Uncovering Order Effects with an IRT Mixture Model

Klint Kanopka

4/22/2020

1 Introduction

Previous work has found that for the NWEA MAP assessment, there are differences in both response accuracy and response speed that evolve over the course of the test. This work uses an IRT mixture model to identify changes in the item response function (IRF) over the course of the test.

The mixture model proposed below requires a three fundamental assumptions:

1. Each item has two difficulties associated with it, one when encountered early in the test and one when encountered late in the test.
2. An individual's probability of correct response depends on three quantities: the difference between their latent ability and the early test difficulty, the difference between their latent ability and the late test difficulty, and the difference between a person-specific "endurance" parameter and the location of the item in the test.
3. The mixture of early and late test contributions is a monotonic function of item location such that as items appear later on the test, their late test parameter contribution is higher.

2 Model

The model fit is a straightforward mixture of two Rasch IRFs:

$$P(X_{ij} = 1|\theta_i, b_{je}, b_{jl}, \pi_{ij}) = \pi_{ij}\sigma\big(\theta_i - b_{je}\big) + (1 - \pi_{ij})\sigma\big(\theta_i - b_{jl}\big)$$

Where:

- θ_i is student i's latent ability
- b_{je} is item j's difficulty when encountered early in the test
- b_{jl} is item j's difficulty when encountered late in the test
- σ is the standard logistic sigmoid function, specified by:

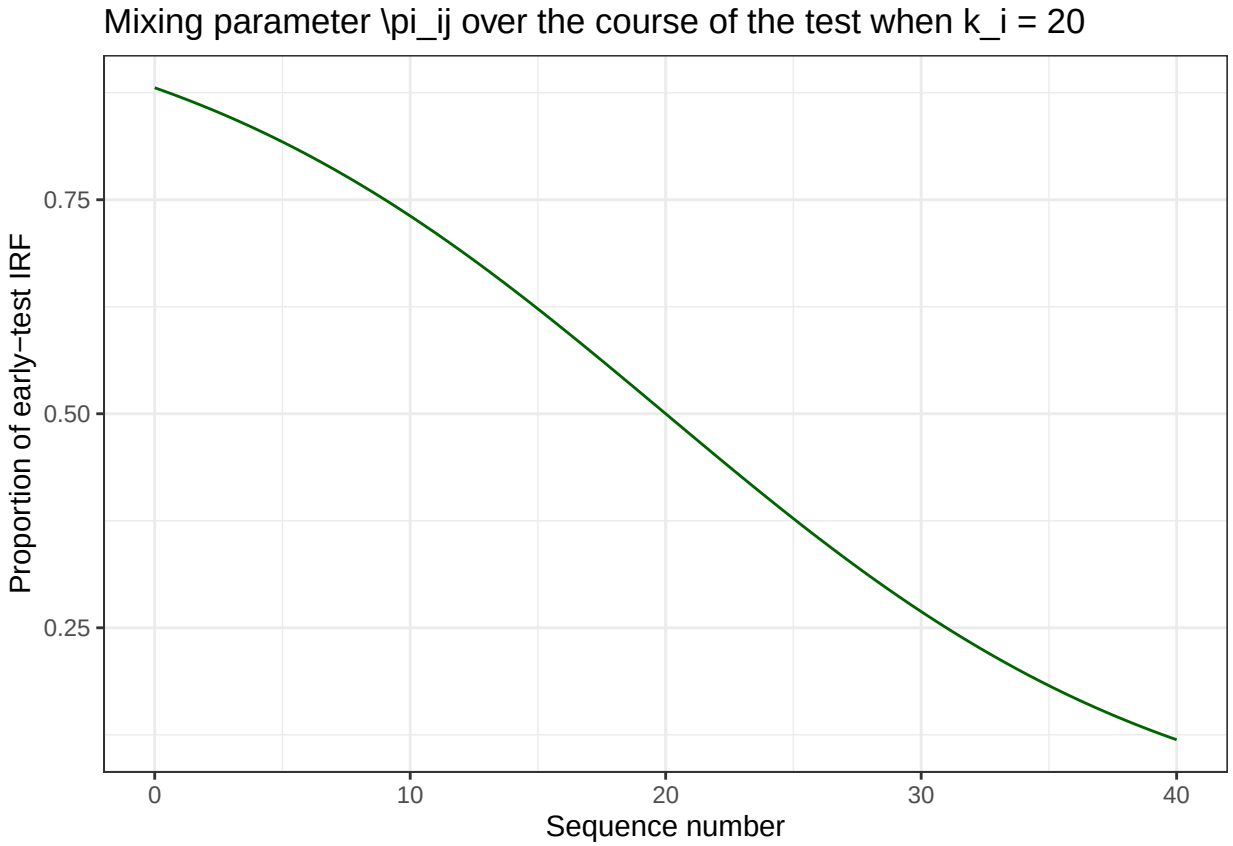
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- π_{ij} is a mixing parameter for the ith student's exposure to item j. This parameter is specified as a function of sequence number:

$$\pi_{ij} = \sigma\left(\frac{k_i - s_{ij}}{c}\right)$$

where s_{ij} is the *sequence number*, or the position in the test sequence where student i saw item j, k_i is the aforementioned "endurance" parameter — learned parameter that corresponds to the sequence number where student i's item response function is a 50/50 mixture of early and late test behavior, and c is a scale parameter that is fixed across all students. For this analysis, the scale parameter was fixed to $c = 10$.

For a respondent with $k_i = 20$, the plot below shows how their π_{ij} changes over the course of a 40-item test.



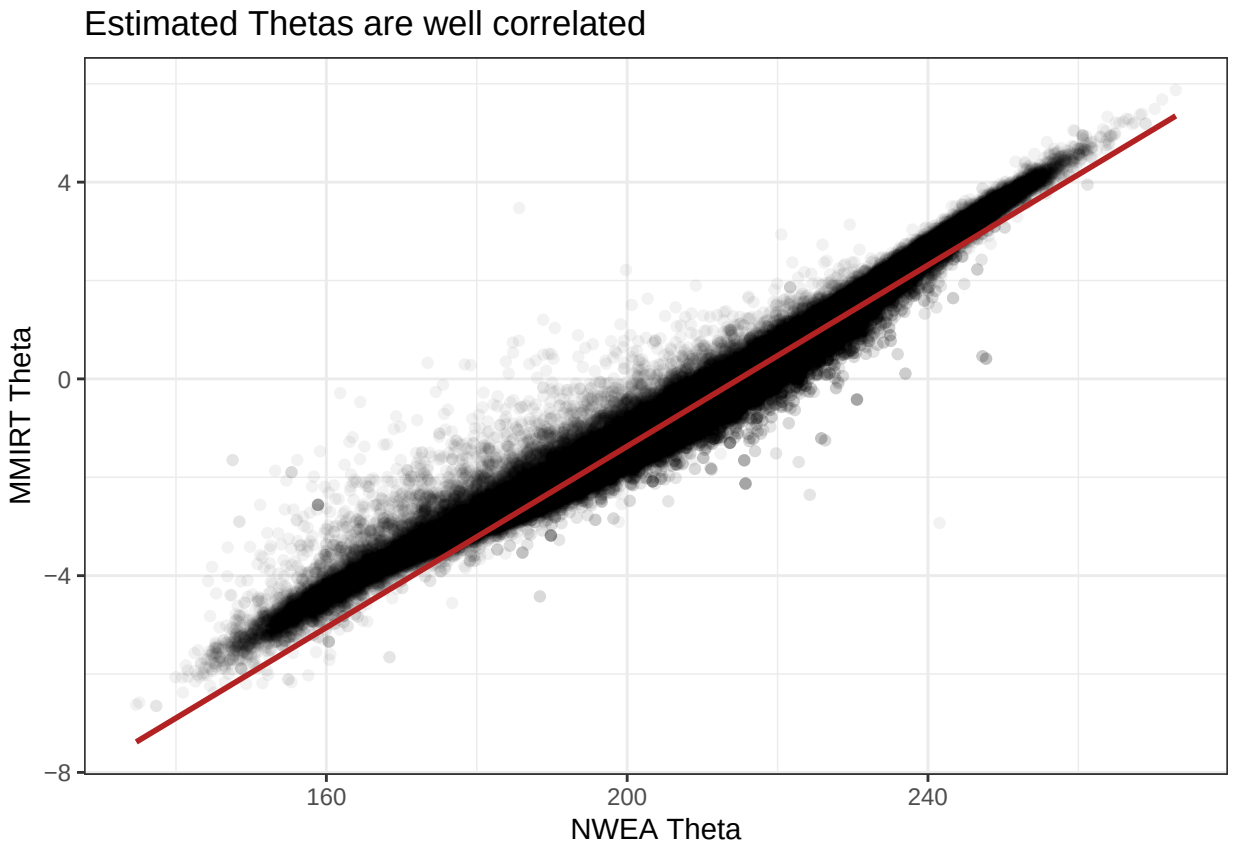
Note that different values of k_i will change where this curve hits 0.5, but the slope is fixed across respondents.

3 Results

3.1 Person-side

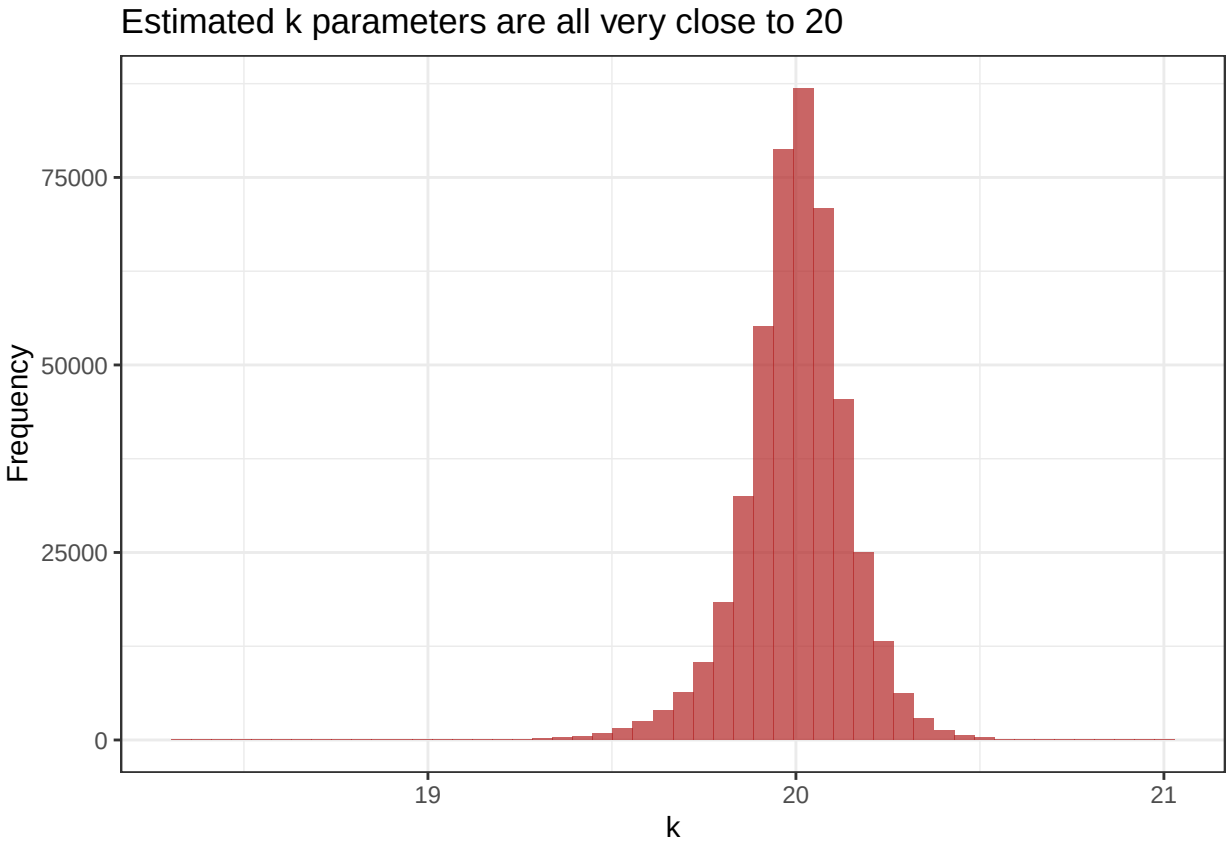
3.1.1 Comparing ability estimates

We see that the θ values estimated from the mixture model are well correlated with the θs estimated by NWEA ($\rho = 0.986$). This is done without using any of the pre-calibrated item information, and estimated from raw item response matrices.



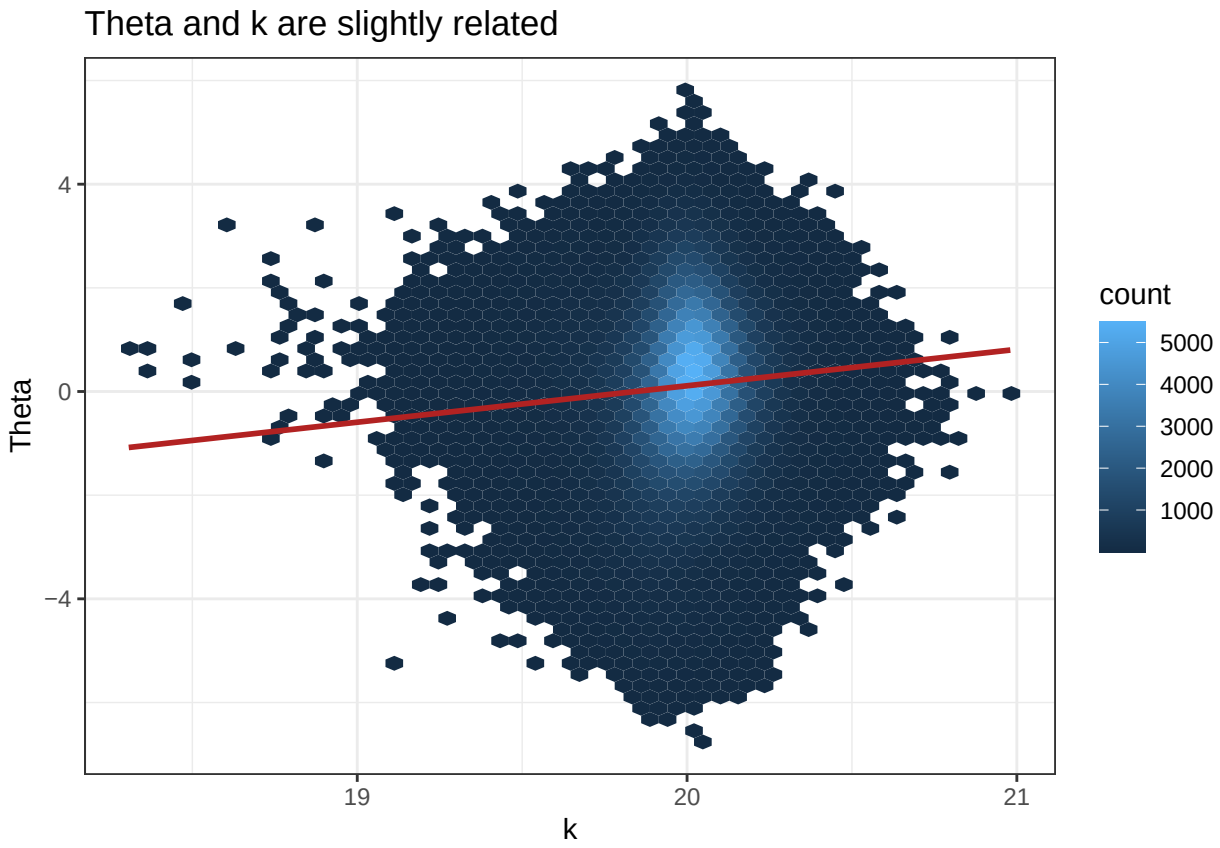
3.1.2 The "endurance" parameter, k

At the start of training, all of the k parameters are initialized to $k = 20$. We see that during training, many students spread away from 20, but not by a wide margin. This is likely due to the use of a fixed scaling constant, $c = 10$, for all of the mixing parameter curves. Future work should explore allowing for a more flexible specification of the mixing parameter, though this would involve estimated three parameters per person instead of only two.



3.1.3 Relationship between k and θ

Intuitively, one might expect that students who maintain early test behavior further into the test will do better. Plotting a hexbin scatter with a linear fit line over it, we see this to be (weakly) the case.



Looking at a linear regression of θ on k, we see that k only explains half a percent of the overall variance in θ. One potential confound is that all grades are pooled together - so high performing younger students may have higher k values and similar θ to lower performing older students. Additionally, with this many observations, any coefficient would likely be significant.

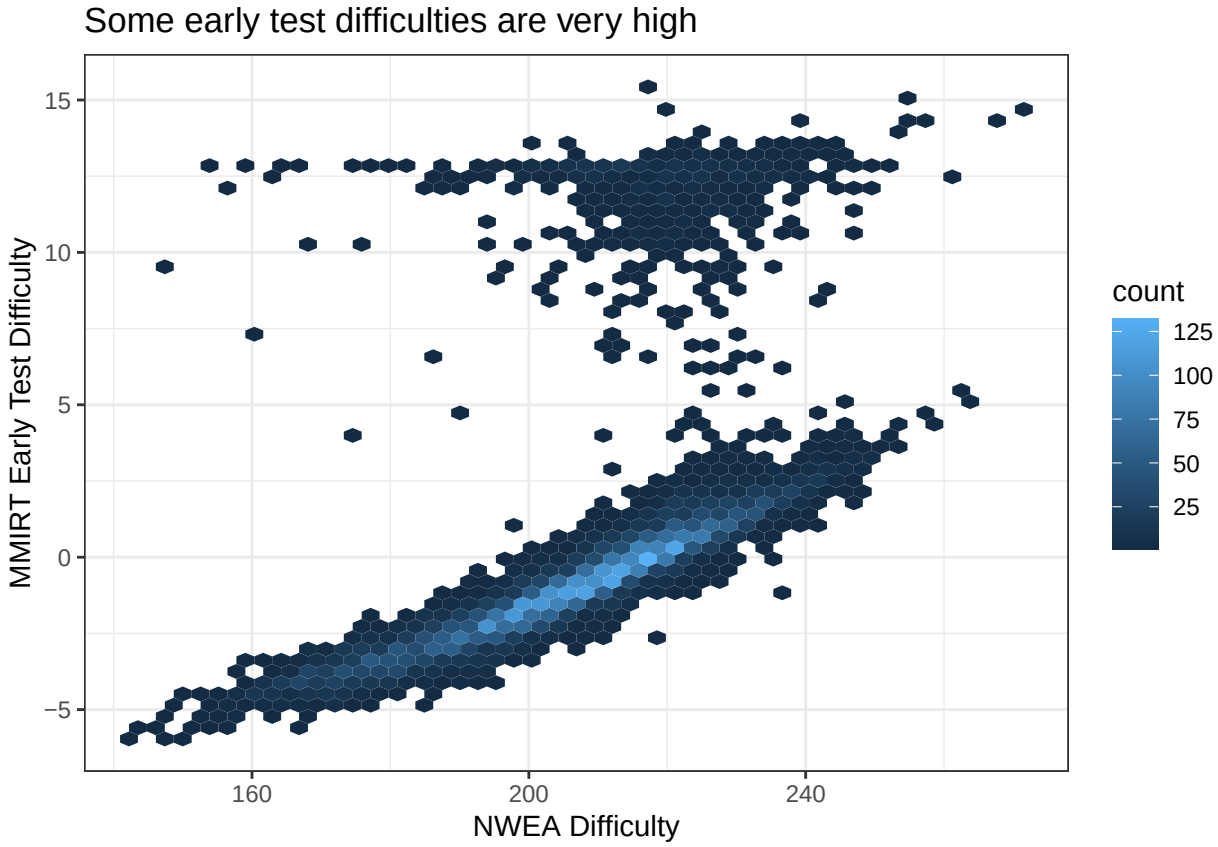
Table 1:

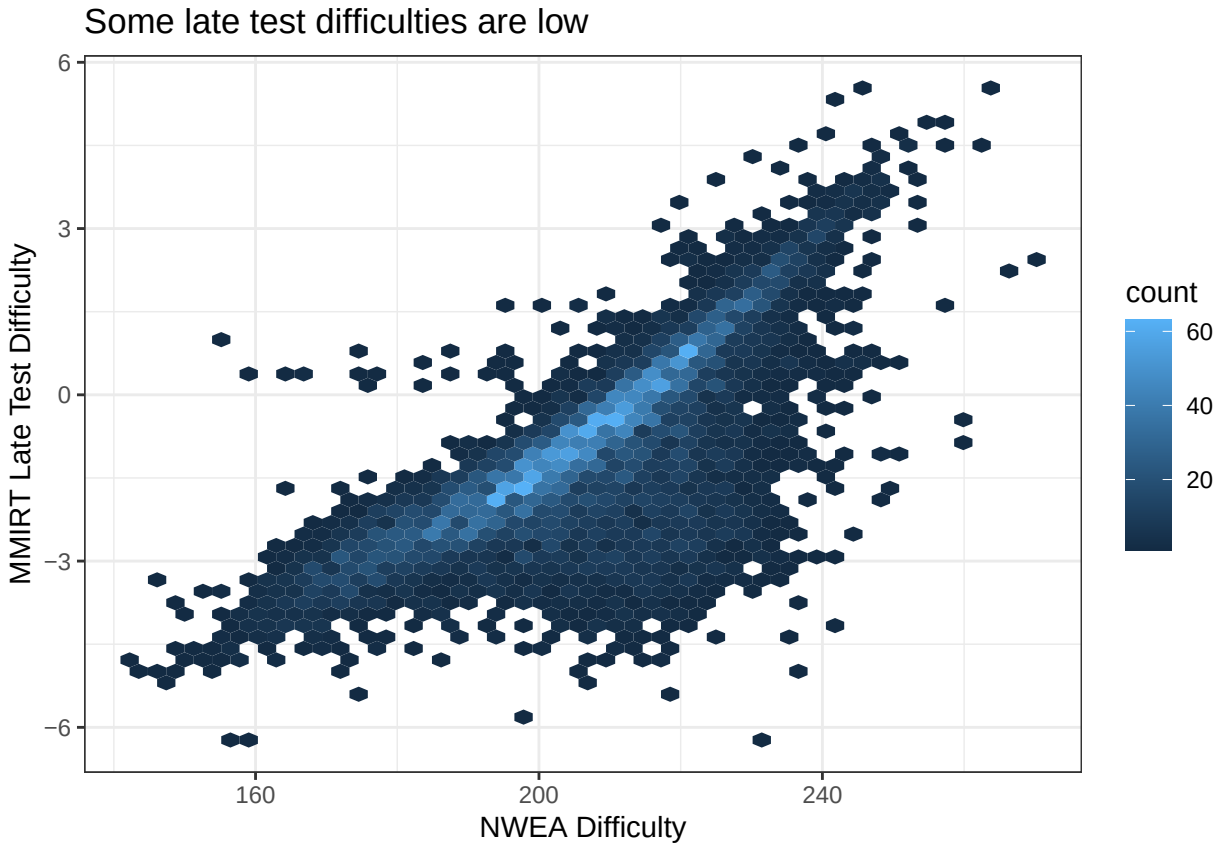
	Dependent variable:
	theta
k	0.703^{***}
	(0.015)
Constant	-13.952^{***}
	(0.295)
Observations	464,117
R^2	0.005
Adjusted R^2	0.005
Residual Std. Error	1.432 (df = 464115)
F Statistic	$2{,}267.438^{***}$ (df = 1; 464115)
Note:	*p<0.1; **p<0.05; ***p<0.01

3.2 Item-side

3.2.1 Comparing difficulties with NWEA

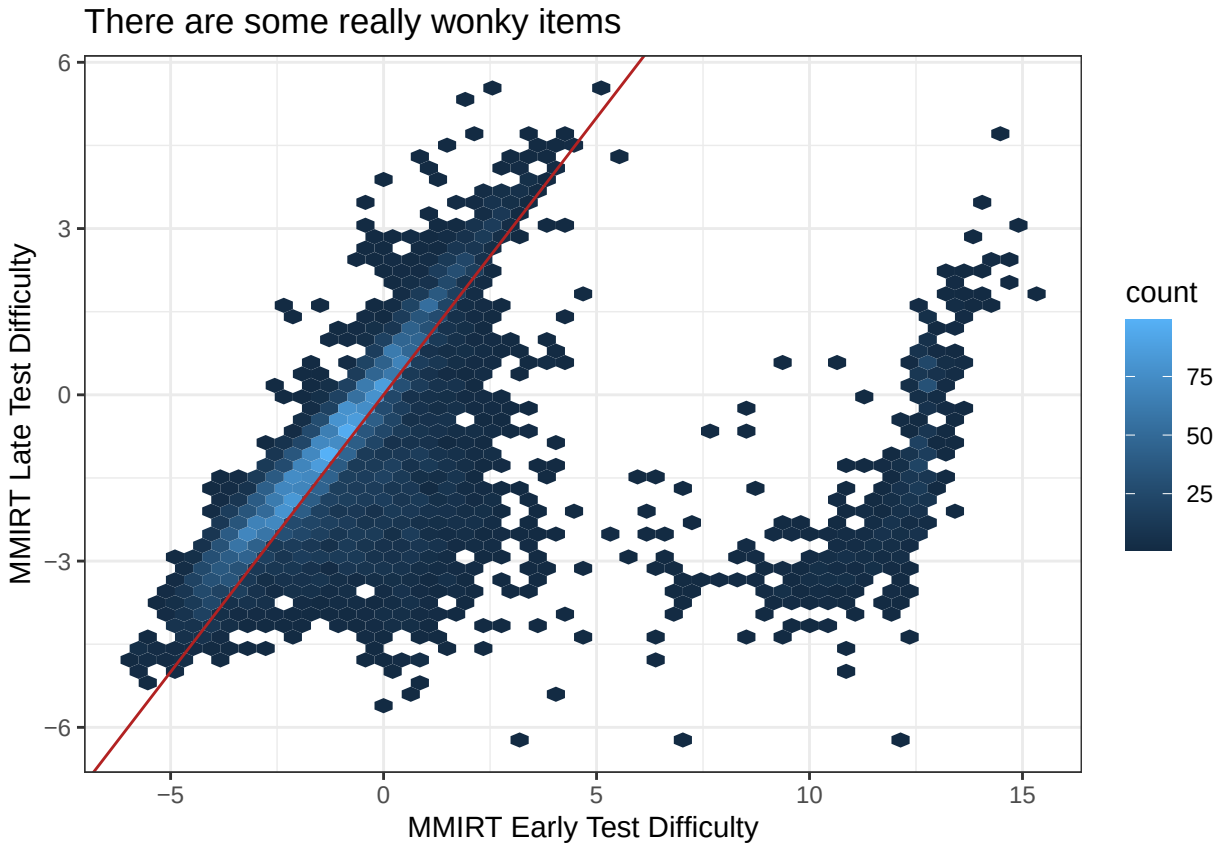
The two plots below compare mixture model estimated early and late test difficulties to the NWEA estimated difficulties. We see that they are generally well correlated in both scenarios, but some mixture model early test difficulties are really high, while there is a chunk of low late test difficulties. We examine the relationship between early and late test difficulty within the mixture model in the next section.



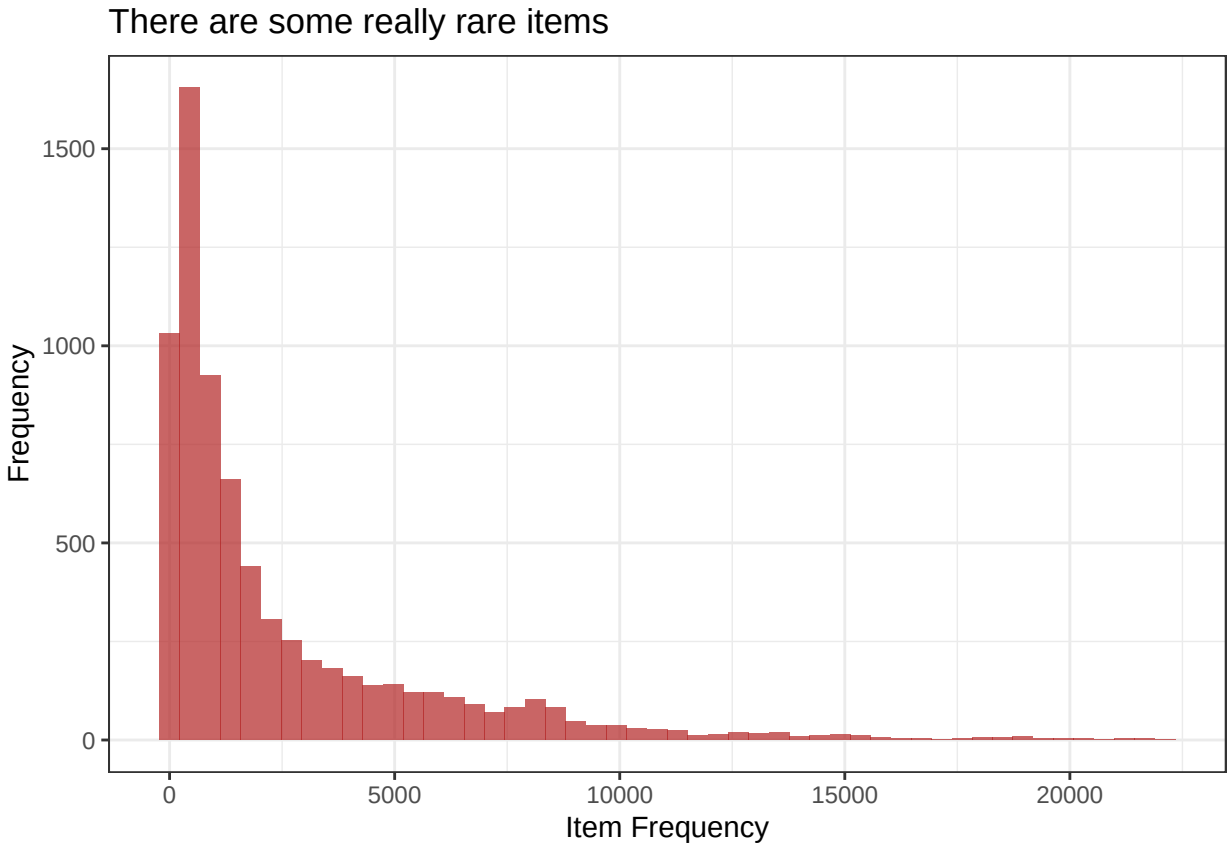


3.2.2 Changing difficulties

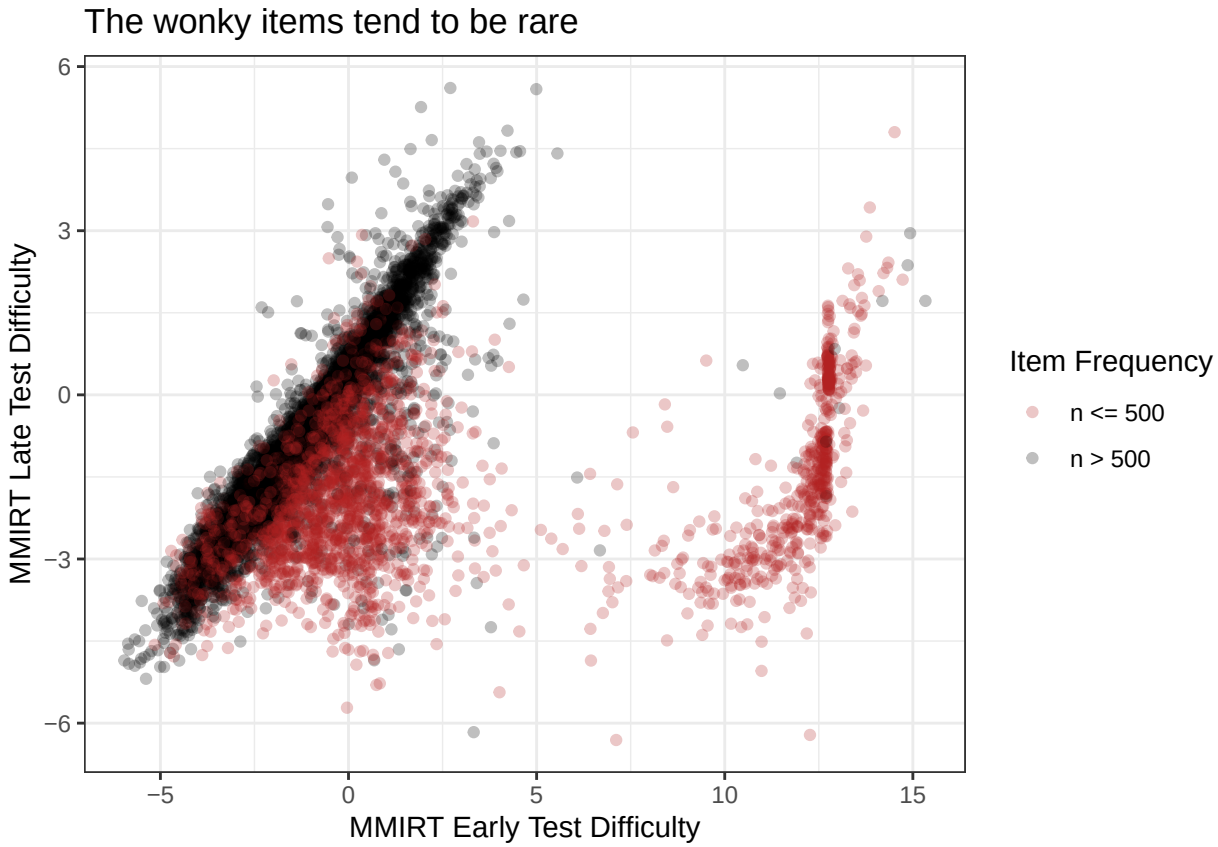
Looking at a hexbin scatter with mixture model early test difficulty on the x axis and late test difficulty on the y axis. A line along $y = x$ has been added to guide the eye. We see that while the bulk of items are, in fact, harder when encountered late in the test, there are two large blobs of items that are estimated to be easier early in the test.



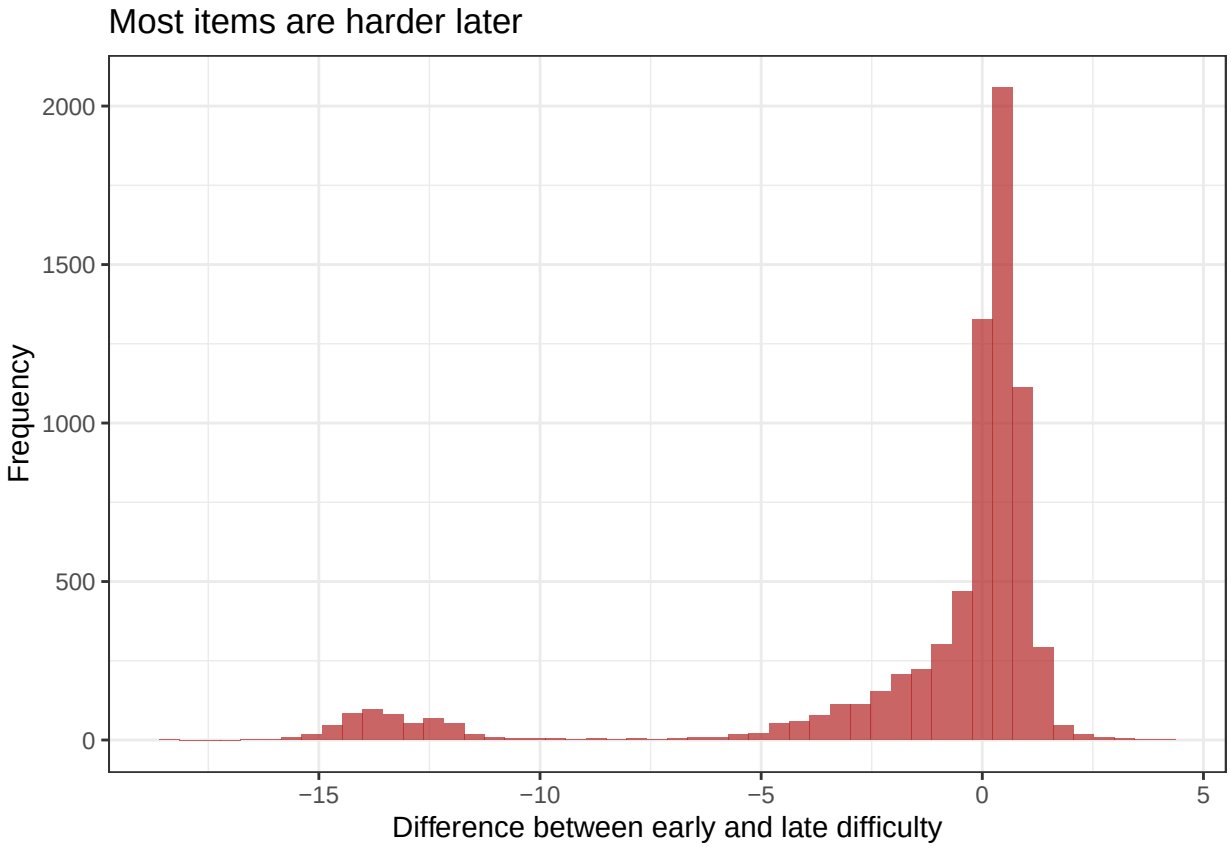
One potential explanation for this could have to do with the relative frequency of individual items.



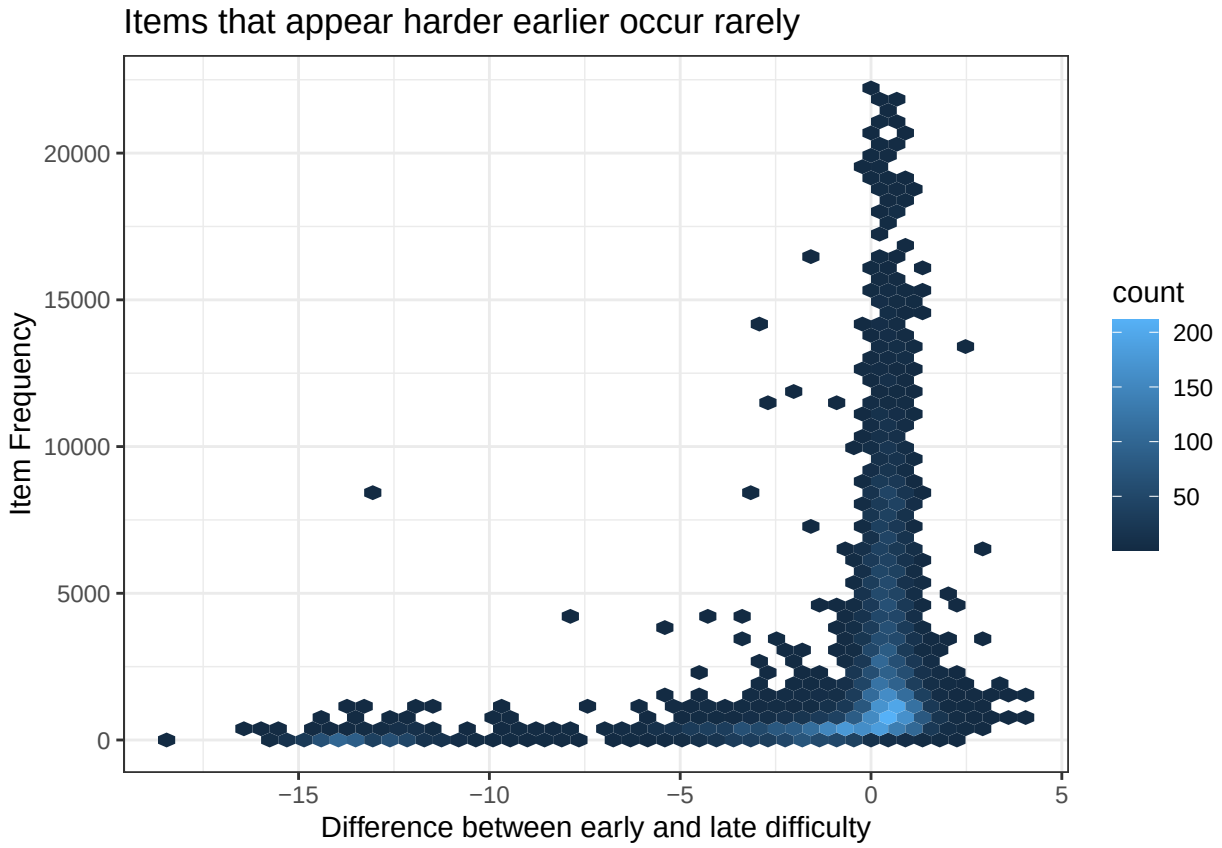
When we color the earlier plot based upon item frequency, we see that items that occur fewer than 500 times account for almost all of the "wonky" items that appear easier in the beginning than they are later.



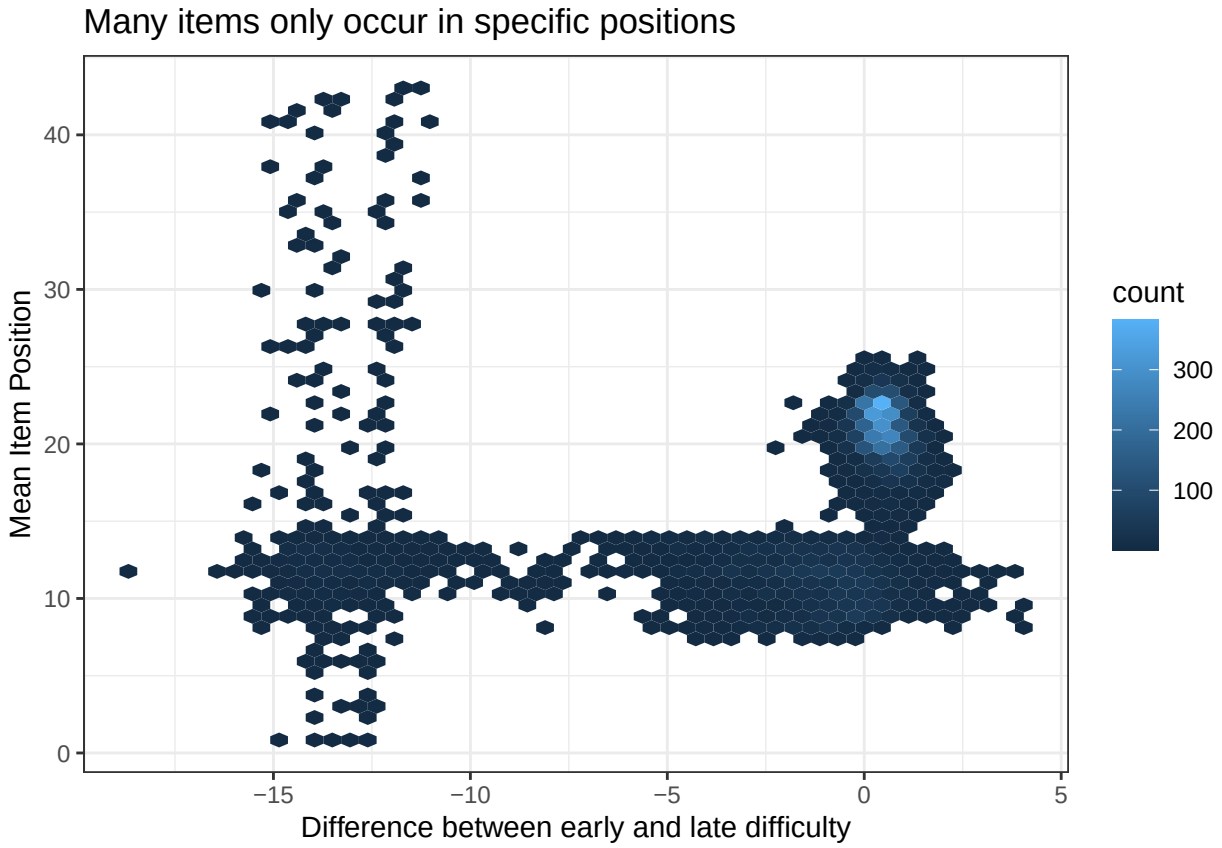
Looking at a histogram of the difference between late test difficulty and early test difficulty, we see that most items do appear harder later in the test.



Additionally, we confirm with a different plot that those items that are harder earlier are, in fact rare. Many of these items appear only a single time in the dataset.

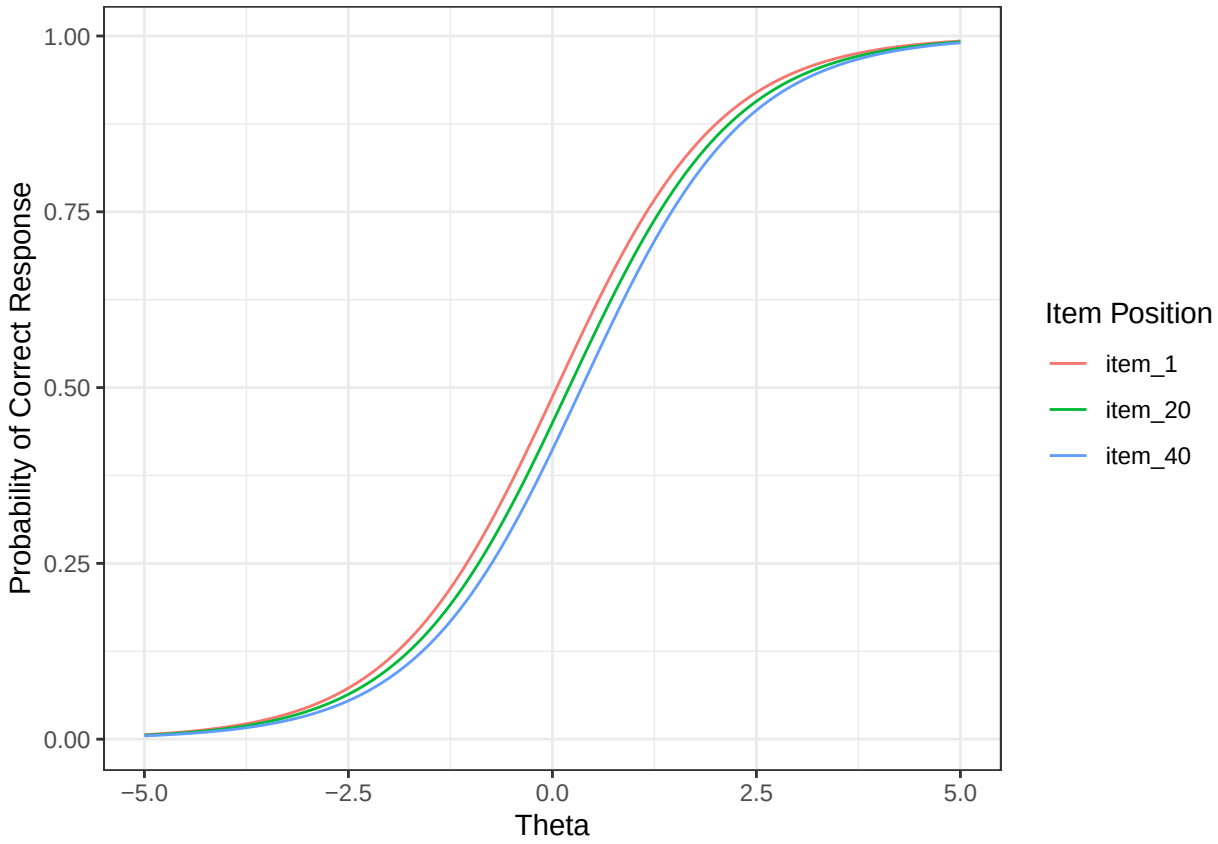


We can also look at the mean sequence position vs the difference between late and early test difficulty. Here we see that many of the items with the largest magnitude differences between early and late test difficulty may only appear at the beginning or end of the test. If all respondents (or, in some cases, they only respondent) answer those items correctly or incorrectly, maximum likelihood estimation will be able to see gains by pushing the item parameters farther and farther apart. This is a potential limitation of the model. For items that tend to appear anywhere and have a large number of responses, parameter estimation seems to reflect intuition.



3.2.3 Effective IRF

Among items occurring $n > 500$ times, the median difference between early and late test difficulty is $\delta = 0$. Next we visualize how the item response function changes as a function of sequence position for a hypothetical item with early test difficulty $b_0 = 0$ and this δ ($b_1 = 0$). We see that for a respondent with $\theta = 0$ and $k = 20$, the probability of correct response when the item is first is $p = 0.487$. When the item is approximately last ($s = 40$), the response probability drops to $p = 0.412$ — a decrease of 0.075. For some items, the magnitude of this position effect is larger, and for others it is smaller.



4 Conclusions

Within the NWEA MAP assessment, using a mixture model efficiently estimates item position effects by including only a single additional parameter per item. Additionally, the general specification is agnostic to the choice of item response function and allows for flexible specification (for example, using a 2PL and estimating a single discrimination per item, but early/late difficulties). This method appears work well assuming that there is a sufficient number of responses to each item and sufficient variability in observed item position.

Additionally, this model allows for estimating person parameters that allow for individuals to progress from "relatively early" interactions to "relatively late" interactions at different rates.

Because the thetas and item difficulties estimated are reasonably well aligned with NWEA's estimates (despite being reported on a wildly different scale), I have confidence that the model hasn't fully gone off the rails and is estimating meaningful parameters. Additionally, because there is evidence the items are subject to position effects of relatively large magnitude, I suspect there may be some (as yet unquantified) problem with the NWEA student ability estimates.

That said, this method does have limitations. First, items need to appear in a wide variety of positions. Second, items that have only a few responses will have outlandish parameter estimates. Third, because of the way that the scaling constant, c is used in the mixing parameter π_{ij}, the person-side "endurance" parameter k is not particularly informative. Future work to improve the model should allow for a more flexible specification of π_{ij}. I see the contribution from this work as specifying a model to estimate position effects in item response datasets where items can occur in multiple positions. The NWEA MAP assessment provides an empirical example the highlight both strengths and weaknesses of the approach.