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

Abstract

Multiple-answers multiple-choice (MAMC) items are widely used in educational testing and social research. However, inadequate studies and resources are allocated to the assessments of MAMC items. In this study, we introduce a new approach to analyze MAMC items using original response data (which alternatives have been selected) without scoring (correct or wrong). It has immense potential to provide rich information relevant for tractable psychological behavior and interpretable educational measurement. We use the signal detection theory (SDT) to measure the decision-making behavior across alternatives. Then, the mixed membership model is applied to capture the grouped data structure in the MAMC item. A simulation study of the HMM-SDT model is presented with a comparison to the tradition treatments in Item Response Theory (IRT).

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). 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 items 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:

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. Duncan and Milton (1978) summarized three benefits of using MAMC items: 1) the MAMC format permits a more convenient and natural wording of questions and alternatives; 2) test construction is simplified; 3) the appearance of distracters like "none of these" is avoided.

In practical, MAMC items commonly can be scored on an all-or-none basis where the student is given one point for selecting all the correct alternatives, and none of the incorrect alternatives, or is given zero points otherwise. Using the all-or-none basis for multiple-choice items inevitably dichotomizes the original response pattern. Thus, the distinct identity of the incorrect alternatives is lost (Thissen & Steinberg, 1984). For MAMC, this approach is even worse because extreme response pattern (get all items wrong) are more likely to appear. The probability of getting a score converges to zero exponentially as the number of alternatives increases. For example, for a MAMC item that contains five alternatives, if the probability of making the correct choice is 0.5 for every alternative independently, the probability of getting score is 0.5^5 (3.125%). All-or-none basis treat MAMC and SAMC items with no difference. The most widely used models are item response models for a binary response (e.g., Rasch Model and the Birnbaum model). To improve testing efficiency in the context of Multiple-Choice items, extracting additional information from item response data have become highly desirable (Bolt et al., 2012). It is also appealing to design the methodological strategies for analyzing these data.

Alternatively, MAMC items can also be scored by giving one point for each correct alternatives selected and one point for each distracter not selected. However, we need to ensure that every item has the same number of alternatives. Moreover, it treats MAMC items with no difference with Likert-type items and ignores the grouped structure of the item response. Since the items can have more than two possible scores, polytomous item response models are usually used. For example, Partial Credit Model (PCM; Masters, 1982), the Generalized Partial Credit Model (GPCM; Muraki, 1992, 1993), the Rating Scale Model (RSM; Andrich, 1978).

Even though MAMC items are frequently used in educational testing, market research, and elections, inadequate attention has been paid in previous research. Increasing available resources are disproportionately allocated towards assessments of SAMC items. Moreover, models for the SAME items usually cannot be easily extended for the MAMC

item without the loss of information. Additionally, item response theory (IRT) assessments can not exam the psychological behaviors in testing. They are *measurement* models rather than *psychological* models (DeCarlo, 2014).

In this study, we illustrate the use of the Hierarchical Mixed Membership Model with Signal Detection Theory (HMM-SDT) in MAMC items. The model follows directly from a conceptualization about examinees' decision-making behavior based on signal detection theory (SDT; DeCarlo, 1998; Ingleby, 2003; Stanislaw & Todorov, 1999; Hautus, 2015). The model will be introduced here is an extension of what is proposed by ? (?) for True–False Exams. A generalization of this model for analyzing other types of selected-response exams (e,g., SAMC) is possible for the future study. We will present the psychological conceptualization underlying the model and its statistical characteristics. We also discuss the implications of the HMM-SDT model, test the reliability of the model design, and compare this model with item response models in the simulation study.

Model Theory

In this chapter, we will explain the details of the Hierarchical Mixed Membership model with Signal Detection Theory (HMM-SDT). Firstly, Signal Detection Theory (SDT) will be discussed in an alternative selection scenario for a single examinee. Then, the SDT model will be extended with a Mixed Membership framework to measure the selection behavior for the whole item for a single examinee. Finally, the model will be extended to measure multiple examinees' behaviors.

Single alternative Selection Behavior at Individual Level. Signal Detection Theory (SDT) is a mathematical framework for the case that reasoning and decision making takes place in the presence of uncertainty (Swets, 1988). It is widely applied in education, psychology, and medical research. In the MAMC scenarios, the examinee's decision is whether to select each alternative or not. Since the number of keys is usually unknown, the selections are independent. The decisions are based on a continuous latent variable of ϕ, which represents the examinee's *perception* of the presented event. *perception* is represented by probability distributions to capture the uncertainty of selection behaviors. ? (?) provided an innovative interpretation of SDT in testing data. This study follows his interpretation of SDT.

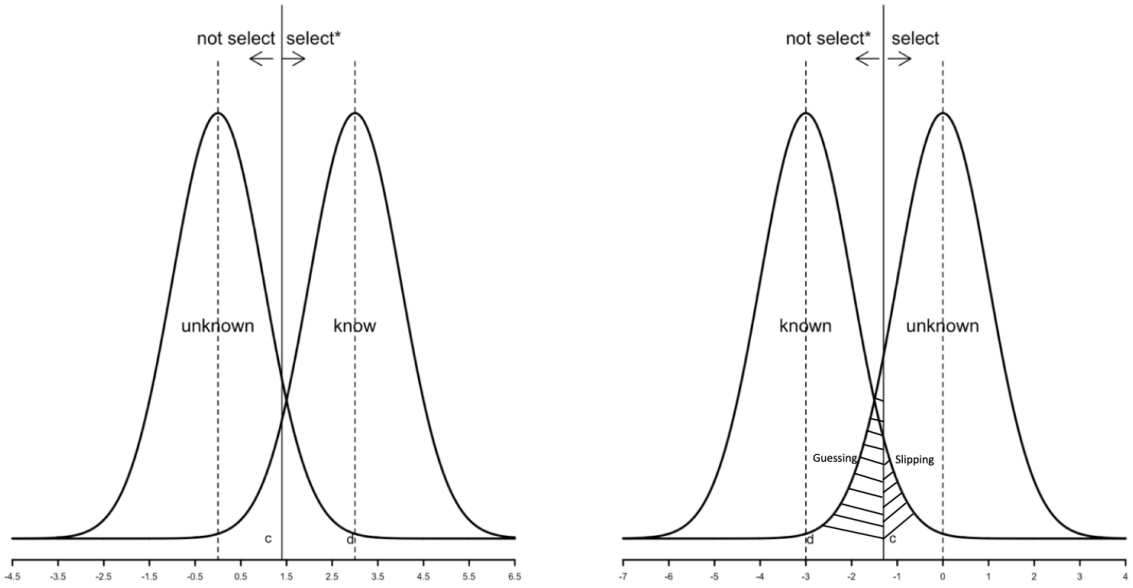

Figure 1. An illustration of signal detection theory with two latent classes and two response categories

Figure 1 illustrates the basic ideas of SDT for the situation where there are two cognitive conditions (known or unknown) and the two observed responses (select or not select). The x-axis represents the level of perception ϕ, and the y-axis represents the probabilities. "*" indicates the correct answer of the decision. First, examinee compares its *perception* to the decision criterion/boundary (c). 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 left plot, examine who know is more likely to select (the correct choice).

The right plot in Figure 1 shows the situation when the correct answer is "not select". The left side of the decision criterion always represents "not select", and the right side always represents "select".

Let δ denote the cognitive condition of the examinee as known ($\delta = 1$) and unknown ($\delta = 0$). Thus, SDT is a latent class model with a categorical latent variable δ. The center of signal distribution for the examinee who does not know is fixed at zero to avoid the identification problem. The distance between two signal distributions is d ($d > 0$), which reflects how well that alternative can discriminate examinees' cognitive conditions. Conceptually, d plays a similar role as discriminating power parameters in IRT models. 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 (e.g., $\epsilon_k \sim Normal(0, \sigma)$).

Let Z indicates the observed correct answer as "select" ($z = 1$) or "not select " ($z = -1$). In the left plot of Figure 1 ($z = 1$), the alternative becomes more difficult since the probability for all examinees to make the wrong decision is increasing as the decision boundary (c) increases. In the right plot of Figure 1 ($z = -1$), the results are the opposite. Thus, the location of c decision criterion and correct answer status z together represent the difficulty of the alternative. We can use cz denote the difficulty of the alternative.

For an examinee who does not know, he/she can only randomly guess if the decision boundary is at the center of signal distribution ($c = 0$). 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. Moreover, the difficulty level keeps decreasing if the absolute value of c is increasing. When c and z have the same signal, the results are the opposite.

Guessing and *Slipping* (see the right plot in Figure 1) can be easily defined in SDT models. For an examinee who does not know, the probability of making the right choice is called guessing. The probability of who knows to make the wrong decision is slipping. Besides, guessing and item difficulty is more consistent In SDT compared with IRT (3PL) approach. Because more difficult alternatives are harder to guess (lower probabilities of being right for the one who does not know). We can denote *gussing* as G, and *sliping* as S.

$$G = \begin{cases} 1 - F(c), & if \quad z = 1 \\ F(c), & if \quad z = -1 \end{cases} \tag{2}$$

$$S = \begin{cases} F(c - d), & if \quad z = 1 \\ 1 - F(c + d), & if \quad z = -1 \end{cases} \tag{3}$$

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 correct answer z. The probability of observed response in a single alternative is:

$$Pr(y|z, \delta, d, c) = F(c - \delta dz)^{1-y}(1 - F(c - \delta dz))^{y} \tag{4}$$

In summary, SDT is a psychological model for decision-making behaviors. The appealing interpretations empower this model to measure the alternative properties (e.g., difficulty cz, discriminating power d) and examinee properties (e.g., cognitive condition δ). The latent class and response patterns are generally ordinal. Additionally, the SDT model can be generalized for the situations where there are more than two latent classes and more than two response categories. Signal distributions can also be captured by logistic and extreme value distribution. To simplify, we assume that signal distributions share the same variance, and the distances between adjacent distributions (d) are fixed. DeCarlo (2003) discussed the challenges of using different variances.

Multiple-Answer Multiple-Choice Behavior at Individual Level. In this section, we extend the model with the Mixed Membership framework for the whole MAMC item. 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). Like 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 is shared across all alternatives. Because item discriminating power is related to the concept of item reliability index in 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). The 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 categories, 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 ($E(\delta_{jk}) = \lambda_j$) independently. 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, a partial membership score captures the grouped structure of MAMC data.

In summary, the alternatives in one MAMC item are exchangeable. 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:

$$\begin{aligned} Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}, \lambda_j) &= Pr(y_{jk} = 1|z_{jk}, \delta_{jk}, d_j, c_{jk}) Pr(\delta_{jk}|\lambda_j) \\ &= \lambda_j(1 - F(c_{jk} - d_j z_{jk})) + (1 - \lambda_j)(1 - F(c_{jk})) \end{aligned} \tag{5}$$

? (?) did the comparison of Mixed Membership SDT model with IRT model in True/False items.

As we mentioned above, the Mixed Membership model can be generalized for the situations where there are more than two latent classes. For example, Latent Dirichlet Al-

location (Blei, Ng, & Jordan, 2000) is an application of Mixed Membership where the latent classes are *topics*, which are nominal. However, criticism arises when an ordinal latent class is extended to be more than two categories in the Mixed Membership model. For example, when the latent classes are "known," "normal," and "unknown," it is abnormal to interpret that one examinee is 20 % known, 30 % normal, and 50 % unknown. Because ordinal latent variables are not exchangeable. In this study, dichotomous memberships (know and unknown) can be viewed as a nominal or ordinal latent variable without apparent differences. "Known" and "unknown" are two extreme situations that rarely happen in reality. Thus, λ_j represents the distance towards these two extreme situations on a probability scale. Moreover, λ_j also indicates examinee is the ability on the jth item.

Multiple-Answer Multiple-Choice Behavior at Group Level. Previous sections focus on grouped data Y_{jk} for a signal examinee. In this section, we extend the model to capture multiple examinees' behaviors with IRT. Let Y_{ijk} denote the observed response of the kth alternative in jth item for i examinee. Similarly, ϕ_{ijk} , δ_{ijk}, and λ_{ij} can all be extended for the kth examinee. We do not extend discriminating power d_j 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:

$$\begin{aligned} Pr(y_{ijk} = 1|z_{jk}, \delta_{ijk}, d_j, c_{jk}, \lambda_{ij}) &= Pr(y_{ijk} = 1|z_{jk}, \delta_{ijk}, d_j, c_{jk}) Pr(\delta_{ijk}|\lambda_{ij}) \\ &= \lambda_{ij}(1 - F(c_{jk} - d_j z_{jk})) + (1 - \lambda_{ij})(1 - F(c_{jk})) \end{aligned} \tag{6}$$

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

However, there are some critical features of interest have not been discussed in this model yet. For example, λ_{ij} can be interpreted as examinee i's ability on item j. However, we can not get the overall ability of an examinee. Similarly, $c_{jk}z_{jk}$ indicates the difficulty of each alternative. However, the overall difficulty of a MAMC item has not been measured.

A straightforward approach to estimate examinees' overall ability is to marginalize the possible membership score of λ_{ij}. We can take the average score of λ_{ij} across different items to measure its ability ($\theta_i = \frac{1}{J}\sum_{j=1}^{J} \lambda_{ij}$). Similarly, item difficulty can be measured by taking the average score of alternatives difficulty ($b_j = \frac{1}{K}\sum_{k=1}^{K} z_{jk}c_{jk}$). However, this approach assumes that the examinees' ability on each item contribute equally towards their overall ability. The difficulty of each alternative contributes equally to the overall difficulty of the item.

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:

$$P_{ijk} = \begin{cases} \lambda_{ij}F(c_{jk} + d_j) + (1 - \lambda_{ij})F(c_{jk}) & , \quad z_{jk} = -1 \\ \lambda_{ij}(1 - F(c_{jk} - d_j)) + (1 - \lambda_{ij})(1 - F(c_{jk})) & , \quad z_{jk} = 1 \end{cases} \tag{7}$$

Given c_{jk}, d_j, and z_{jk}, P_{ijk} is a linear transformation of λ_{ij}. The probabilities for ith examinee to get score on the jth item (select all alternatives correctly), which is used in IRT models, can be expressed as: $\prod_{n=1}^{m_j} P_{ijk}$. It is a non-linear transformation from λ_{ij}. Compared with $\prod_{n=1}^{m_j} P_{ijk}$, λij is less likely to be close to zero, and easier to be estimated in MCMC. Though HMM-SDT needs extra information, it does not need the extra data

process to determine "right" or "wrong" of each decision as in IRT. The initially observed responses and accurate answers for each alternative are the input data required for HMM-SDT.

In summary, decision-making behavior on every single alternative is independently measured by Signal Detection Theory. Decision boundary c_{jk} is used to capture the alternative difficulty Distance parameter d_j is used to measure the item discriminating power. δ_{ijk} is used to capture the examinees' cognitive condition on every alternative. The alternatives within the same MAMC item are measured by the Mixed Membership model design. They share the same partial membership score λ_{ij}, which measure the ability at item level independently. The variables in this model and their interpretations can be summarized in Table 1.

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	$(0, \infty)$
	λ_{ij}	Partial Membership Score (Item Ability)	[0,1]

* observed variable

From Theory to Model

In this article, we would take a Bayesian approach to design the model. There are three levels of independence based on the framework of Mixed Membership model: 1) the conditional independence among the membership assignment parameters δ_{ijk} given the partial membership scores λ_{ij} ; 2) the conditional independence among the observed response Y_{ijk} given the membership assignment parameters δ_{ijk}. Besides, there are two levels of independence based on SDT model: 1) the independence among distance parameters d_j; 2) the independence among the criterion parameters c_{jk}.

We select the week priors based on the suggestions from Gelman (2006). See the appendix to get more detail about the Stan code. Latent class δ_{ijk} are marginalized in the model to apply the no-U-turn sampler (NUTS, Hoffman & Gelman, 2014) in stan. The generative process and the setting of hyperparameter are:

1. Draw components, for each examinee i:
   (a) For each pair of item j and alternative k:
   draw the decision criterion c_{jk}: $c_{jk} \sim Normal(\frac{1}{2}d_j z_{jk}, 1)$.
   (b) Calculate the conditional probabilities of making right decision ρ_{ijk}:
   $\rho_{ijk} = Pr(Y_{ijk} = 1|\lambda_{ij}, z_{jk}, c_{jk}, d_j)$
2. For each item j:
   (a) Draw a difficulty parameter d_j: $d_j \sim Lognormal(0, 0.5)$
   (b) Draw proportions λ_{ij}: $\lambda_{ij} = beta(1, 1)$
   (c) For each observed response Y_{ijk} scored by rater j:
   Draw the data point Y_{ijk}: $Y_{ijk} \sim Bernoulli(\rho_{ijk})$.

Theoretically, the best decision boundary c location is the intersection point of two signal distribution. This decision boundary ensures that people who know always have a higher probability of making the correct choice. Given the variance of two signal distributions are fixed, the intersection point is alway locate at $\frac{1}{2}d_j z_{jk}$.

The total number of the possible pattern is 2^{IJK}. The number of parameter under HMM-SDT model would be: $J \times K$ decision criterion c_{jk} parameters, J distance parameters d_j, and $I \times J$ mixture assignment parameters λ_{ijk}. As long as:

$$IJ + JK + J < 2^{IJK} \tag{8}$$

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

$$\begin{aligned} log\ p(\cdot) = \sum_i \sum_j \sum_n (log\ p(Y_{ijk}|\rho_{ijk}) + log\ p(\rho_{ijk}|\delta_{ijk}, z_{jk}, c_{jk}, d_j) + log\ p(\delta_{ijk}|\lambda_{ij})) \\ + \sum_j \sum_n log\ p(c_{jn}|d_j) + \sum_i \sum_j log\ p(\lambda_{ij}) \end{aligned} \tag{9}$$

Simulation Study

In this section, we will use R (R Development Core Team, 2008) to simulate the fake data. Stan (Carpenter et al., 2017) is used for model estimation. Since MCMC generally requires complex computation, NUTS sampler in Stan takes a long time to estimate. Variational inference can provide a rough approximation of the posterior distributions and takes much shorter estimation time with an acceptable loss of estimation accuracy. In terms of finial inference, the NUTS sampler is generally recommended. In this section, we use mean-field variational inference to measure the tendency of parameter recovery under different usage situations.

We take the Expected A-Posterior (EAP) as the estimated value since this usually gives a more robust estimation. We use Root Mean Square Error (RMSE) and average posterior standard deviation (PSD) as the measurements for parameter recovery. Three examinee sample size (500 and 1000) and two exam item sample size (20 and 40) are tested. Every item is assumed to contain four alternatives.

Table 2: Summary of parameter recovery

		Measurements					
I	J	RMSE($\boldsymbol{c}$)	PSD($\boldsymbol{c}$)	RMSE($\boldsymbol{d}$)	PSD($\boldsymbol{d}$)	RMSE($\boldsymbol{\lambda}$)	PSD($\boldsymbol{\lambda}$)
500	20	0.308	0.062	0.891	0.050	0.150	0.235
	40	0.290	0.058	0.792	0.045	0.159	0.235
1000	20	0.291	0.040	0.895	0.032	0.159	0.235
	40	0.267	0.040	0.809	0.031	0.164	0.235

As we can see from Table 2, the parameter generally recovers to the real value well. Generally, increasing the sample size of items and examinees will improve the estimation accuracy of decision boundary and item discriminating power.

Model Implications

Data are simulated for 500 examinees, 20 MAMC items, and four 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 $500 \times 20 \times 4 = 400000$ decision-making behaviors. Four parallel chains were simulated out to 2000 iterations with the estimates calculated from the last 1000 iterations. For the fake data we get, 40 alternatives should be selected, and 40 should not be selected based on the correct answers (z_{jk}). According to the observed response from examinees, there are 51.58 % alternatives have been selected.

To measure the estimation convergence, we pick the standard criteria: R-hat statistic by Gelman and Rubin (1992). The R-hat statistic measures the degree of multiple parallel Markov Chain, which is run with starting values that are over-dispersed relative to the posterior distribution. If the chains have converged, R-hat is close to one, and estimates are unbiased. 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.

The relative location of the decision boundary and two centers of signal distributions provide evidence about the quality of alternatives. 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 the signal distribution of the examinee who does not know, the alternative is "too easy." Slipping is converge to zero, and guessing is more significant than random guess (50%). In contrast, if the decision boundary goes beyond the center of signal distribution of the examinees who know, the alternative is "too hard." Figure 2 shows the visualization of the difficulty analysis with our estimates from HMM-SDT. The dashed line separates the alternatives into different items (four alternatives in one item). When c is close to zero, the difficulty is less likely to be "too hard" since the examinees have a random chance to get it correct.

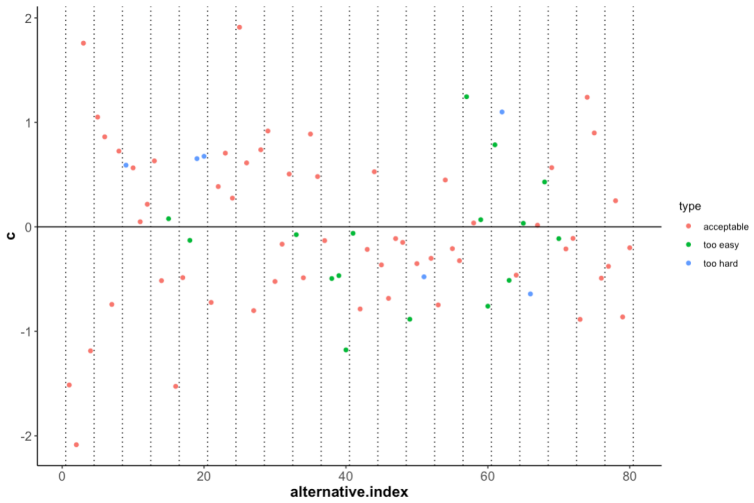

Figure 2. Difficulty Analysis

When there is at least one alternative "too hard" in the item, all-or-none basis scoring tends to underestimate examinee' ability. When there is at least one alternative "too easy" in the item, Likert-type basis scoring will have less discriminating power. For example, the 10th item (alternative index from 36 to 40) needs to be improved (three alternatives are "too easy").

SDT guarantee that $1 - s > g$, which means people who know always have a higher probability of making the right decision than those who do not know. To abstain from the personalized "Guessing" and "1 - Slipping", partial membership score λ_{ij} is required.

$$\begin{aligned} G_{ijk} &= (1 - \lambda_{ij})(1 - F(c_{jk})) \\ S_{ijk} &= \lambda_{ij}(1 - F(c_{jk - d_j z_{jk}})) \end{aligned} \tag{10}$$

Figure 3 shows the "Guessing" and "1 - Slipping" for the examinee with the high overall ability (above) and low ability (below). When "1 - Slipping" is smaller than 0.5, the corresponding alternative is "too difficult," even if the examinee knows this alternative. The probability of whether the examinee knows or does not know the alternative is captured by λ_{ij}. When "Guessing" is smaller than 0.5, the corresponding alternative is harder than random guess if the examinee does not know this alternative. For the examinee with high overall ability, the estimated "Guessing" (most below 0.5) and "1-Slipping" (most above 0.5) are evenly distributed. When λ_{ij} is high, "Guessing" and "1-Slipping" are more clearly separated. For the examinee with low overall ability, most "Guessing" and "1-Slipping" are below 0.5 and mixed.

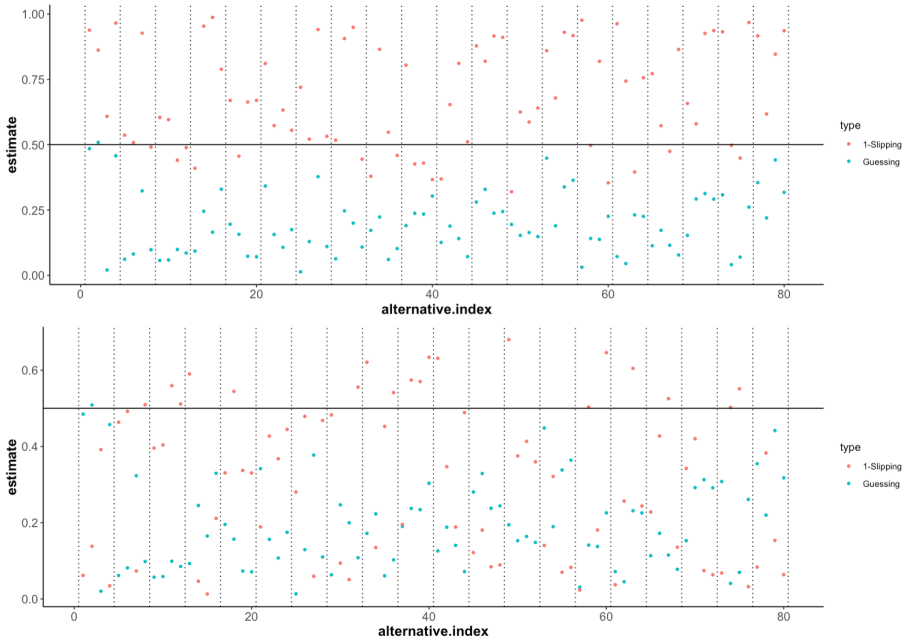

Figure 3. Guessing and Slipping Analysis

Model Comparison

In this section, we will compare the HMM-SDT with the IRT models with the simulated data in the last section. For the most common approach of handling the MAMC item, we take the 3PL model with an all-or-none basis. In the 3PL model, the probability of ith examinee to get the score on the jth item is:

$$Pr(x_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]} \tag{11}$$

, where a are the discriminating power parameters, b are the item difficulty parameters, c are the pseudo-guessing parameters, and θ are the examinee ability parameters.

Alternatively, we also use a Likert-type basis and take the Generalized Partial Credit Model (GPCM).

$$Pr(x_{ij} = k|\theta_i, a_j, \mathbf{b_j}) = \frac{exp[\sum_{v=1}^{k} a_j(\theta_i - b_j + d_v)]}{\sum_{c=1}^{N} exp[\sum_{v=1}^{c} a_j(\theta_i - b_j + d_v)]} \tag{12}$$

, where b are item location parameters and d are the threshold parameter. We assume that $d_1 = 0$ to avoid identification issue. Item location parameters and threshold parameters together capture the probabilities of how many numbers of choices the examinee would make right across different alternatives in a MAMC item. However, they cannot tell the difficulty of each alternative individually.

If we take the all-or-none basis, 32 examinees get the 0 out of 20 (lowest), and six examinee gets 7 out of 20 (highest). The average score is 2.46, the mode score is 2, and the median score is 2. The standard deviation of the score is 1.45. If we take the Likert-type basis, the possible score for every item can be 0 to 5, and the overall highest score is $40 \times 5 = 200$. One examinee gets the 25 (lowest), and two examinee gets 58 (highest). The average score is 40.14, the mode score is 36, and the median score is 40. The standard deviation of the score is 6.32. As we can see from Figure 3, the distribution of the score is more symmetric and dispersed using a Likert-type basis. The score distribution of an all-or-none basis is right-skewed. An extreme lower score is more likely to appear on an all-or-none basis. Thus, a Likert-type basis is more reliable. To some extent, HMM-SDT utilizes a similar basis as Likert-type scoring, which incorporates the decision-making observations on each alternative. However, HMM-SDT does not require to score the alternative or items at all. Since the observed response Y in SDT is the original decision (*select* or *not select*) examinee made on each alternative. Thus, more information can be kept from the original data without going through the scoring process.

All of these models can be used to measure: item difficulty, item discriminating power, and examinee overall ability. However, GPCM cannot measure alternative difficulty, guessing and slipping, and examinee ability on an item. 3PL measure the guessing and slipping on item level, and can not measure alternative difficulty and examinee ability on an item. An increasing number of parameters benefits much more measurement power in HMM-SDT.

As we can see from Table 3, the 3PL model does not perform well in terms of correlation with the actual parameter value. For average Posterior Standard Deviations (PSD), all three models have a similar level performance. All-or-none scoring basis wastes much

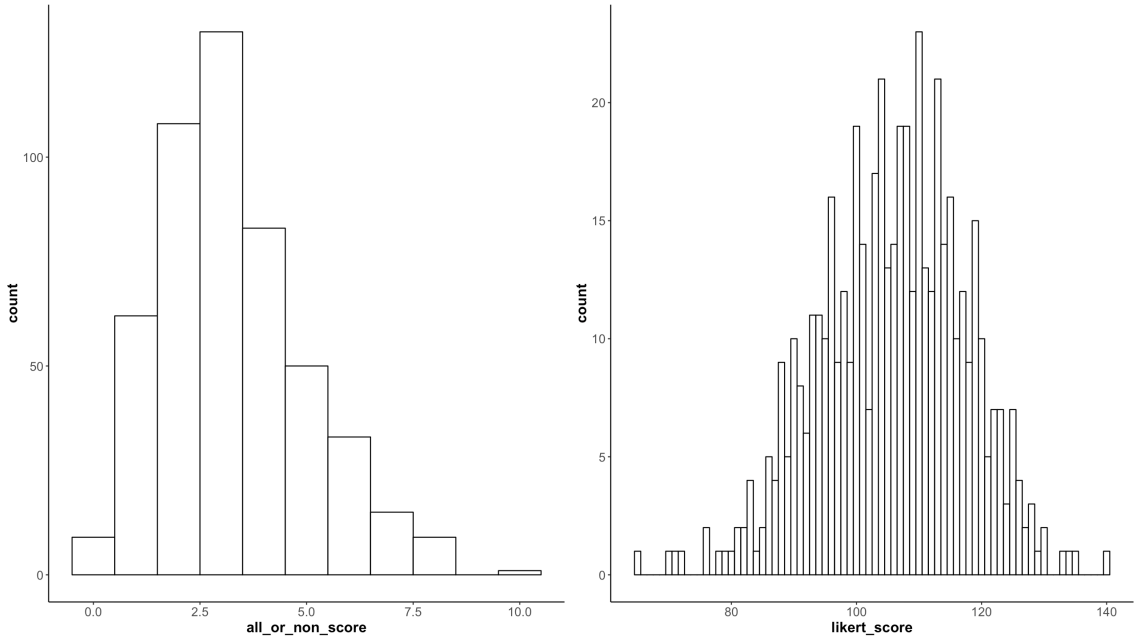

Figure 4. Histogram for all-or-none score and Likert-type score

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.

Table 3: Model Comparison

parameter	model	Correlation	average PSD
Ability	MHH-SDT	91.85%	0.009
	3PL	29.93%	0.014
	GPCM	53.9%	0.003
Difficulty	MHH-SDT	93.68%	0.086
	3PL	71.87%	0.020
	GPCM	81.97%	0.004
Discriminating power	MHH-SDT	85.90%	0.011
	3PL	15.48%	0.014
	GPCM	42.8%	0.003

Summary & Discussion

In this study, we have designed a hierarchical mixed membership model with a signal detection theory for the MAMC item. Signal detection theory is a psychology model and captures the decision-making behavior for every alternative. The mixed membership model captures the grouped data structure in the MAMC item. Different alternatives share the same partial membership scores, which measure the examinee's ability on different items.

Same as the IRT models, there are two underlying assumptions for HMM-SDT model: 1) Monotonicity: as the overall ability is increasing or the item difficulty is decreasing, the item level ability is increasing. 2) Local independence: The response of separate alternatives in an item is mutually independent, given a certain level of ability, and the ability of separate items are mutually independent, given a certain level of overall ability.

HMM-SDT provides a flexible framework for handling almost all kinds of multiple-choice items without requiring every item to has the same length or scoring process. The information and structure of response patterns are captured more completely. Exact–False item is a particular case where every MAMC item contains only one alternative. To analyze the SAMC items, we need to break the local independence assumption at the alternative level. Since the distracters' effect plays an important role and the decision-making behaviors are consequently not independent anymore. The limitation of current HMM-SDT is complete local independence at the alternative level. In reality, there is usually at least one essential alternative that should be selected in the MAMC item. However, our model allows the extreme situation that none of the alternatives are chosen. It leads to a slight underestimation of ability and overestimation of difficult.

Based on the simulation study, HMM-SDT has several advantages over the traditional IRT approach. The future study will test and compare the model in more dimensions and under more different simulation conditions with real data. MAMC items are widely used in model testing and have distinct benefits and flexibility in application. We hope that the present article will encourage researchers to use and do more research on the HMM-SDT model and MAMC item. The result will be a more in-depth and more informative analysis.

Appendix

Stan Code for HMM-SDT.

```
data {
  int<lower=0> K; // #  alternative
  int<lower=0> I; // #  examinee
  int<lower=0> J; // #  item
  int y[I,J,K];   // observations
  int z[J,K];     // true answer
}

parameters {
  real c[J,N];
  real<lower=0> d[J];
  real<lower=0,upper=1> lambda[I,J];
}

model {
  d ~ lognormal(0,1);
  for (j in 1:J){
    for (n in 1:N){
      c[j,n] ~ normal((1/2)*d[j]*z[j,n],1);
     for (i in 1:I){
            y[i,j,n] ~ bernoulli( lambda[i,j] *
            (1 - normal_cdf(c[j,n] - d[j] * z[j,n],0,1)) +
            (1 - lambda[i,j]) * (1 - normal_cdf(c[j,n],0,1)));
          }
        }
    }
}
```

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*. doi: 10.1007/BF02293814

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2000). Latent Dirichlet Allocation David. *CrossRef Listing of Deleted DOIs*. doi: 10.1162/jmlr.2003.3.4-5.993

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*. doi: 10.18637/jss.v076.i01

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. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods*. doi: 10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2003). Using the PLUM procedure of SPSS to fit unequal variance and generalized signal detection models. *Behavior Research Methods, Instruments, and Computers*. doi: 10.3758/BF03195496

DeCarlo, L. T. (2014). STUDIES OF A LATENT-CLASS SIGNAL-DETECTION MODEL FOR CONSTRUCTED-RESPONSE SCORING. *ETS Research Report Series*. doi: 10.1002/j.2333-8504.2008.tb02149.x

Duncan, G. T., & Milton, E. O. (1978). Multiple-answer multiple-choice test items: Responding and scoring through bayes and minimax strategies. *Psychometrika*. doi: 10.1007/BF02294088

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

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences Linked references are available on JSTOR for this article : Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*.

Hautus, M. (2015). Signal Detection Theory. In *International encyclopedia of the social & behavioral sciences: Second edition.* doi: 10.1016/B978-0-08-097086-8.43090-4

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler : Adaptively Setting Path Lengths. *Journal of Machine Learning Research*.

Ingleby, J. (2003). Signal detection theory and psychophysics. *Journal of Sound and Vibration*. doi: 10.1016/0022-460x(67)90197-6

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*. doi: 10.1007/BF02296272

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*. doi: 10.1177/014662169201600206

Muraki, E. (1993). Information Functions of the Generalized Partial Credit Model. *Applied Psychological Measurement*. doi: 10.1177/014662169301700403

R Development Core Team. (2008). Computational Many-Particle Physics. *R Foundation for Statistical Computing*, *739*. Retrieved from `http://link.springer.com/10.1007/978-3-540-74686-7` doi: 10.1007/978-3-540-74686-7

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*. doi: 10.3758/BF03207704

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science Science*. doi: 10.1126/science.3287615

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*. doi: 10.1007/BF02302588