A PRELIMINARY MODEL FOR JOINT INFERENCE OF HEARING CONDITIONS AND AGE THRESHOLDS

YONGFU LIAO

November 23, 2024

1. Introduction

Constructing items for assessing and diagnosing young children can be challenging due to the rapid pace of cognitive and neuromuscular development during this period. Typically, in a scale designed for such purposes, the latent conditions of interest are assumed to be signaled by the presence (or absence) of certain behaviors as indicated by the scale's items. However, this assumption may not hold when assessing very young children, as the presence of a behavior could depend more on the child's development than their latent condition. For example, consider items for detecting hearing loss in children. In such a case, an item like "My child often fails to understand long sentences" is only valid for detecting hearing loss in children who have developed the cognitive competence for understanding longer sentences. Therefore, to correctly infer a child's latent condition, it is essential to also consider their developmental status.

To address this, we developed a model that accounts for the influences of both development and latent conditions on item responses. The model enables one to (a) estimate the age boundaries within which items are valid and (b) classify individuals into binary conditions by weighing information across items according to their age.

2. Motivating Context

We begin by describing four types of items whose discriminative power (i.e., validity) changes with development, for which we use child age as a proxy. The subplots in Figure 1 correspond to the four item types, each specifying the relationship between child age and the probability of a "Yes" response[1]. The red curves plot the trajectories of children with hearing loss ($z = 1$) across ages, whereas the blue curves plot the trajectories for children with typical hearing ($z = 0$). Type-1 and Type-2 items (top row in Figure 1) represent items that are valid for discriminating hearing loss from typical hearing only *above* a certain age threshold, as there is minimal difference between children in the two conditions at younger ages. In contrast, Type-3 and Type-4 items are valid *below* a certain age threshold, often reflecting behaviors that fade out

[1]For the sake of illustration, it is assumed here that no item response needs reverse coding. That is, assuming the items are valid, the presence of hearing loss ($z = 1$) always leads to higher probabilities of "Yes" responses, compared to "No" responses, for all items.

as children mature, such as startle reflexes and breastfeeding-related behaviors. Some hypothetical items are provided at the bottom of the subplots in Figure 1.

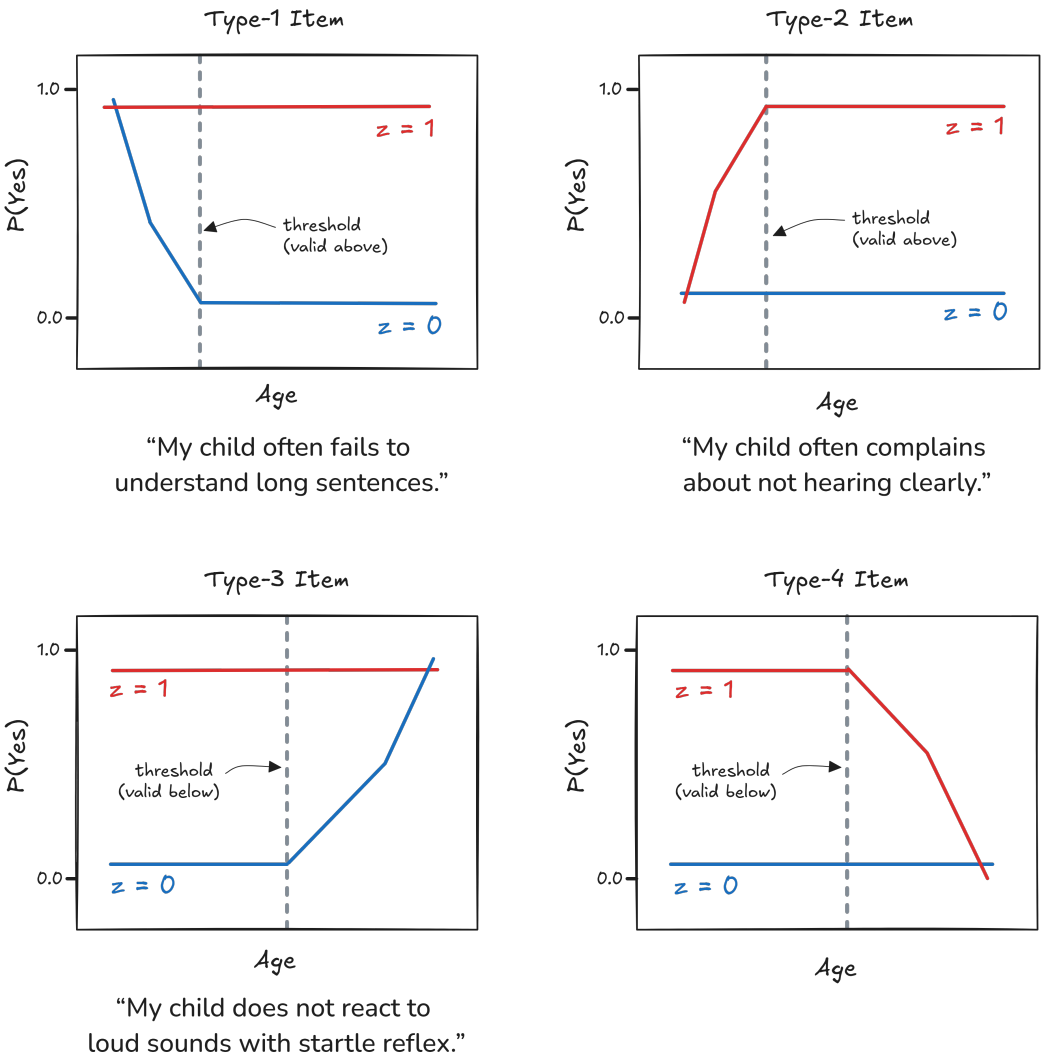

FIGURE 1. The four types of age-restricted items addressed in this manuscript. The horizontal axes in the plots represent child development, with age used as a proxy. The vertical axes represent the probability of a "Yes" item response. The red lines graph the trajectories of hearing loss ($z = 1$) children, and the blue lines represent typical hearing ($z = 0$) children. The four item types are therefore defined by two independent properties: (a) whether an item becomes discriminative (i.e., valid) *below* or *above* the age threshold, and (b) whether an invalid item results uniformly in a *high* or *low* probability of a "Yes" response in different groups.

3. Model Specification

With the four item types in mind, we now introduce our model. The model's logic is illustrated in the tree diagrams in Figure 2, which represent the (idealized) data-generating processes that map a child onto an expected item response according to two latent variables, k and z. The variable k specifies the *discriminative power* (validity) of an item given a child's age,

where $k = 1$ indicates maximal discrimination, and $k = 0$ indicates no discrimination. The variable z denotes a child's latent condition, which, in this context, is either hearing loss ($z = 1$) or typical hearing ($z = 0$). Thus, k, z, and the item's age-dependent response pattern codetermine the expected item response. Specifically, the left tree diagram depicts the data-generating process for Type-1 and Type-3 items (the left column in Figure 1), while the right diagram depicts the process for Type-2 and Type-4 items (the right column in Figure 1). For instance, one would expect a "Yes" response when the item is discriminative ($k = 1$) and the child has hearing loss ($z = 1$). However, if the item is non-discriminative ($k = 0$), a similar response would be expected regardless of the child's hearing condition. The description so far is *idealized* and *deterministic*; in practice, noise is expected. Therefore, at the bottom of the tree diagrams, we list the probability of a "Yes" response for each condition modeled. These probabilities correspond to the model's parameters, which we now turn to.

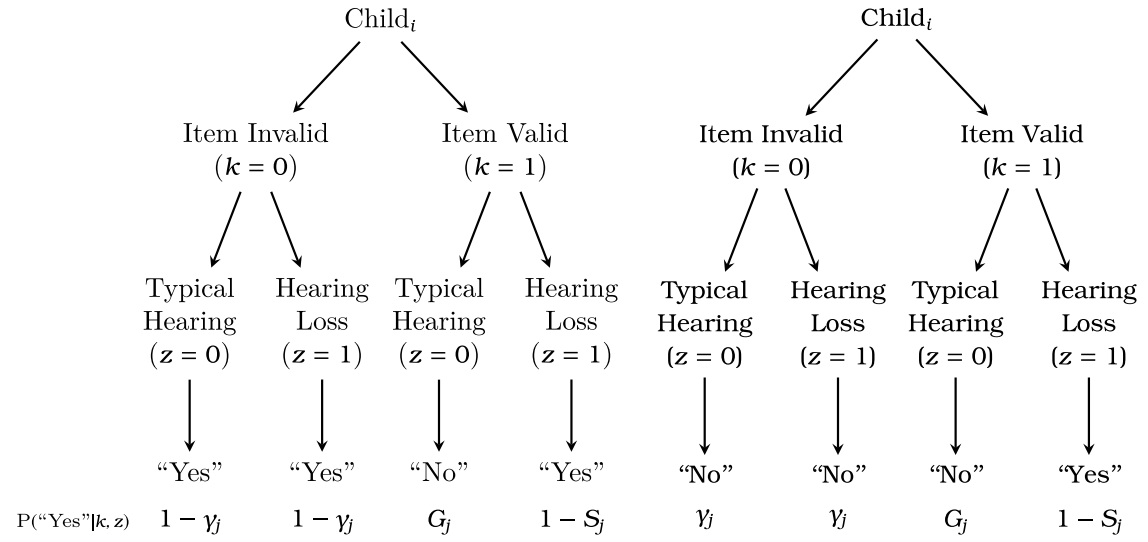

FIGURE 2. Idealized data-generating processes mapping a child to an expected item response according to the discriminative power k and hearing condition z. The left diagram corresponds to the process for Type-1 and Type-3 items, and the right diagram for Type-2 and Type-4 items. The terms at the bottom represent probabilistic versions of the corresponding idealized item responses to account for noise in real data.

In our model, we used the Bernoulli distribution to link a binary item response $\mathrm{Y}_{i,j}$, collected from item j for child i, to an underlying probability p.

$$\mathrm{Y}_{i,j} \sim \mathrm{Bernoulli}(p)$$

The probability p is determined by the item parameters S_j, G_j, and γ_j, the child's latent condition parameter z_i, the discriminative power k, and the item type, as shown below.

$$p = \begin{cases} \big(1 - S_j\big)^{z_i k} G_j^{(1-z_i)k} \big(1 - \gamma_j\big)^{1-k} & \text{[Type-1 \& Type-3 items]} \\ \big(1 - S_j\big)^{z_i k} G_j^{(1-z_i)k} \gamma_j^{1-k} & \text{[Type-2 \& Type-4 items]} \end{cases}$$

The terms making up p essentially formalize the relationships represented in the tree diagrams in Figure 2. By specifying the item type and substituting combinations of 0 and 1 for z_i and k (e.g., $z_i = 1$ and $k = 0$) into the above equation, p simplifies to a term corresponding to one of the eight probabilities at the bottom of Figure 2.

The item parameters S_j, G_j, and γ_j model deviations from ideal probabilities of 1 or 0. Specifically, S_j and G_j can be respectively thought of as the false-negative rate (or, 1 − sensitivity) and the false-positive rate (or, 1 − specificity) in a signal-detection context, or as the "slip" and "guess" parameters in diagnostic classification models (Rupp et al., 2010).

To link the discriminative power k to child age, k is modeled as a function of age and the item age threshold parameter δ_j. The function is set up so that, as the difference between the child's age and the item age threshold increases, k approaches either 0 or 1, depending on the item type. How fast k approaches 0 and 1 is governed by the parameter D (fixed across items), akin to the discrimination parameter in traditional item response models.

$$k = \begin{cases} \text{logit}^{-1}\big(D\big(\text{Age}_i - \delta_j\big)\big) & \text{[Type-1 \& Type-2 items]} \\ \text{logit}^{-1}\big(D\big(\delta_j - \text{Age}_i\big)\big) & \text{[Type-3 \& Type-4 items]} \end{cases}$$

Note that in the discussion of the data-generating process in Figure 2, k is assumed to be binary for simplicity. From here on, we treat k as continuous and bounded between 0 and 1.

By collecting the terms above and including age-unrestricted items, we arrive at the full model in (1). The final term, $z_i \sim \text{Bernoulli}(\pi)$, indicates that the latent condition z_i is generated from an underlying prevalence parameter π.

$$\begin{aligned} \text{Y}_{i,j} &\sim \text{Bernoulli}(p) \\ p &= \begin{cases} \big(1 - S_j\big)^{z_i k} G_j^{(1-z_i)k} \big(1 - \gamma_j\big)^{1-k} & \text{[Type-1 \& Type-3 items]} \\ \big(1 - S_j\big)^{z_i k} G_j^{(1-z_i)k} \gamma_j^{1-k} & \text{[Type-2 \& Type-4 items]} \\ \big(1 - S_j\big)^{z_i} \ G_j^{(1-z_i)} & \text{[Age-unrestricted items]} \end{cases} \\ k &= \begin{cases} \text{logit}^{-1}\big(D\big(\text{Age}_i - \delta_j\big)\big) & \text{[Type-1 \& Type-2 items]} \\ \text{logit}^{-1}\big(D\big(\delta_j - \text{Age}_i\big)\big) & \text{[Type-3 \& Type-4 items]} \end{cases} \\ z_i &\sim \text{Bernoulli}(\pi) \end{aligned} \tag{1}$$

Finally, the priors are specified in (2). Two points are worth noting. First, γ_j is constrained[2] to be bounded between 0 and 0.5 in order to consistently differentiate Type-1/3 items from Type-2/4 items, enabling model identification during fitting. Second, the mean and standard deviation for the normal prior of the delta parameter are set so that roughly 95% of the prior density encompasses the full age range of the data.

$$\begin{aligned} S_j, G_j, &\sim \text{Beta}(2,2) \\ 2\gamma_j &\sim \text{Beta}(2,2) \quad {\scriptstyle [0\le\gamma_j\le 0.5]} \\ \pi &\sim \text{Beta}(2,2) \\ \delta_j &\sim \text{Normal}\Big(\frac{1}{2}\max_i \text{Age}_i, \frac{1}{4}\max_i \text{Age}_i\Big) \\ D &\sim \text{Exponential}(1) \end{aligned} \tag{2}$$

3.1. **Connections to the DINA model.** Our model can be viewed as a modification of a two-attribute non-compensatory diagnostic classification model (Rupp et al., 2010). Specifically, it largely resembles the deterministic inputs, noisy "and" gate model (a.k.a. the DINA model) (Haertel, 1989; Junker & Sijtsma, 2001) in structure, where z_i and k are the two attributes. The difference is that, in our model, k is not a *static* attribute tied to a person but a joint function of both the person's age and the item's age threshold. The interaction between the two attributes also differs between the models, as reflected in the exponents of the G_j parameters below.

$$\text{P}\big(\text{Y}_{i,j}=1 \mid z_i, k_i, S_j, G_j\big) = \big(1-S_j\big)^{z_i k_i} G_j^{1-(z_i k_i)} \quad \text{[DINA Model]}$$

$$\text{P}\big(\text{Y}_{i,j}=1 \mid z_i, k, S_j, G_j, \gamma_j\big) = \big(1-S_j\big)^{z_i k} G_j^{(1-z_i)k}(...) \quad \text{[Our Model]}$$

3.2. **Modifications to meet practical demands.** We used several synthetic datasets to test the model's ability to recover the parameters of the data-generating process. Initial attempts revealed that the item parameters were recovered with poor precision when fitting the model specified in (1). The latent condition parameters z_i, on the other hand, were still correctly inferred, despite the high posterior variances in the item parameters.

As one of the primary goals of our model is to obtain precise item age threshold estimates for assessing item quality (i.e., to check whether these age estimates align with the literature and, if not, identify potential causes), we slightly modify the model to reduce the posterior variation in item parameters. Specifically, in our Stan (Carpenter et al., 2017) implementation of the model, the z_i's are *partially observed*, such that z_i is treated as *data* when person i's true condition is known and as a *latent discrete parameter* to be inferred when

[2]Through a scale transformation, $\gamma'_j \sim \text{Beta}(2,2)$, where $\gamma_j = \frac{1}{2}\gamma'_j$.

the true condition is unknown. This approach allows for a train-test split in which all z_i's are treated as data in the training phase, alleviating the burden of simultaneously estimating item and person parameters and consequently resulting in more precise recovery of the item parameters. This approach also aligns well with the applied scenario our model is ultimately targeting. Specifically, a trained model is *necessary* in such a context, where it provides predictions based on individuals' responses without refitting the full Bayesian model each time new data arrive.

Another modification to (1) is that we fix the prevalence parameter π to 0.5. This adjustment forces the model to rely solely on the information in the item responses, excluding any reliance on the latent conditions' base rates in the population when computing the posterior probabilities of the conditions. This decision reflects that the prevalence of hearing loss in the sample used to fit the model differs from that in the general population. Furthermore, since our questionnaire is intended as a checklist-like resource for concerned parents and practitioners, we do not have prior knowledge of hearing loss prevalence in such a context, nor do we plan to estimate it. Therefore, we believe it is reasonable for the model to disregard the base rates of the latent conditions.

4. Parameter Recovery Study

We now describe the parameter recovery study and discuss several prominent properties of our model that we have observed.

Parameter/Variable	**Simulated values**	**N**
Child age (Age_i)	Uniform(0, 36)	Train: 300 / Test: 300
Hearing condition (z_i)	0 or 1 (50% each)	Train: 300 / Test: 300
Age threshold (δ_j)	Fixed to 3, 9, 9, 9, 9, 12, 12, 12, 15, 15, 15, 24, 24, 30, 30, 36	16
Slip (S_j)	Uniform(.35, .9)	20
Guess (G_j)	Uniform(.02, .4)	20
γ_j	Uniform(0, .3)	20

Table 1. Parameter values used in the simulation. The sixteen age-restricted items are all set as Type-1 items.

4.1. **Simulation.** Three hundred subjects were simulated for training the model, and another three hundred for testing. The age distributions of the subjects were generated from a uniform distribution ranging from zero to thirty-six months. Sixteen items with age thresholds covering a similar age

range, along with four additional age-unrestricted items, were simulated. The simulation, summarized in Table 1, was designed to closely match our planned data collection scenario. In the simulation, we assign all items with age restrictions as Type-1 items.

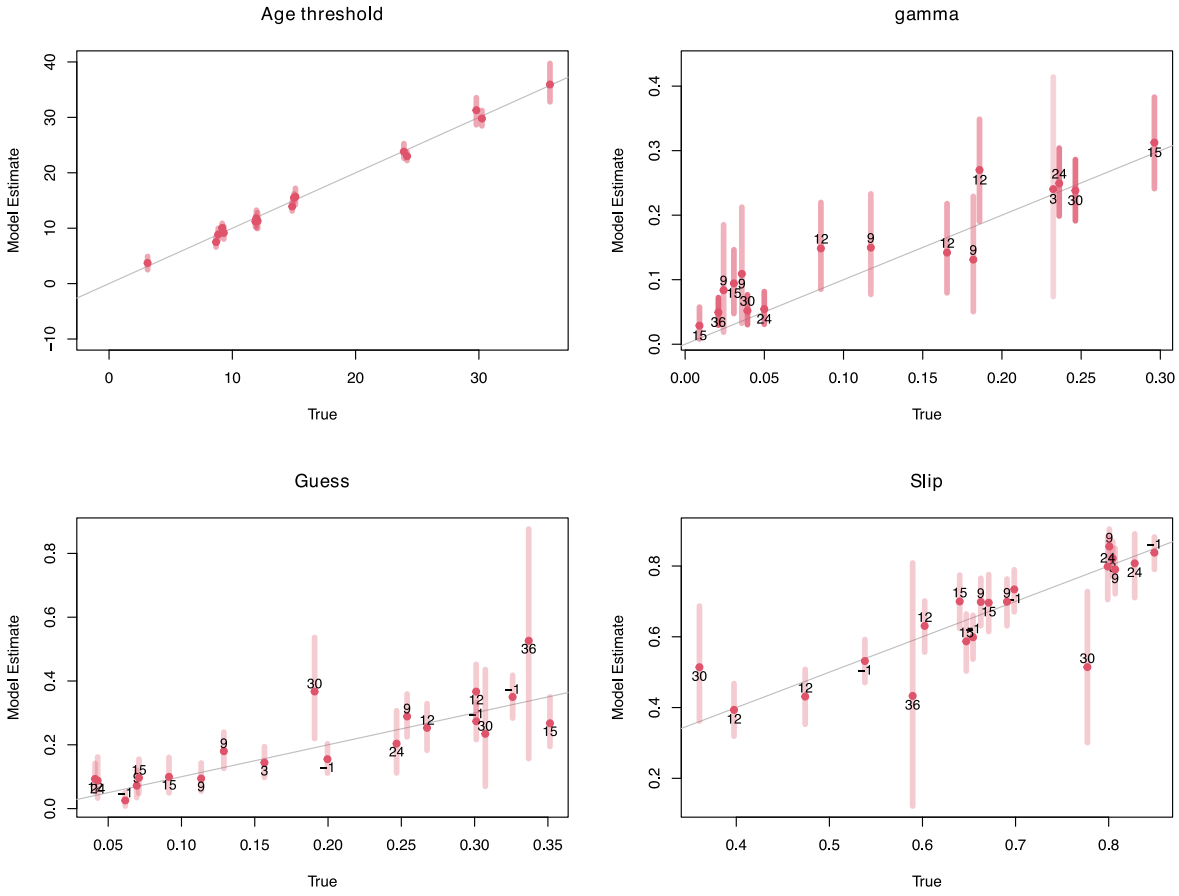

FIGURE 3. Recovery of the item parameters from the simulated data in Table 1. The horizontal axes show true values, and the vertical axes show corresponding posterior estimates. Red dots indicate the posterior means, and bars represent the central 95% posterior densities. True age thresholds are labeled[3] near the means in the plots for the gamma (γ_j), Guess (G_j), and Slip (S_j) parameters. As shown in these plots, the posterior variances of the gamma parameters are *negatively* associated with age thresholds, while those of the Guess and Slip parameters are *positively* associated with age thresholds.

4.2. **Parameter recovery.** Figure 3 plots the recovery of the item parameters by comparing the true (i.e., simulated) parameter values (horizontal axis) with their corresponding posterior estimates (vertical axis). The dots indicate posterior means, and the bars represent the central 95% posterior densities. Among the four types of item parameters listed, the age threshold parameters (top-left subplot) are the most reliably recovered, with the posterior means aligning closely with the gray identity line.

[3]Those labeled with −1 indicate age-unrestricted items.

The remaining parameters are generally recoverable, though extremely wide posteriors are observed for some. Indeed, for the γ_j, G_j, and S_j parameters, the posterior variances correlate with the item age thresholds (labeled as numbers next to the dots in the subplots): larger variances appear for items with higher age thresholds in the "Guess" and "Slip" parameters, while lower age thresholds show greater variance in the γ_j parameter. This effect arises from all items in the current simulation being Type-1 items. When Type-1 items have higher age thresholds, the "Guess" and "Slip" parameters have less available information, as younger samples cannot be used for estimation by model design. Conversely, more samples are available for estimating the γ_j parameter when a Type-1 item has a high age threshold. The pattern would reverse if Type-3 items were used. This is illustrated in Figure 4, which depicts the same parameter recovery as in Figure 3 but with Type-3 items used in the simulation instead.

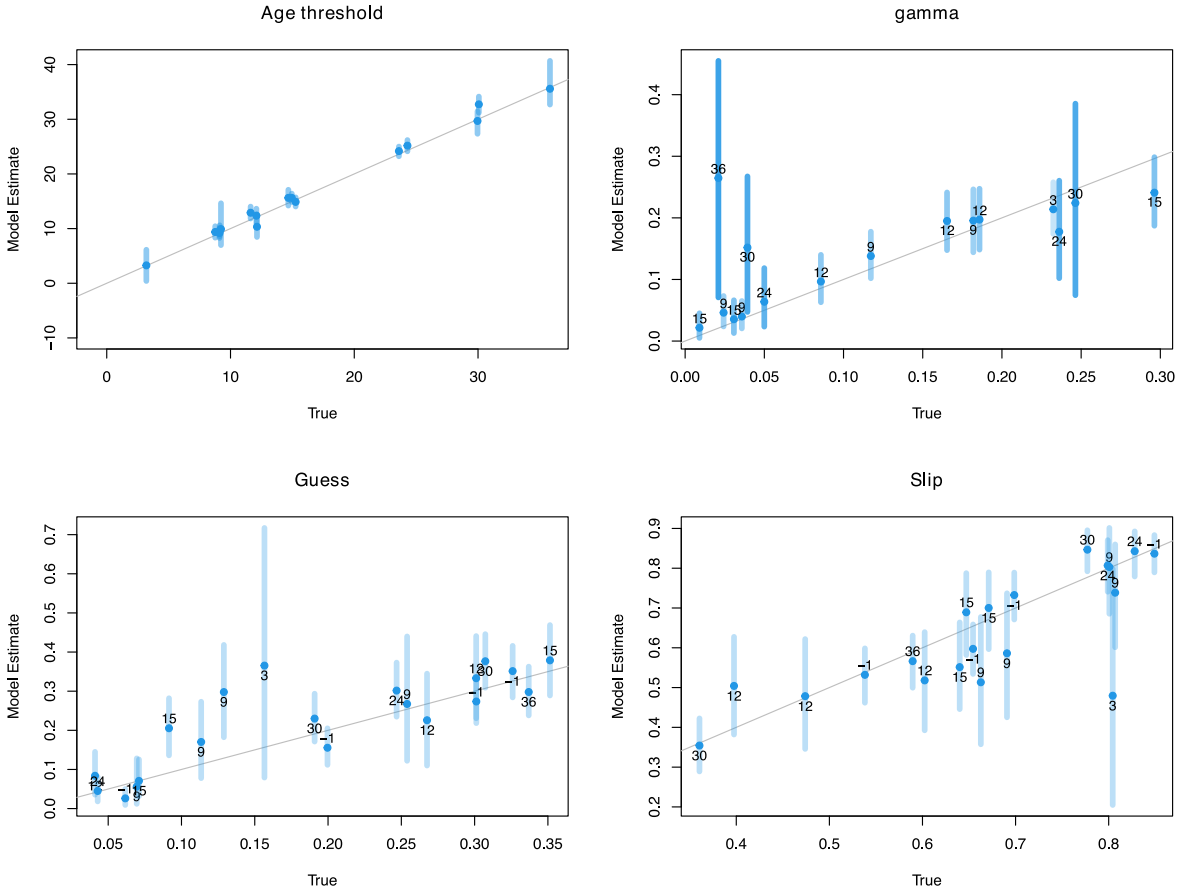

FIGURE 4. This figure shows the recovery of item parameters similar to that in Figure 3, with the only difference being that all age-restricted items are set to Type-3 instead of Type-1. The pattern of associations between posterior variances and age thresholds is now reversed.

4.3. **Predictions on hearing conditions.** After training the model, it is applied to the testing dataset to infer the subjects' hearing conditions. In this prediction phase, the item parameters are kept fixed at the values estimated

during the training phase. The individuals' hearing conditions (z_i) are now *unobserved* latent discrete parameters to be inferred, as discussed in Section 3.2.

Figure 5 depicts the inference of hearing conditions on the testing dataset. The horizontal axes in the plots indicate the true hearing conditions assigned in the simulation, and the vertical axes represent the mean of the posterior hearing loss probability, $\mathrm{P}(z_i = 1|\mathcal{D}, \mathcal{M})$, obtained from the model. To evaluate how well the hearing conditions are inferred, we set a mean posterior probability of 0.5 as the criterion for assigning individuals to a prediction of either hearing loss or typical hearing. This enables us to calculate the model's sensitivity and specificity, where sensitivity is defined as the probability of a positive test case given an individual with hearing loss, $\mathrm{P}(+|\mathrm{HL})$, and specificity is defined as the probability of a negative test case given an individual with typical hearing, $\mathrm{P}(-|\mathrm{TH})$. These are shown in the panels in Figure 5, where the three panels differ only in terms of the populations plotted: the left panel includes all subjects from the data, the central panel includes only subjects under 12 months old, and the right panel includes only those over 24 months old.

As can be seen from the plots in Figure 5, higher accuracies (in terms of both sensitivity and specificity) are observed for older subjects. This phenomenon follows naturally from the fact that the prediction of an older subject's hearing condition is based on information from more items compared to younger individuals. Similar to the discussion in Section 4.2, if Type-3 items had been used instead in the simulation, we would expect a reversed pattern, with higher accuracies observed for *younger* subjects.

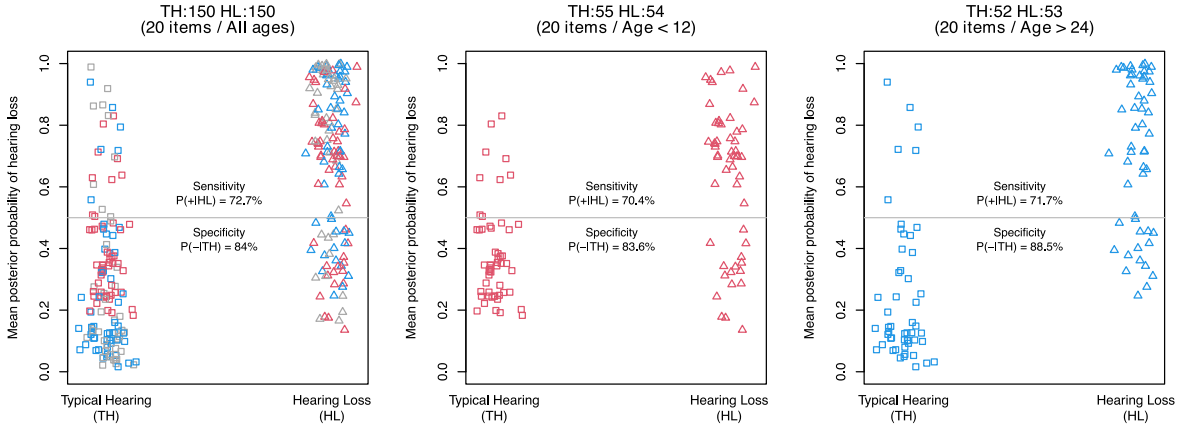

FIGURE 5. Predictions of subjects' hearing conditions in the testing dataset. The three panels each plots the predictions (based on the posterior probability of hearing loss) and their accuracies (sensitivity and specificity) for a subpopulation of the data (left: all subjects; center: subjects under 12 months old; right: subjects over 24 months old). The horizontal gray lines indicate a mean posterior probability of hearing loss of 0.5, which is the criterion set for assigning predicted labels.

5. Implications for Scale Development

To utilize our model effectively in real-world applications, several subtleties need to be addressed. First, item construction requires careful attention. A review of items in published questionnaires and milestone checklists (e.g., Wachtlin et al. (2017)) revealed that age-restricted items do not always correspond to any of the four item types in Figure 1. We found that items developed for young children often target behaviors that exist only within specific developmental stages (e.g., babbling). Such items are not properly handled by our model and therefore cannot be included in the scale as is. A workaround is to modify these items by incorporating an "or" statement to remove the upper or lower boundary of the developmental stage. For instance, the item "Makes a lot of different sounds like 'mamamama' and 'babababababa'"[4] could be appended with "*or* produces complete words" to create a Type-1 item. It is the authors' responsibility to ensure that items modified in this way are supported by the literature and backed by the empirical data collected.

Item selection also warrants careful consideration. Since parameter recovery depends on item type and age threshold (see Section 4.2), it is crucial to avoid constructing a scale by selecting items without considering how their discriminative power functions with age. Doing so is likely to result in inefficient (or even complete failure of) parameter estimation. Importantly, when an item's age threshold is close to the maximum or minimum age in the data, trade-offs arise in how precisely the error rate parameters (S_j and G_j) and the γ_j parameters can be recovered. One solution is to recruit participants with a broader age range than the expected range of all items' age thresholds. However, this may not always be feasible, particularly when the population of interest is very young children, making it impossible to cover ages below zero. In such cases, items expected to have an age threshold near zero might better be avoided.

Alternatively, strategically concentrating participants within specific age ranges can be effective. For instance, recruiting more participants with children under 3 months old could work in principle, but there are additional trade-offs to consider. Concentrating participants can improve parameter estimation for certain items but may worsen the estimation for other items with very different age thresholds. For example, if participants are concentrated to enhance estimates for items with a low age threshold, fewer participants remain available to estimate the error rate parameters for items with a higher age threshold, assuming all items are Type-1. That being said, when a scale includes multiple item types, such as a mix of Type-1 and Type-3 items, concentrating participants within particular age ranges might be advantageous. Therefore, whether

[4]https://www.cdc.gov/ncbddd/actearly/milestones/milestones-9mo.html

to concentrate, and which age range(s) to target, depends on the collective properties of the items involved. Simulations and parameter recovery studies are essential for addressing these complexities and provide a general approach to exploring the implications of various plausible item parameter patterns.

Finally, we highlight an unavoidable property of our model that becomes apparent in hindsight. Compared to a model with items that do not depend on age, our model requires a larger sample size to achieve the same level of precision. This is because the age restrictions on the items' discriminative power always result in fewer available samples for estimating item parameters than in the unrestricted case. Consequently, one should compensate for this limitation by maximizing sample-use efficiency through carefully considering the potential interactions among item types, age thresholds, and participants' age distribution. The exact effects of these interactions are case-specific and can only be reliably assessed through simulations and recovery studies. Therefore, model-based analysis should not only inform but also be integrated into conventional scale development practices to effectively address the additional challenges posed by items with age thresholds.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Haertel, E. H. (1989). Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items. *Journal of Educational Measurement*, *26*(4), 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

Wachtlin, B., Brachmaier, J., Amann, E., Hoffmann, V., & Keilmann, A. (2017). Development and Evaluation of the LittlEARS® Early Speech Production Questionnaire – LEESPQ. *International Journal of Pediatric Otorhinolaryngology*, *94*, 23–29. https://doi.org/10.1016/j.ijporl.2017.01.007

Children's Hearing Foundation, Taipei, Taiwan
Email address: tomliao@chfn.org.tw
URL: https://yongfu.name