Supplementary Materials for "Learning Large Q-matrix by Restricted Boltzmann Machines"

Chengcheng Li, Chenchen Ma, and Gongjun Xu

Department of Statistics, University of Michigan*

This file contains additional simulation results in Section 1 and the proofs of all lemmas and propositions in Section 2.

1 Additional Simulation Studies

1.1 Estimating Randomly Sampled Q-Matrix

In this section, we consider randomly sampled Q-matrix in a way that can simulate potentially more challenging scenarios. In specific, we include the one-, two- and three-attribute item designs. The exact construction of the Q-matrix is as follows. Similar to the construction in the main article, we still fix the dimension of the Q-matrix to be $3K$ by K, i.e. $3K$ items with K attributes. For each row j, we first determine which item design it will take by a random sampling scheme. Let $M = \binom{K}{1} + \binom{K}{2} + \binom{K}{3}$. The number of required attributes (denoted by n) for each item is randomly sampled from $\{1, 2, 3\}$ with probabilities $\{\binom{K}{1}/M, \binom{K}{2}/M, \binom{K}{3}/M\}$. Then, n attributes are sampled without replacement from $\{1, 2, ..., K\}$ with equal probabilities, the corresponding entries in $\mathbf{q}_j$ will be set to 1 and the rest to 0. Note that this random construction of the Q-matrix would somewhat simulate the extreme situations where the easiest learned one-attribute items will be sampled with the smallest probabilities. For example, when $K = 15$, the probability to select a one-attribute item is only 0.0261. Furthermore, we also point out that under this random design, there will be a high chance the sampled Q-matrix is not identifiable, making the estimation even more difficult.

*This research is partially supported by NSF CAREER SES-1846747, DMS-1712717, SES-1659328.

100 replications for each of $K = 5, 10, ..., 25$ are considered and the average results are presented in Figure 1. For illustration purpose, we only consider the settings when $N = 2000$ and when the attributes are independent, for the DINA, the ACDM, and a mixture of the DINA, ACDM, and DINO data. For the data from a mixture of three models, the data are generated from the DINA, ACDM and DINO models with proportions 0.35, 0.35, and 0.3 respectively, respectively. All the other set-ups remain the same as the independent settings in Section 4 of the main article.

From Figure 1, we can observe that the OE's of our proposed method remain controlled for three types of data. However, we can also see that the OE's worsen and the OTP's become much more volatile compared to the fixed Q-matrix design in Section 4 of the main article. This is not surprising because of the increased difficulty in the design where the Q-matrices contain more two- and three-attribute items and the number of non-identifiable Q-matrices increases significantly. In line with our observations in the main article, we also observe the increased uncertainty level impact most negatively on the OTP. However, overall, the proposed method still possesses certain degrees of learning power of the Q-matrix even in such extreme situations.

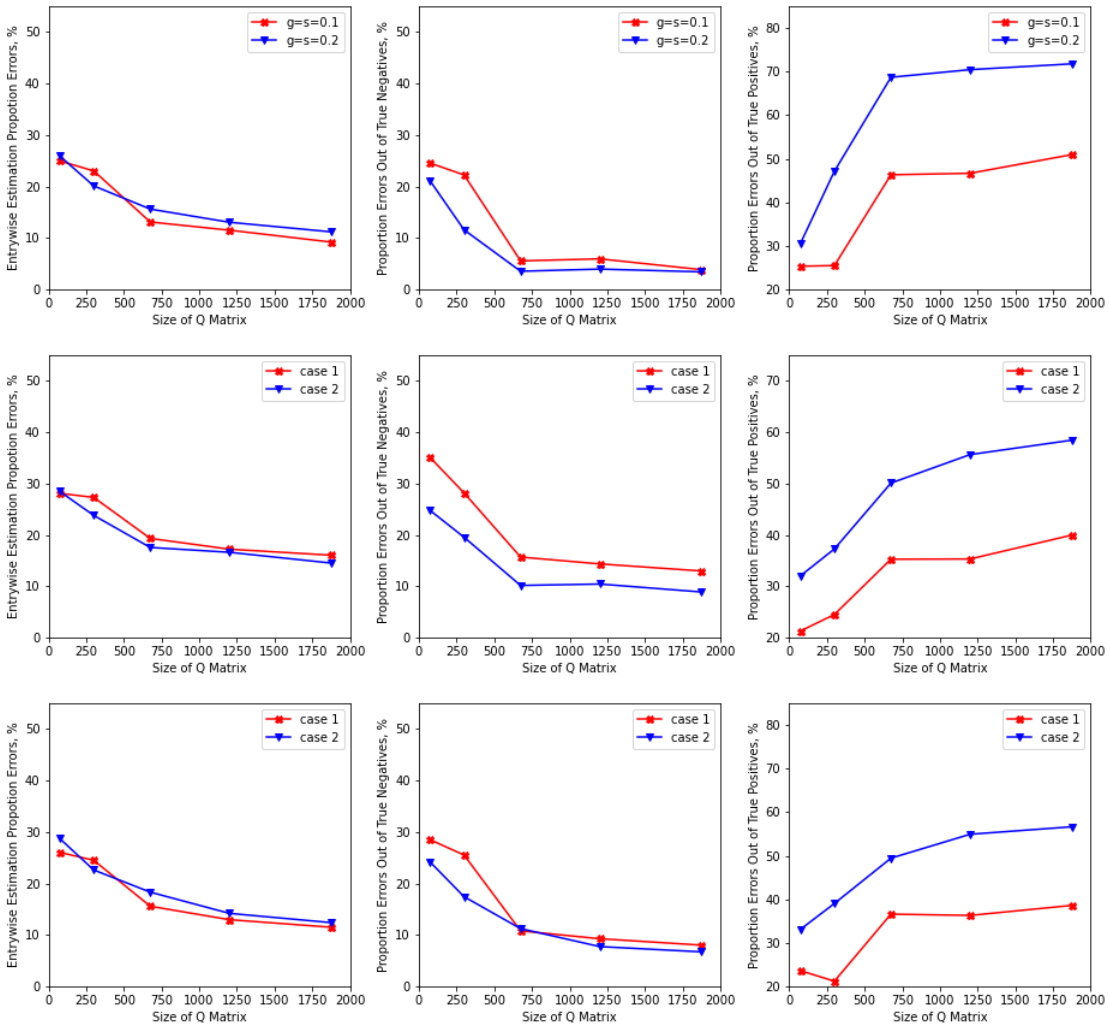

Figure 1: Plots of different performance metrics against the sizes of the Q-matrix. Rows 1 to 3 correspond to the DINA data, the ACDM data and a mixture of the DINA, ACDM and DINO data, respectively. For the DINA and DINO data, two uncertainty levels are represented by $g_j = s_j = 0.1$ and $g_j = s_j = 0.2$ for all items j, where subscripts j are omitted in the legends. For both the ACDM data and the GDINA data, cases 1 and 2 represent the settings when $\delta_{j,0} = 0.1$, $p_j = 0.9$ and $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$ respectively.

1.2 Attribute Classifications in Correlated Settings

In this section, we explore the potential of our proposed method in learning the latent attribute patterns. As discussed in the main article, the marginal distributions of the latent attributes are mis-specified in RBMs. Therefore, we would like to explore to what extent our proposed method can perform latent attribute classifications directly when the conditional independence assumption is intensely violated. Similarly, ACC rate is used to assess the performance. Recall that the ACC

of the k'th attribute is defined as

$$ACC(k) := \frac{1}{N}\sum_{i=1}^{N}|\hat{\alpha}_{ik} - \alpha_{ik}|,$$

where $\hat{\alpha}_{ik}$ and α_{ik} represent the estimated value and the true value respectively.

The simulation set-ups remain the same as the dependent settings in Section 4 of the main article. The recovered latent attribute matrix corresponding to the optimal estimated Q-matrix is returned. All the DINA, ACDM and GDINA data are considered. For each of the 100 replications, the ACC rate for every attribute in each of the settings with $K = 5, 10, ..., 25$ is evaluated. The setting-wise average ACC rate is evaluated by computing the average ACC for each attribute out of 100 repetitions first, and then averaging out of all the K latent attributes for each settings of $K = 5, 10, ..., 25$. The results are summarized in Table 1.

Overall, we can see that the proposed method performs well in attribute classifications with all ACC rates above 0.85. Furthermore, we also observe that the ACC rates drop as the number of attributes increases in the model. The attribute patterns would increase as the number of attributes increments, making the estimation more difficult. Similar to the observations made in the main article, we see the ACC rates are generally higher when the correlations amongst attributes are higher. We also point out that increasing sample size can in general improve ACC rates using the proposed method. The performance of the proposed method is better on the ACDM data and the GDINA data than on the DINA data. This is especially obvious when K is relatively small at 5 and 10. This observation is in line with our discussions in Section 2.3 of the main article.

$N = 2000$						$N = 10000$					
$\rho = 0.25$			$\rho = 0.75$			$\rho = 0.25$			$\rho = 0.75$		
DINA	ACDM	GDINA	DINA	ACDM	GDINA	DINA	ACDM	GDINA	DINA	ACDM	GDINA
0.898	0.916	0.916	0.917	0.927	0.924	0.903	0.916	0.917	0.918	0.932	0.931
0.897	0.896	0.900	0.888	0.902	0.903	0.901	0.907	0.911	0.885	0.911	0.912
0.878	0.876	0.880	0.880	0.888	0.893	0.891	0.887	0.893	0.880	0.897	0.900
0.875	0.863	0.869	0.879	0.885	0.889	0.883	0.879	0.882	0.874	0.894	0.893
0.866	0.853	0.857	0.875	0.883	0.887	0.877	0.868	0.874	0.874	0.887	0.890

Table 1: Average ACC rates for using RBM on the DINA data, the ACDM data and the GDINA data. Rows 1 to 5 correspond to the settings with $K = 5, 10, ..., 25$ respectively.

2 Proofs of Lemmas and Propositions

Before proving our main propositions 2.1 and 2.2, we first give a lemma which would be used in the proof of the main propositions.

Lemma 1. *Assume* $\boldsymbol{\alpha}$ *are independent and* $\alpha_k \sim Ber(p_k)$ *for* $k = 1, ..., K$. *If true model with response* R *satisfies either the GDINA model Equation (3) or the DINA model* $P(R = 1 \mid \boldsymbol{\alpha}) = g + (1 - s - g)\alpha_1\alpha_2...\alpha_{K^*}$ *for some* s, g *satisfying* $g < 1 - s$, *then the mis-specified linear additive model of* R *regressed on* $(\alpha_1, \alpha_2, ..., \alpha_K)$ *has the corresponding mean function in the form of* $\mathbb{E}^*[R \mid \boldsymbol{\alpha}] = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + ... + \beta_K\alpha_K$ *with* $\beta_k = 0$ *for* $k = K^* + 1, ..., K$.

Proof of Lemma 1. By the independence assumption and the linear regression theory, we have for $k = 1, \ldots, K$,

$$\begin{aligned}\beta_k &= \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, R\big) \\ &= \frac{1}{p_k(1-p_k)} Cov\big(\alpha_k, R\big).\end{aligned}$$

Denote $\alpha_{1,\ldots,K^*} := \{\alpha_1, ..., \alpha_{K^*}\}$, then by the Law of Total Covariance, we have for $k = K^*+1, ..., K$,

$$Cov\big(\alpha_k, R\big) = \mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\ldots,K^*}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*}\right], \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*}\right]\big). \tag{1}$$

Applying the independence assumption again, we have

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*}\right], \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*}\right]\big) = Cov\big(p_k, \mathbb{E}[R \mid \alpha_{1,\ldots,K^*}]\big) = 0.$$

Hence, we only need to consider the first term of (1). Referring to Figure 2, we know that in both the DINA and the GDINA model setting, $R \perp\!\!\!\perp \alpha_k \mid \alpha_{1,\ldots,K^*}$ for all $k = K^* + 1, ..., K$.

$$\mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\ldots,K^*}\big)\right] = 0.$$

Therefore,

$$\beta_k = \frac{0}{p_k(1-p_k)} = 0 \quad \forall k = K^* + 1, ..., K.$$

□

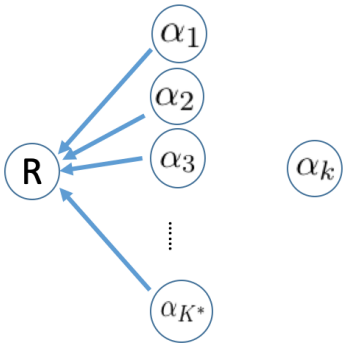

Figure 2: Illustration of the conditional independence relationship between R and α_k given $\alpha_1, ..., \alpha_{K^*}$ for all $k = K^* + 1, ..., K$

.

Next we give the proofs of our main propositions.

Proof of Proposition 1. First note that by Lemma 1, we have $\beta_k = 0$ for $k = K^* + 1, ..., K$.

In the DINA setting, we have

$$P(R = 1 \mid \boldsymbol{\alpha}) = \begin{cases} 1 - s & \text{if } \boldsymbol{\alpha} \succcurlyeq \mathbf{1}_{K^*} \\ g & \text{otherwise,} \end{cases}$$

or,

$$R \mid \boldsymbol{\alpha} \sim \begin{cases} \text{Ber}(1 - s) & \text{if } \boldsymbol{\alpha} \succcurlyeq \mathbf{1}_{K^*} \\ \text{Ber}(g) & \text{otherwise.} \end{cases} \tag{2}$$

Under the independence condition, for any $k = 1, ..., K^*$, we have

$$\beta_k = \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, R\big) = \frac{1}{p_k(1 - p_k)} Cov(\alpha_k, R).$$

Consider the following two events which partition the sample space of $\boldsymbol{\alpha}$, $E_{0,k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*} \mid \prod_{i=1, i\neq k}^{K^*} \alpha_i = 0\big\}$ and $E_{1,k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*} \mid \prod_{i=1, i\neq k}^{K^*} \alpha_i = 1\big\}$. Denote $\alpha_{1,...,K^*\setminus k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*}\big\}$. By the Law of Total Covari-

ance, we have

$$Cov\big(\alpha_k, R\big) = \mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\dots,K^*\setminus k}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k|\alpha_{1,\dots,K^*\setminus k}\right], \mathbb{E}\left[R \mid \alpha_{1,\dots,K^*\setminus k}\right]\big). \quad (3)$$

Applying the independence condition,

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\dots,K^*\setminus k}\right], \mathbb{E}\left[R \mid \alpha_{1,\dots,K^*\setminus k}\right]\big) = Cov\big(p_k, \mathbb{E}\left[R \mid \alpha_{1,\dots,K^*\setminus k}\right]\big) = 0.$$

Hence, we only need to consider the first term of (3),

$$\mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\dots,K^*\setminus k}\big)\right] = \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid \alpha_{1,\dots,K^*\setminus k}\right] - \mathbb{E}\left[\alpha_k \mid \alpha_{1,\dots,K^*\setminus k}\right] \cdot \mathbb{E}\left[R \mid \alpha_{1,\dots,K^*\setminus k}\right]\right]. \quad (4)$$

For a fixed k, define another two events: $E_{2,k} := \{\boldsymbol{\alpha} \mid \alpha_k = 0\}$ and $E_{3,k} := \{\boldsymbol{\alpha} \mid \alpha_k = 1\}$. Then in the event of $E_{0,k}$,

$$\begin{aligned}
(4) &= \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_{0,k}\right] - \mathbb{E}\left[\alpha_k \mid E_0\right]\mathbb{E}\left[R \mid E_{0,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_{0,k}, E_{3,k}\right]P(E_{3,k}) + \mathbb{E}\left[\alpha_k R \mid E_{0,k}, E_{2,k}\right]P(E_{2,k}) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E_{0,k}\right]\right] \\
&= \mathbb{E}\left[g \cdot p_k - p_k \cdot g\right] \\
&= 0.
\end{aligned}$$

In the event of $E_{1,k}$,

$$\begin{aligned}
(4) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_{1,k}\right] - \mathbb{E}\left[\alpha_k \mid E_{1,k}\right]\mathbb{E}\left[R \mid E_{1,k}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_{1,k}, E_{3,k}\right]P(E_{3,k}) + \mathbb{E}\left[\alpha_k R \mid E_{1,k}, E_{2,k}\right]P(E_{2,k})\right. \\
&\left. - \mathbb{E}\left[\alpha_k\right] \cdot \mathbb{E}\left[R \mid E_{1,k}, E_{3,k}\right] \cdot P(E_{3,k}) - \mathbb{E}\left[\alpha_k\right] \cdot \mathbb{E}\left[R \mid E_{1,k}, E_{2,k}\right] \cdot P(E_{2,k})\right] \\
=& \mathbb{E}\left[(1-s)p_k + 0 - p_k(1-s)p_k - p_k g(1-p_k)\right] \\
=& p_k(1-p_k)(1-s-g).
\end{aligned}$$

Since the above reasoning works for any $k = 1, 2, ..., K^*$, we must have for each $k = 1, 2, ..., K^*$,

$$\begin{aligned}
\beta_k &= \frac{1}{p_k(1-p_k)} Cov\big(\alpha_k, R\big) \\
&= \frac{1}{p_k(1-p_k)} \big(0 \cdot P(E_{0,k}) + p_k(1-p_k)(1-s-g) \cdot P(E_{1,k})\big) \\
&= (1-s-g) \prod_{i=1, i\neq k}^{K^*} p_i \\
&\neq 0.
\end{aligned}$$

□

Proof of Proposition 2. Note that by Lemma 1, we have $\beta_k = 0$ for $k = K^* + 1, ..., K$.

Under the independence condition, for any $k = 1, ..., K^*$, we have

$$\begin{aligned}
\beta_k &= \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, R\big) \\
&= \frac{1}{p_k(1-p_k)} Cov(\alpha_k, R). \qquad (5)
\end{aligned}$$

Denote $S := \big\{1, 2, 3, ..., K^*\big\}$. We consider the following 2^{K^*} events: $E_0 := \big\{\boldsymbol{\alpha} \mid \alpha_l = 0, \forall l \in S\big\}$, $E_{1,i} := \big\{\boldsymbol{\alpha} \mid \alpha_i = 1, \alpha_j = 0, \forall j \neq i \in S\big\}$ for some $i \in S$ (i.e. events that only one of the required variables taking value of 1 and all others being 0), $E_{2,(i,j)} := \big\{\boldsymbol{\alpha} \mid \alpha_i = \alpha_j = 1, \alpha_k = 0, \forall k \neq i, j \in S\big\}$ for some $i \neq j \in S$ (i.e. events that any two of the required variables are 1 and all others being 0), ..., $E_{K^*} := \big\{\boldsymbol{\alpha} \mid \alpha_l = 1, \forall l \in S\big\}$. Note that $E_0, E_{1,i}$ for $i \in S$, $E_{2,(i,j)}$ for some $i \neq j \in S$, ..., E_{K^*} partition the sample space of $\boldsymbol{\alpha}$. The response R would have the following distribution.

$$R|\boldsymbol{\alpha} \sim \begin{cases}
\text{Ber}(\delta_0) & \text{if } E_0 \\
\text{Ber}(\delta_0 + \delta_i) & \text{if } E_{1,i} \\
\text{Ber}(\delta_0 + \delta_i + \delta_j + \delta_{i,j}) & \text{if } E_{2,(i,j)} \\
... & \\
\text{Ber}\big(\delta_0 + \sum_{k=1}^{K^*} \delta_k + ... + \delta_{12...K^*}\big) & \text{if } E_{K^*}.
\end{cases} \qquad (6)$$

By the Law of Total Covariance, we have

$$Cov\big(\alpha_k, R\big) = \mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\ldots,K^*\backslash k}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k|\alpha_{1,\ldots,K^*\backslash k}\right], \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*\backslash k}\right]\big). \quad (7)$$

Similar to the DINA case, we also have

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*\backslash k}\right], \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*\backslash k}\right]\big) = Cov\big(p_k, \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*\backslash k}\right]\big) = 0.$$

Hence, we only need to consider the first term of (7),

$$\mathbb{E}\left[Cov\big(\alpha_k, R \mid \alpha_{1,\ldots,K^*\backslash k}\big)\right] = \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid \alpha_{1,\ldots,K^*\backslash k}\right] - \mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*\backslash k}\right] \cdot \mathbb{E}\left[R \mid \alpha_{1,\ldots,K^*\backslash k}\right]\right]. \quad (8)$$

Fix a $k \in S$. Let $S' := \big\{1, 2, ..., k-1, k+1, ..., K^*\big\}$. We can define new 2^{K^*-1} events: $E_0^* := \big\{\alpha_{1,\ldots,K^*\backslash k} \mid \alpha_l = 0 \quad \forall l \in S'\big\}$, $E_{1,i}^* := \big\{\alpha_{1,\ldots,K^*\backslash k} \mid \alpha_i = 1, \alpha_l = 0, \forall l \neq i \in S'\big\}$ for some $i \in S'$, $E_{2,(i,j)}^* := \big\{\alpha_{1,\ldots,K^*\backslash k} \mid \alpha_i = \alpha_j = 1, \alpha_l = 0, \forall l \neq i, j \in S'\big\}$ for some $i \neq j \in S'$,..., $E_{K^*-1}^* := \big\{\alpha_{1,\ldots,K^*\backslash k} \mid \alpha_l = 1 \quad \forall l \in S'\big\}$. And define $E_0' := \big\{\boldsymbol{\alpha} \mid \alpha_k = 0\big\}$ and $E_1' := \big\{\boldsymbol{\alpha} \mid \alpha_k = 1\big\}$.

In the event of E_0^*,

$$\begin{aligned}
(8) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_0^*\right] - \mathbb{E}\left[\alpha_k \mid E_0^*\right]\mathbb{E}\left[R \mid E_0^*\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E_0^*, E_1'\right]P(E_1') + \mathbb{E}\left[\alpha_k R \mid E_0^*, E_0'\right]P(E_0')\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E_0^*, E_1'\right]P(E_1') - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E_0^*, E_0'\right]P(E_0')\right] \\
=& \mathbb{E}\Big[(\delta_0 + \delta_k)p_k + (1-p_k)\cdot 0 - (\delta_0+\delta_k)p_k^2 - \delta_0(1-p_k)p_k\Big] \\
=& p_k(1-p_k)\delta_k.
\end{aligned}$$

In the event of $E^*_{1,i}$ for some $i \in S'$,

$$
\begin{aligned}
(8) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E^*_{1,i}\right] - \mathbb{E}\left[\alpha_k \mid E^*_{1,i}\right]\mathbb{E}\left[R \mid E^*_{1,i}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E^*_{1,i}, E'_1\right]P(E'_1) + \mathbb{E}\left[\alpha_k R \mid E^*_{1,i}, E'_0\right]P(E'_0)\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E^*_{1,i}, E'_1\right]P(E'_1) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E^*_{1,i}, E'_0\right]P(E'_0)\right] \\
=& \mathbb{E}\left[(\delta_0 + \delta_i + \delta_k + \delta_{ik})p_k + (1-p_k)\cdot 0 - (\delta_0 + \delta_i + \delta_k + \delta_{ik})p_k^2 - (\delta_0 + \delta_i)(1-p_k)p_k\right] \\
=& p_k(1-p_k)(\delta_k + \delta_{ik}).
\end{aligned}
$$

In the event of $E^*_{2,(i,j)}$ for some $i \neq j \in S'$,

$$
\begin{aligned}
(8) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E^*_{2,(i,j)}\right] - \mathbb{E}\left[\alpha_k \mid E^*_{2,(i,j)}\right]\mathbb{E}\left[R \mid E^*_{2,(i,j)}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k R \mid E^*_{2,(i,j)}, E'_1\right]P(E'_1) + \mathbb{E}\left[\alpha_k R \mid E^*_{2,(i,j)}, E'_0\right]P(E'_0)\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E^*_{2,(i,j)}, E'_1\right]P(E'_1) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[R \mid E^*_{2,(i,j)}, E'_0\right]P(E'_0)\right] \\
=& \mathbb{E}\left[(\delta_0 + \delta_i + \delta_j + \delta_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk})p_k + (1-p_k)\cdot 0\right. \\
& \left. - (\delta_0 + \delta_i + \delta_j + \delta_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk})p_k^2 - (\delta_0 + \delta_i + \delta_j + \delta_{ij})(1-p_k)p_k\right] \\
=& p_k(1-p_k)(\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk}).
\end{aligned}
$$

Continuing this process and substitute the relevant values into Equation (5), we can show that

$$
\beta_k = \begin{cases}
\delta_k & \text{if } E^*_0 \\
\delta_k + \delta_{ik} & \text{if } E^*_{1,i} \\
\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk} & \text{if } E^*_{2,(i,j)} \\
... & \\
\delta_k + \sum_{i=1, i\neq k}^{K^*} \delta_{ik} + ... + \delta_{1...K^*} & \text{if } E^*_{K^*-1}.
\end{cases} \tag{9}
$$

Since the above holds for all $k = 1, 2, 3, ...K^*$, we have for each $k = 1, 2, 3, ...K^*$,

$$\begin{aligned}\beta_k =&\delta_k \cdot P(E_0^*) + \sum_{i \in S'} (\delta_k + \delta_{ik}) \cdot P(E_{1,i}^*) + \sum_{i,j \in S', i \neq j} (\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk}) \cdot P(E_{2,(i,j)}^*) + ... \\ &+ \big(\delta_k + \sum_{i=1, i \neq k}^{K^*} \delta_{ik} + ... + \delta_{1...K^*}\big) \cdot P(E_{K^*-1}^*) \end{aligned} \tag{10}$$

Assuming monotonicity in acquiring an additional skill, we can show all the terms in (10) are greater than 0. The first term is positive as both δ_k and $P(E_0^*)$ are positive. To see why the second term is positive, consider two examinees, one with skill set $\boldsymbol{\alpha}_1 = \big\{\boldsymbol{\alpha} \mid \alpha_i = 1, \alpha_l = 0, \quad \forall l \neq i \in S\big\}$ while the other with skill set $\boldsymbol{\alpha}_2 = \big\{\boldsymbol{\alpha} \mid \alpha_i = \alpha_k = 1, \alpha_l = 0, \quad \forall l \neq i, k \in S\big\}$. Then we know according to Equation (3), $P(R = 1 \mid \boldsymbol{\alpha_1}) = \delta_0 + \delta_i$ and $P(R = 1 \mid \boldsymbol{\alpha_2}) = \delta_0 + \delta_i + \delta_k + \delta_{ik}$. The monotonicity assumption then implies $P(R = 1 \mid \boldsymbol{\alpha_2}) - P(R = 1 \mid \boldsymbol{\alpha_1}) = \delta_k + \delta_{ik} > 0$. Hence the second term is positive. We can use a similar strategy to show all the terms in (10) are positive and thus reach the conclusion that $\beta_k \neq 0$ for each $k = 1, 2, 3, ...K^*$. □

Discussion of Remark 2. Conditional on $\alpha_1, \alpha_2, ..., \alpha_{K^*}$, consider adding one α_k, for any $k = K^* + 1, ..., K$, into the main effect regression model, then its coefficient can be expressed as

$$\beta_k = \frac{Cov\Big(R - \mathbb{E}^*[R \mid \alpha_1, ..., \alpha_{K^*}], \quad \alpha_k - \mathbb{E}^*[\alpha_k \mid \alpha_1, ..., \alpha_{K^*}]\Big)}{Var\Big(R - \mathbb{E}^*[R \mid \alpha_1, ..., \alpha_{K^*}]\Big)},$$

where $\mathbb{E}^*[A \mid B]$ is the the regression mean function of A on B. In the special case when $K^* = 1$, we seek to show $\beta_k = 0$. When $K^* = 1$, note that we must have $\mathbb{E}^*[R \mid \alpha_1] = \mathbb{E}[R \mid \alpha_1]$. This is because α_1 can only take values of 0 or 1. These two variability's can be modeled exhaustively by the free intercept and the only coefficient in the regression mean function. Note that when $K^* > 1$, this may not hold in general. Note by the Law of Total Covariance,

$$\begin{aligned} &Cov\Big(R - \mathbb{E}^*[R \mid \alpha_1], \quad \alpha_k - \mathbb{E}^*[\alpha_k \mid \alpha_1]\Big) \\ =&\mathbb{E}\Big\{Cov\Big(R - \mathbb{E}[R \mid \alpha_1], \quad \alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big)\Big\} \qquad (11) \\ &+ Cov\Big\{\mathbb{E}\big(R - \mathbb{E}[R \mid \alpha_1] \mid \alpha_1\big), \quad \mathbb{E}\big(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\big)\Big\}. \qquad (12) \end{aligned}$$

Note (12) $= 0$ and

$$
\begin{aligned}
(11) &= \mathbb{E}\Big\{\mathbb{E}\Big[\big(R - \mathbb{E}[R \mid \alpha_1]\big)\big(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1]\big) \mid \alpha_1\Big] + \mathbb{E}\Big[\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big]\mathbb{E}\Big[\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big]\Big\} \\
&= \mathbb{E}\Big\{\mathbb{E}\Big[(R - \mathbb{E}[R \mid \alpha_1])(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1]) \mid \alpha_1\Big]\Big\} \\
&= \mathbb{E}\Big\{\mathbb{E}\Big[R\alpha_k - R\mathbb{E}(\alpha_k \mid \alpha_1) - \alpha_k\mathbb{E}(R \mid \alpha_1) + \mathbb{E}(R \mid \alpha_1)\mathbb{E}(\alpha_k \mid \alpha_1) \mid \alpha_1\Big]\Big\} \\
&= \mathbb{E}\Big\{\mathbb{E}[R\alpha_k \mid \alpha_1] - \mathbb{E}[R\alpha_k \mid \alpha_1] - \mathbb{E}[R\alpha_k \mid \alpha_1] + \mathbb{E}[R\alpha_k \mid \alpha_1]\Big\} \\
&= 0.
\end{aligned}
$$

Where the second line follows from $\mathbb{E}\Big[\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big] = 0$ and the third line follows from the fact that $\mathbb{E}[R \mid \alpha_1]\mathbb{E}[\alpha_k \mid \alpha_1] = \mathbb{E}[R\alpha_k \mid \alpha_1]$ by the conditional independence between R and α_k given α_1. Therefore, $\beta_k = 0$. □