International Society of the Learning Sciences

An Item Response Theory Model for Multiple Resubmissions

Xiang Liu, Mo Zhang, Chen Li, Hongwen Guo
xliu003@ets.org, mzhang@ets.org, cli@ets.org, hguo@ets.org
Educational Testing Service
Amy J. Ko, Min Li
ajko@uw.edu, minli@uw.edu
University of Washington

Abstract: Item response theory models (IRT) are commonly used to model students' responses to test items. Under IRT, the associations among item responses are modeled by a latent variable. However, most IRT models are not suitable for modeling multiple resubmissions which are common in a learning environment. Students may be allowed to revise and resubmit multiple times after each submission. In this paper, we develop an IRT model for multiple resubmissions. The utility of this model is demonstrated through analyzing a real data set from a recently developed novel authentic assessment of programming skills

Introduction

Item response theory (IRT; Baker & Kim, 2004; Hambleton & Swaminathan, 1985) is widely used to model students' responses to assessment items by relating observed item responses to an underlying latent proficiency. Under standard IRT assumptions, each student produces a single response to each item, and conditional on item parameters such as difficulty and discrimination, responses are locally independent and driven by a static latent trait. This framework has proven especially useful in summative assessment contexts, where responses are typically collected once per item and used to rank examinees along a unidimensional proficiency continuum.

In contrast, many contemporary learning environments—particularly computer-based formative assessments—allow students to revise and resubmit their responses multiple times, often with feedback provided after each submission. Such revise-and-resubmit workflows are increasingly common in online homework systems, programming assessments, and virtual learning environments, where the primary goal is to support learning rather than to measure end-state performance alone. In these settings, students' response processes generate rich multiple-attempt data that reflect not only their initial proficiency, but also how their performance evolves across attempts.

A growing body of psychometric research has recognized that traditional single-response IRT models are not well suited for such data. One prominent line of work extends IRT using sequential models, in which multiple attempts are treated as ordered responses within an item. For example, Bergner, Choi, and Castellano (2019) proposed sequential IRT models for constructed-response items with multiple attempts and demonstrated that explicitly modeling attempt sequences leads to better fit and ability estimation, particularly when students may stop attempting an item before reaching a correct response. Related work by Li et al. (2022) examined the performance of sequential 2PL IRT models under unstructured multiple-attempt data from virtual learning environments, showing that such models can recover ability reasonably well, but may exhibit bias when ability growth across attempts is substantial and not adequately captured.

Parallel developments have occurred in the context of multiple-choice items administered under answer-until-correct or multiple-attempt designs. For instance, Lu, Fowler, and Cheng (2025) developed a family of sequential IRT models that account for guessing and partial knowledge in multiple-choice, multiple-attempt settings, demonstrating that early incorrect attempts contain valuable psychometric information. Together, these studies provide strong evidence that multiple attempts are not psychometric noise, but rather an important source of information that should be modeled explicitly.

Despite these advances, most existing multiple-attempt IRT models do not directly parameterize students' propensity to improve across resubmissions. In sequential models, improvement is typically implicit in the attempt structure, while person parameters are still interpreted primarily as static proficiency. However, in authentic learning tasks—such as programming problems that require iterative debugging—students may differ not only in their initial proficiency, but also in how effectively they use feedback and revise their solutions. Treating these two aspects as indistinguishable may obscure meaningful individual differences in learning processes.

In this paper, we propose an extended IRT model for multiple resubmissions that explicitly separates these two components. In addition to a latent proficiency parameter, the model introduces a second person-level parameter that captures each student's propensity for improvement across successive attempßßßßts. The proposed

model is designed for constructed-response tasks with revise-and-resubmit behavior and constrains improvement to be non-decreasing across attempts. We demonstrate the utility of the model using data from a computer-based Python programming assessment, showing how it provides a richer characterization of student performance than traditional IRT approaches.

A Python Programming Assessment

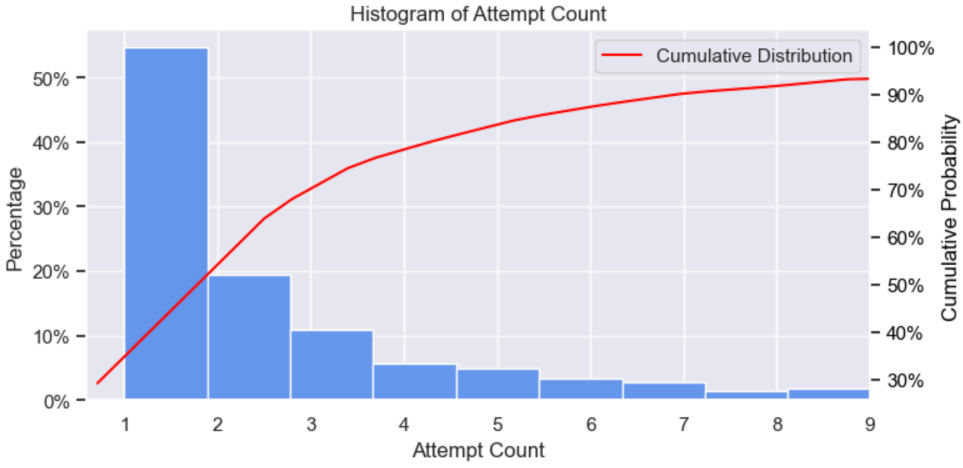
A test was designed to evaluate students' proficiency in basic Python programming. Participants were required to solve 14 to 15 programming tasks by submitting Python-coded solutions. These programs were executed against multiple sets of test cases. If a solution failed any test case, students received feedback indicating which cases failed and were allowed to revise and resubmit their code as many times as needed. The digital testing platform not only recorded the final responses but also logged detailed process data, including all keystrokes, program runs, window focus changes, and copy-paste actions.

The study involved 180 college students enrolled in an introductory Python course. All participants initially completed a set of 9 medium-difficulty items. Based on their performance, they were then directed to either an easier block of 6 items or a more challenging block of 6 items. Each solution was scored as correct (1 point) only if it passed all test cases; otherwise, it received 0 points. Many students opted to revise and resubmit their solutions until they achieved a correct result. The test had no time limit and allowed unlimited resubmissions. Nonetheless, most correct submissions occurred within the first five attempts (see Figure 1). As expected, the likelihood of a correct solution generally increased with more attempts. The data confirmed this trend, showing that students' total scores improved with additional revisions and resubmissions (see Figure 2).

There are some intricacies when modeling this type of item response. In contrast to the traditional single submission item response, multiple resubmissions add an extra dimension - the number of attempts. Consequently, it is no longer sufficient to just consider the propensity of getting items correct when measuring the students' performance, but we also need to take into account how many attempts a student makes before getting that correct response.

Figure 1

Total number of attempts until a correct response



An Extended IRT Model for Resubmissions

Let $X_{ij} = x_{ij}$ denote the item response from the ith student to the jth item, for $i = 1, 2, \ldots, N$ students and $j = 1,2, \ldots, K$ items. If the response is correct, $x_{ij} = 1$; otherwise, $x_{ij} = 0$. Under a two-parameter logistic IRT model, the probability of a correct response is given by

$$P(X_{ij} = 1|\theta_i) = \frac{\exp\left(a_j(\theta_i - b_j)\right)}{1 + \exp\left(a_j(\theta_i - b_j)\right)}, \quad (1)$$

where θ_i describes the proficiency level of the *i*th student, and a_j and b_j are the item discrimination and the difficulty parameters for the *j*th item. Sometimes, especially under small sample sizes, the item discrimination parameters a_j are constrained to be the same for all $j = 1,2,\dots,K$, i.e. $a_j = a, \forall j \in \{1,2,\dots,K\}$. Then the IRT model becomes

$$P(X_{ij} = 1|\theta_i) = \frac{\exp\left(a(\theta_i - b_j)\right)}{1 + \exp\left(a(\theta_i - b_j)\right)}, \quad (2)$$

To accommodate item responses with multiple resubmissions, we extend the IRT model. Let $X_{ijt} = x_{ijt}$ denote the *t*th submission from the *i*th student to the *j*th item. $x_{ijt} = 1$ if the submission is correct and $x_{ijt} = 0$ if it is incorrect. In addition to the latent proficiency parameter θ_i, We introduce another person parameter γ_i which describes the propensity of the *i*th student getting improved probability of a correct response during multiple resubmissions. Furthermore, we constrain the parameter to be nonnegative, i.e. $\gamma_i \geq 0$, so that a student's correct response probability is non-decreasing in multiple resubmissions. Under the model, the probability of a correct response is now given by

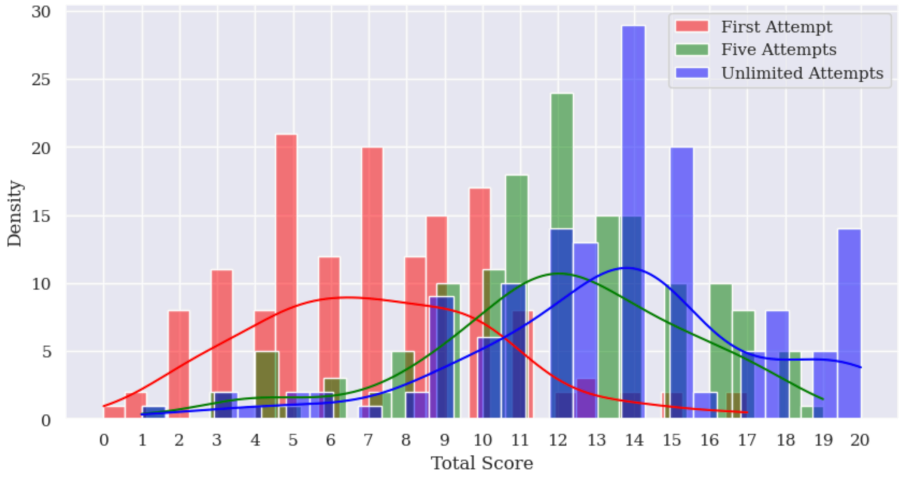
$$P(X_{ijt} = 1|\theta_i, \gamma_i) = \frac{\exp\left(a(\theta_i - b_j + (t-1)\gamma_i)\right)}{1 + \exp\left(a(\theta_i - b_j + (t-1)\gamma_i)\right)}, \quad (3)$$

Where the person proficiency parameter, θ_i, is assumed to follow a standard normal distribution $\theta_i \sim N(0,1)$.

Under this extended IRT model, a student is described by a pair of latent variables – (θ_i, γ_i). For two students who may be located differently in this two-dimensional space, if $\theta_i \geq \theta_j$, the *i*th student is more proficient than the *j*th student. As a result, the more proficient student is more likely to get items

Figure 2

Total score by number of attempts



correct on the first submission. But the *j*th student may be more likely to improve during subsequent resubmissions, if $\gamma_i < \gamma_j$. Notice that the improvement is in the logit scale, not the probability scale.

Real Data Analysis

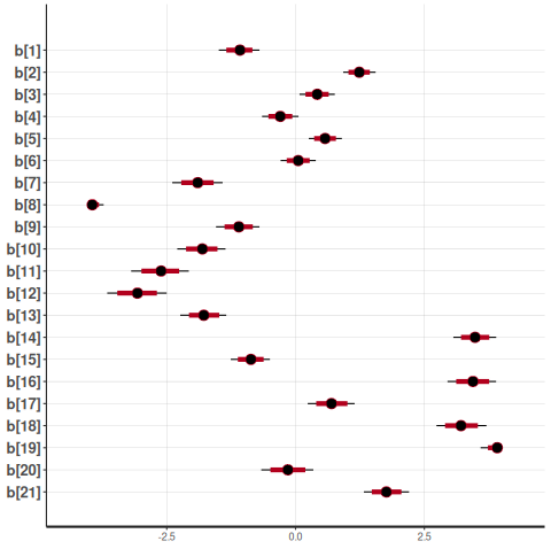
To demonstrate the utilities of our proposed approach, we fit the model to the Python programming assessment dataset. The model parameters are estimated using a Bayesian approach. The Bayesian estimation is carried out

using STAN (Gelman et al., 2015) - a general-purpose Bayesian inference software. Interested readers may find the code in the appendix.

We used 5000 iterations in the Markov chain Monte Carlo (MCMC; e.g. Liu and Johnson, 2019) algorithm, of which half were treated as burn-ins and the rest were used for summarizing posteriors. Figure 3 shows the estimated item difficulty parameters. The difficulty levels of the items are clearly differentiated. The item difficulties of the easier items are estimated to be around –2.5, those of medium level items are around 0, and the more difficult items have item difficulties above 2.5.

Figure 3

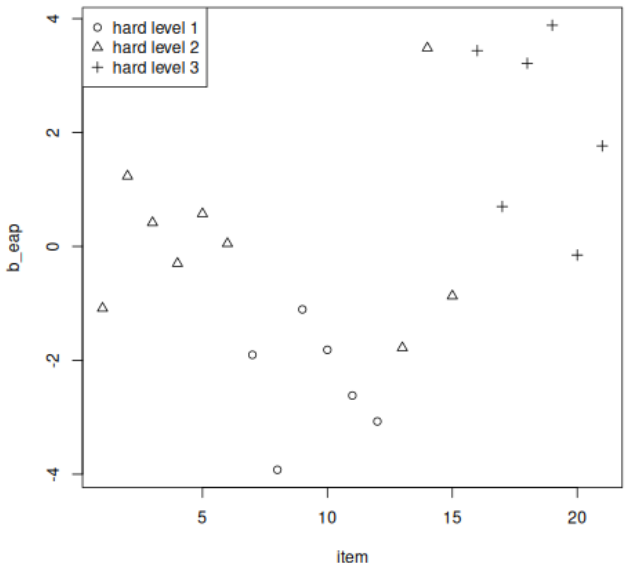
Posterior credible intervals of item difficulty parameters



In Figure 4, the estimated item difficulty parameters are compared against the expert judged difficulty levels during the item development. The results of fitting the data to the extended IRT model empirically validated the expert judgement.

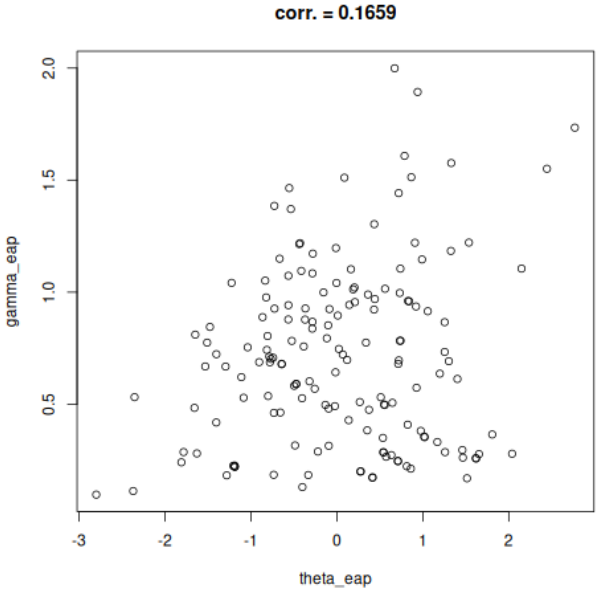
Figure 4

Estimated item difficulties vs. Expert judgement



In a traditional IRT model, the main utility is to rank order students along a single dimensional latent continuum. However, the extended IRT model offers some important additional utilities, such as profiling students on a two-dimensional space which may help gauge students' proficiency levels together with their tendencies to improve over multiple resubmissions. Figure 5 shows that, even though the overall correlation between programming proficiency and the propensity of improvement is not particularly high at 0.1659, there are some recognizable patterns for the extremes. The lower proficient groups tend to have lower propensity for improvement during resubmissions, while the most proficient groups tend to have higher propensity for improvement.

Figure 5
2D profile of proficiency and propensity for improvement



Discussion

In this paper, we proposed an extended IRT model for modeling multiple resubmissions. The model is demonstrated in the context of analyzing a computer programming assessment dataset. However, the introduced model can be used in other types of classroom assessments where multiple resubmissions are common. The results from fitting the model can help practitioners refine items and provide opportunities for delivering individualized interventions.

While the proposed model offers a useful framework for capturing students' improvement across multiple resubmissions, it simplifies the underlying dynamics by assuming a linear improvement propensity. Students' revise-and-resubmit behaviors can be far more complex. For instance, some students may exhibit rapid improvement early on and plateau later, while others may show sporadic or delayed gains. Factors such as motivation, feedback quality, cognitive load, and prior knowledge can all influence the trajectory of improvement in non-linear ways.

The current model assumes that the probability of a correct response increases linearly with each additional attempt, governed by a single parameter γ_i. This assumption, while computationally convenient, may be too restrictive for capturing the nuanced patterns observed in authentic learning environments. Future research could explore more flexible formulations, such as non-linear growth functions, piecewise models, or hierarchical structures that allow for varying improvement rates across different items or student subgroups.

By incorporating richer behavioral dynamics, extended models could offer deeper insights into learning processes and better support adaptive instructional strategies. Such developments would enhance the interpretability and applicability of IRT-based approaches in formative assessment contexts.

References

Baker, F.B., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques. CRC Press.

Bergner, Y., Choi, I., & Castellano, K. E. (2019). Item response models for multiple attempts with incomplete data. Journal of Educational Measurement, 56(2), 415–436. https://doi.org/10.1111/jedm.12214

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic programming language for Bayesian inference and optimization. Journal of Educational and Behavioral Statistics, 40(5), 530-543.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Kluwer-Nijhoff Publishing.

Li, Z., Huggins-Manley, A. C., Leite, W. L., Miller, M. D., & Wright, E. A. (2022). Assessing ability recovery of the sequential IRT model with unstructured multiple-attempt data. Educational and Psychological Measurement, 82(6), 1203–1224. https://doi.org/10.1177/00131644211058386

Lu, Y., Fowler, J., & Cheng, Y. (2025). A family of sequential item response models for multiple-choice, multiple-attempt test items. Psychometrika. Advance online publication. https://doi.org/10.1017/psy.2024.18

Liu, X., & Johnson, M.S. (2019). Estimating CDMs using MCMC. In M. von Davier & Y.-S. Lee (Eds.), Handbook of Diagnostic Classification Models (pp. 629-646). Springer

Appendix

Code for fitting the extended IRT model in STAN

```
data {
        i n t <l o w e r=1> J ; // # s t u d e n t s
        i n t <l o w e r=1> K; // # i t e m s
        i n t <l o w e r=1> N; // # obs
        i n t <l o w e r=1> M; // # max a t t e m p t s
        i n t <l o w e r =1, upper=J> j j [N ] ; // s t u d e n t f o r obs . n
        i n t <l o w e r =1, upper=K> kk [N ] ; // item f o r obs . n
        i n t <l o w e r =1, upper=M> mm[N ] ; // attempt f o r obs . n
        i n t <l o w e r =0, upper=1> y [N ] ; // obs
}
parameters {
```

```
        real theta [ J ] ;
        real <lower=-4, upper=4> b [K ] ;
        real <lower=0, upper=4> gamma [ J ] ;
        real <lower=0.1 , upper=3> a ;
}
model {
        theta ~ std_normal ( ) ;
        for ( n in 1 :N)
                y [ n ] ~ bernoulli_logit ( a * ( theta [ jj [ n ] ] - b [ kk [ n ] ] + (mm[ n ] - 1 ) *
        gamma [jj[n]]));
}
```