply fractions". In HMM-SDT, alternatives are allowed to have different difficulty levels. If the $\lambda_{ij} = 0.8$, we may not see exactly four out of five alternatives are selected correctly. However, we assume that: as the number of alternatives is increasing, the proportion of correct selection converge to λ_{ij}. Secondly, every single decision-making is full of uncertainties beyond the cognitive condition (*known* or *unknown*). For example, time pressure and instantaneous change in the environment of examination may influence how the decisions are made. These uncertainties together can be measured by *guessing* and *slipping* in probability scale. In Figure 4, we pick the decision-making behavior of the examinee with the lowest ability ($\theta = -2.603$). To calculate the guessing and slipping in the 40 items, we use the formula (2) and (3) with λ_{ij} probability of known and $1 - \lambda_{ij}$ probability of unknown. Generally, the slipping is between 50% and 75% and guessing is around 20%. 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. We can analyze the testing result in much more dimensions with HMM-SDT

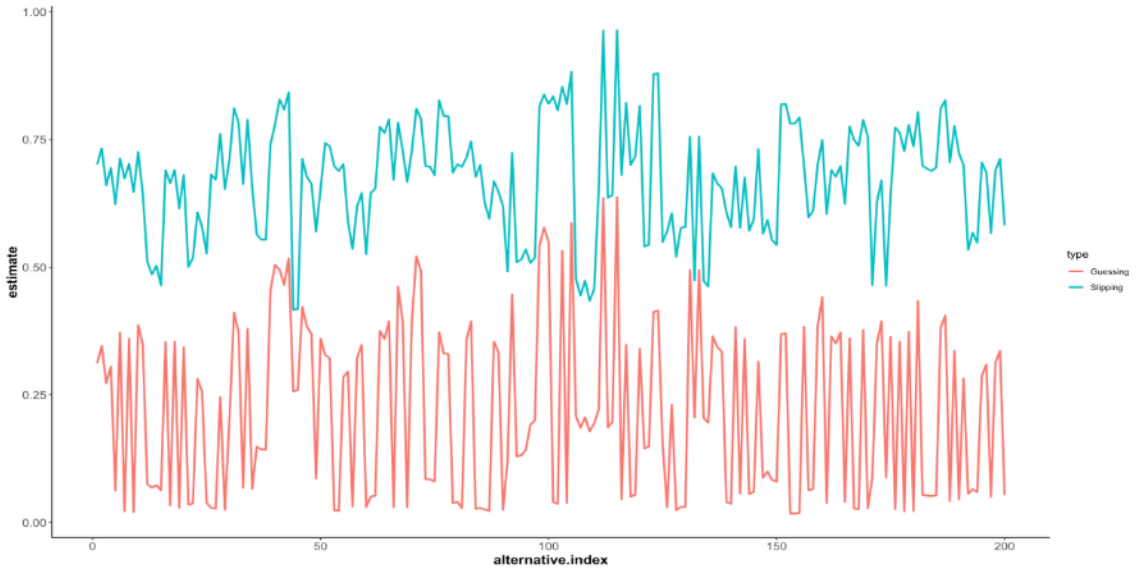

Figure 4. A Random Sample of Guessing & Slipping Estimates for the Examinee with Lowest Ability

We hope that the present article will encourage researchers to use and do more research on HMM-SDT model and MAMC item. The result will be a deeper and more informative analysis.

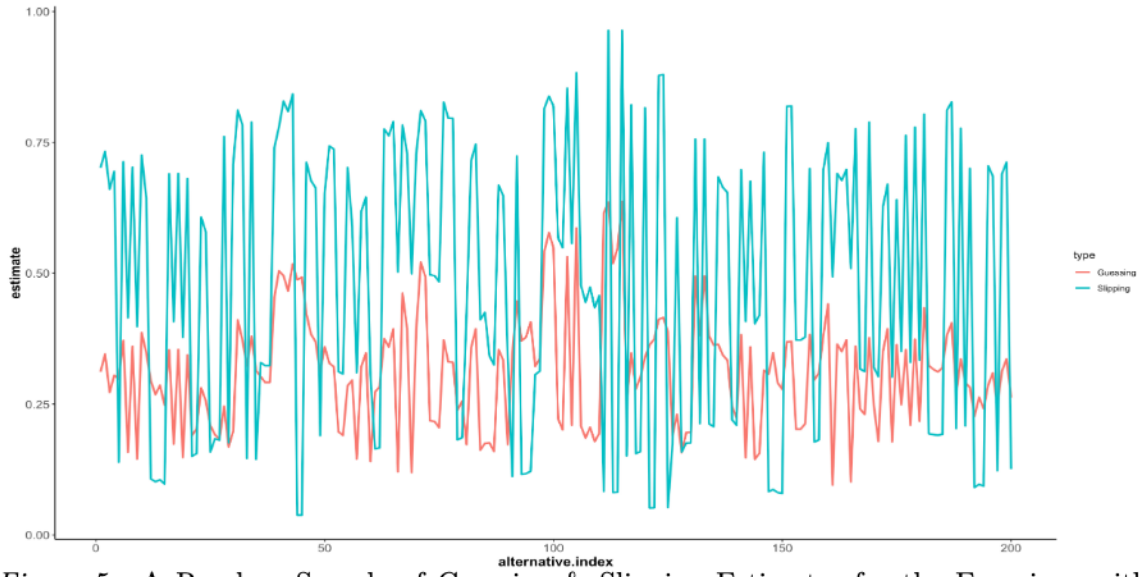

Figure 5. A Random Sample of Guessing & Slipping Estimates for the Examinee with Highest Ability

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

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