

FedIRT: An R package and shiny app for estimating federated item response theory models

Biying Zhou¹ and Feng Ji¹

¹ Department of Applied Psychology & Human Development, University of Toronto, Toronto, Canada
Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Open Journals](#)

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

We developed an R package FedIRT, to estimate traditional IRT models, including 2PL and the graded response models with additional privacy, allowing parameter estimation in a distributed manner without compromising estimation accuracy. Numerical experiments demonstrate that Federated IRT estimation achieves comparable statistical performance to mainstream IRT packages in R, with the benefits of privacy preservation and minimal communication costs. The R package also includes a user-friendly Shiny app that allows clients (e.g., individual schools) and servers (e.g., school boards) to apply our proposed method in a user-friendly manner.

Statement of Need

IRT ([Embretson & Reise, 2013](#)) is a statistical modeling framework grounded in modern test theory, frequently used in the educational, social, and behavioral sciences to measure latent constructs through multivariate human responses. Traditional IRT estimation mandates the centralization of all individual raw response data in one location, thereby potentially compromising the privacy of the data and participants ([Lemons, 2014](#)).

Federated learning has emerged as a field addressing data privacy issues and techniques for parameter estimation in a decentralized, distributed manner. However, there is currently no package available in psychometrics, especially in the context of IRT, that integrates federated learning with IRT model estimation.

Mainstream IRT packages in R, such as `mirt` ([Chalmers, 2012](#)) and `ltm` ([Rizopoulos, 2007](#)) require storing and computing all data in a single location, which can potentially lead to violations of privacy policies when dealing with highly sensitive data (e.g., high-stakes student assessments).

We have therefore developed a specialized R package, FedIRT, to integrate federated learning with IRT. We have also developed an accompanying Shiny app to recognize real-world challenges and aim to reduce the burden of learning R programming for applying this package. This app implements the method in a user-friendly and accessible manner.

Method

Here we briefly introduce the key idea behind integrating federated learning with IRT. For details, please refer to our methodological discussions on Federated IRT ([Zhou & Ji, 2024+](#), [2023](#)).

Model formulation

The two-parameter logistic (2PL) IRT model is often considered the most popular IRT model. In 2PL, the response by person i for item j is often binary: $X_{ij} \in \{0, 1\}$, and the probability of person i answering item j with discrimination α_j and difficulty β_j correctly:

$$P(X_{ij} = 1|\theta_i) = \frac{e^{\alpha_j(\theta_i - \beta_j)}}{1 + e^{\alpha_j(\theta_i - \beta_j)}}$$

To make our package available for polytomous response, we also developed a federated learning estimation algorithm for the Generalized Partial Credit Model (GPCM) in which the probability of a person with the ability θ_i obtaining x scores in item j is:

$$P^{\text{GPCM}}(X_{ij} = x|\theta_i) = \frac{e^{\sum_{h=1}^x \alpha_j(\theta_i - \beta_{jh})}}{\sum_{c=0}^m e^{\sum_{h=1}^c \alpha_j(\theta_i - \beta_{jh})}}$$

In this function, β_{jh} is the difficulty of scoring level h for item j , and for each item j , all difficulty levels have the same discrimination α_j . m_j is the maximum score of item j .

Model estimation

In both 2PL and GPCM, often we assume the ability follows a standard normal distribution, thus we can apply MMLE.

We use a combination of traditional MMLE with federated average (FedAvg) and federated stochastic gradient descent (FedSGD) (McMahan et al., 2017). In our case, the log-likelihood and partial gradients are sent from the clients to the server. Then, the server uses FedSGD to update the item parameters and send them back to clients. By iterations, the model converges and displays the estimates on the interface.

Taking the 2PL model as an example, which has a marginal log-likelihood function l for each school k that can be approximated using Gaussian-Hermite quadrature with q (by default, $q = 21$) equally-spaced levels, and let $V(n)$ to be the ability value of level n , and $A(n)$ is the weight of level n .

$$l_k \approx \sum_{i=1}^{N_k} \sum_{j=1}^J X_{ijk} \times \log\left[\sum_{n=1}^q P_j(V(n))A(n)\right] + (1 - X_{ijk}) \times \log\left[\sum_{n=1}^q Q_j(V(n))A(n)\right]$$

By applying FedAvg, the server collects the log-likelihood values from all k schools and then sums up all the likelihood values to get the overall log-likelihood value: $l = \sum_{k=1}^K l_k$.

The server collects a log-likelihood value l_k and all derivatives $\frac{l_k}{\partial \alpha_j}$ and $\frac{l_k}{\partial \beta_j}$ from all clients, then observe that $\frac{\partial l}{\partial \alpha_j} = \sum_{k=1}^K \frac{l_k}{\partial \alpha_j}$ and $\frac{\partial l}{\partial \beta_j} = \sum_{k=1}^K \frac{l_k}{\partial \beta_j}$ by FedSGD, the server sum up all log-likelihood values and derivative values.

Also, we implemented a Federated Median method, which uses the median of the likelihood values to replace the sum of likelihood values in Fed-MLE. It is more robust when there are outliers in input data.

With estimates of α_j and β_j in 2PL or β_{jh} in GPCM, empirical Bayesian estimates of students' ability can be obtained (Bock & Aitkin, 1981).

68 Comparison with existing packages

69 We showcase that our package could generate the same result as traditional IRT packages,
70 for example, mirt (Chalmers, 2012). Take 2PL as an example, we use a synthesized dataset
71 with 160 students and 10 items. %For traditional packages, the whole dataset is used. For our
72 package, the dataset was separated into two parts, which contain 81 and 79 students.

73 Figure 1 and Figure 2 show the comparison of the discrimination and difficulty parameters
74 between mirt and FedIRT based on example_data_2PL in our package.

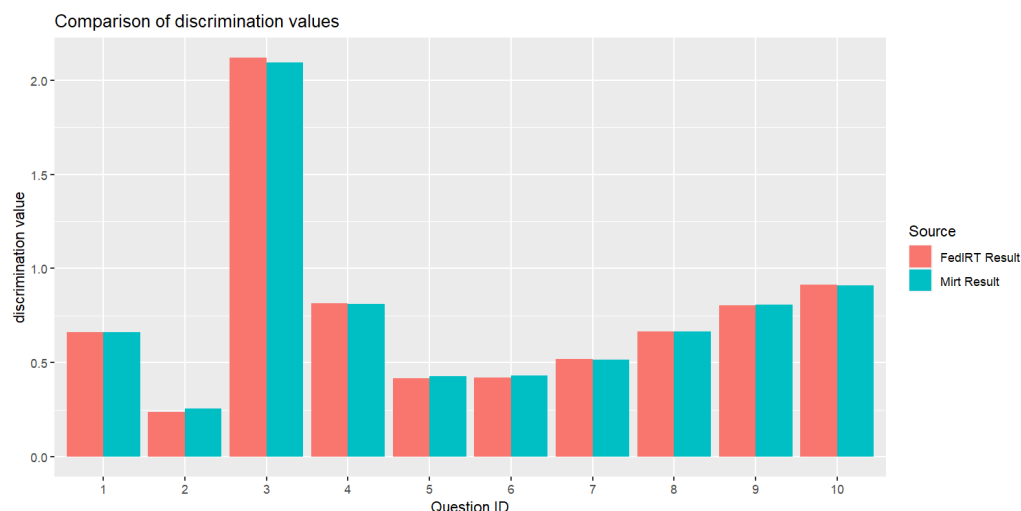


Figure 1: Discrimination comparison

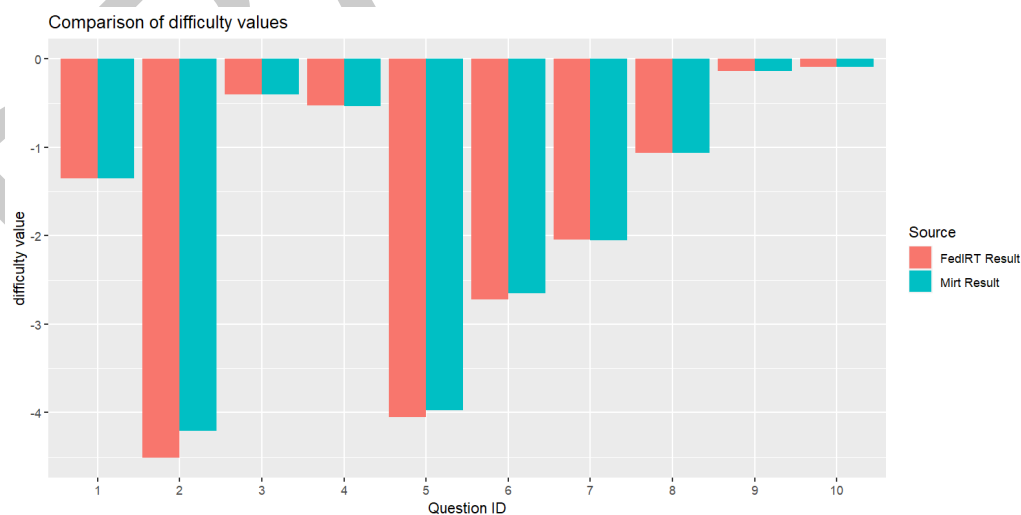


Figure 2: Difficulty comparison

75 Availability

76 The R package FedIRT is publicly available on [Github](#). It could be installed and run by using
77 the following commands:

```
devtools::install_github("Feng-Ji-Lab/FedIRT")
library(FedIRT)
```

78 To provide wider access for practitioners, we include the Shiny user interface in our package.
79 A detailed manual was provided in the package. Taking the 2PL as an example, we illustrate
80 how to use the Shiny app below.

81 In the first step, the server end (e.g., test administrator, school board) can be launched by running
82 the Shiny app (`runserver()`) with the interface shown below:

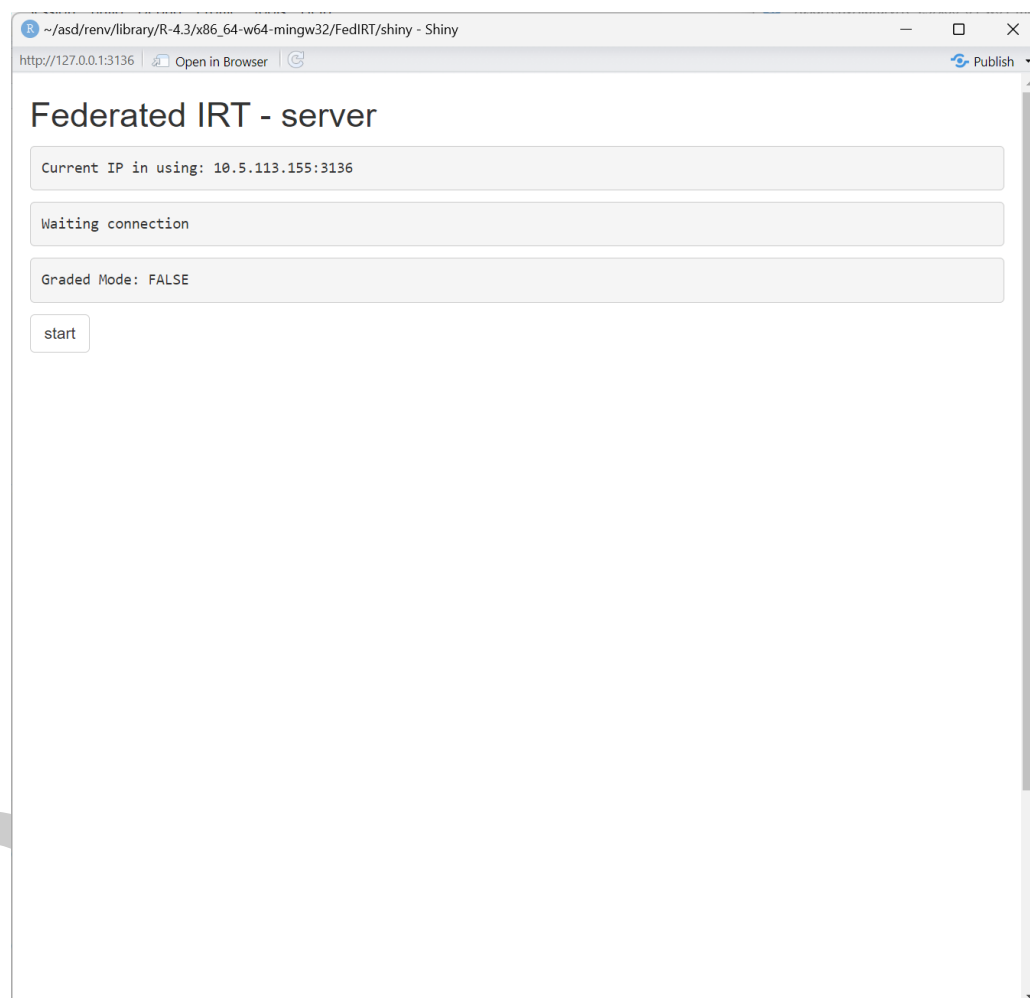


Figure 3: The initial server interface.

83 Then, the client-end Shiny app can be initialized (`runclient()`).

84 When the client first launches, it will automatically connect to the localhost port 8000 as
85 default.

86 If the server is deployed on another computer, type the server's IP address and port (which
87 will be displayed on the server's interface), then click "reconnect". The screenshots of the user
88 interface are shown below.

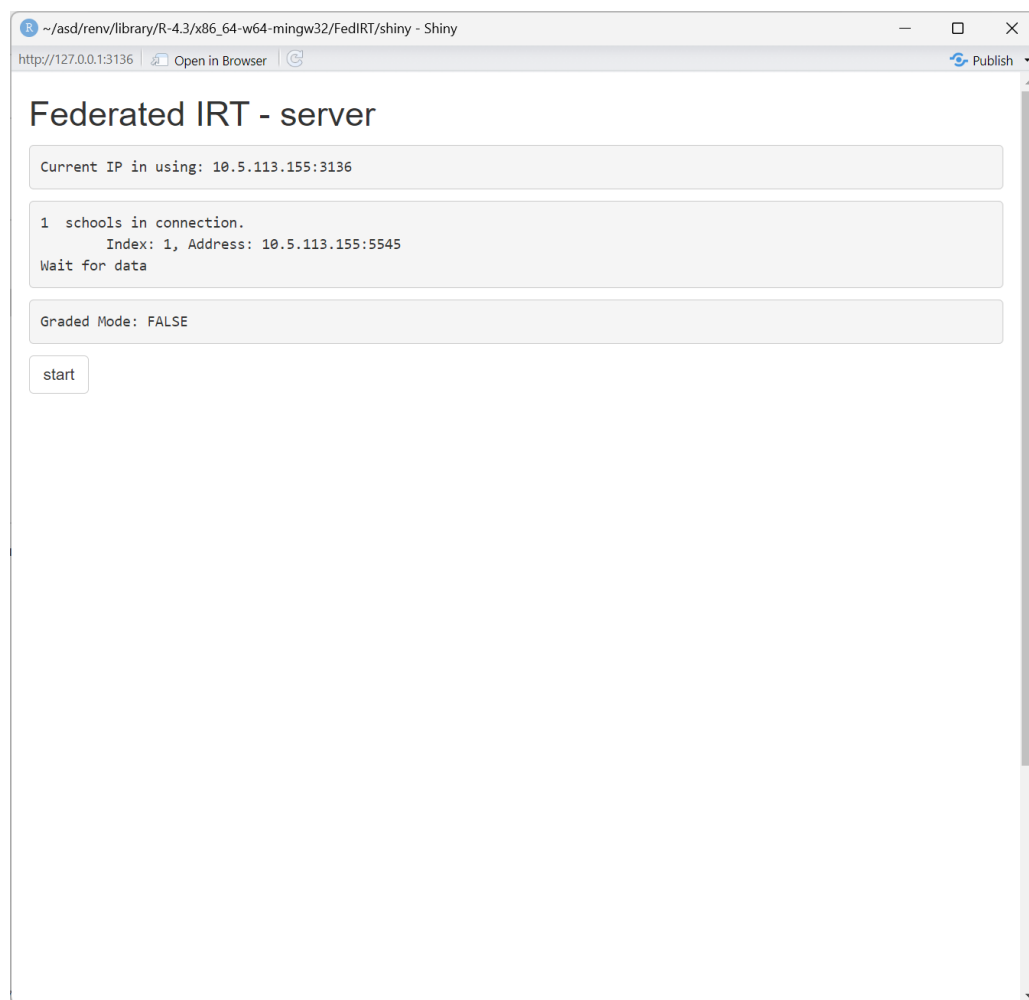


Figure 4: Server interface when 1 school is in connection.

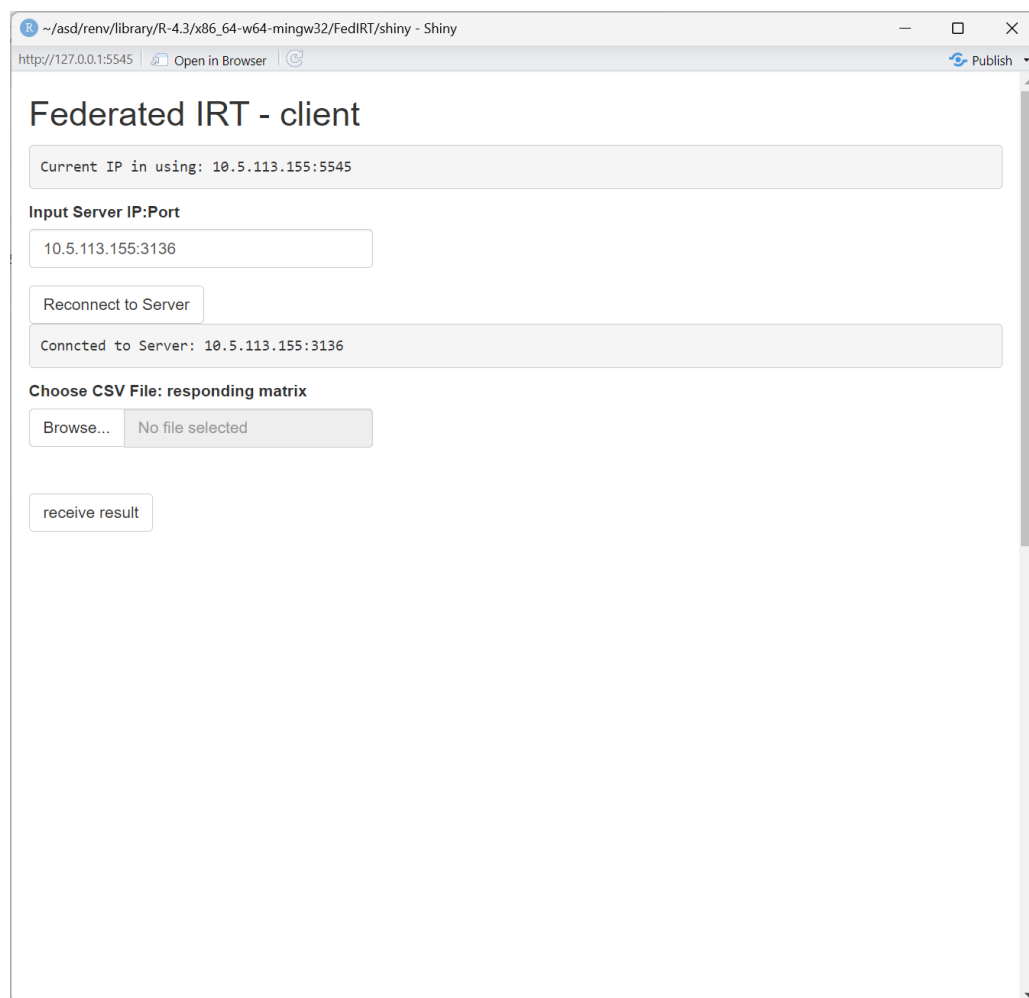


Figure 5: Client interface when connected to server.

89 Then, the client should choose a file to upload to the local Shiny app to do local calculations,
 90 without sending it to the server. The file should be a .csv file, with either binary or graded
 91 response, and all clients should share the same number of items, and the same maximum
 92 score in each item (if the answers are polytomous), otherwise, there will be an error message
 93 suggesting to check the datasets of all clients.

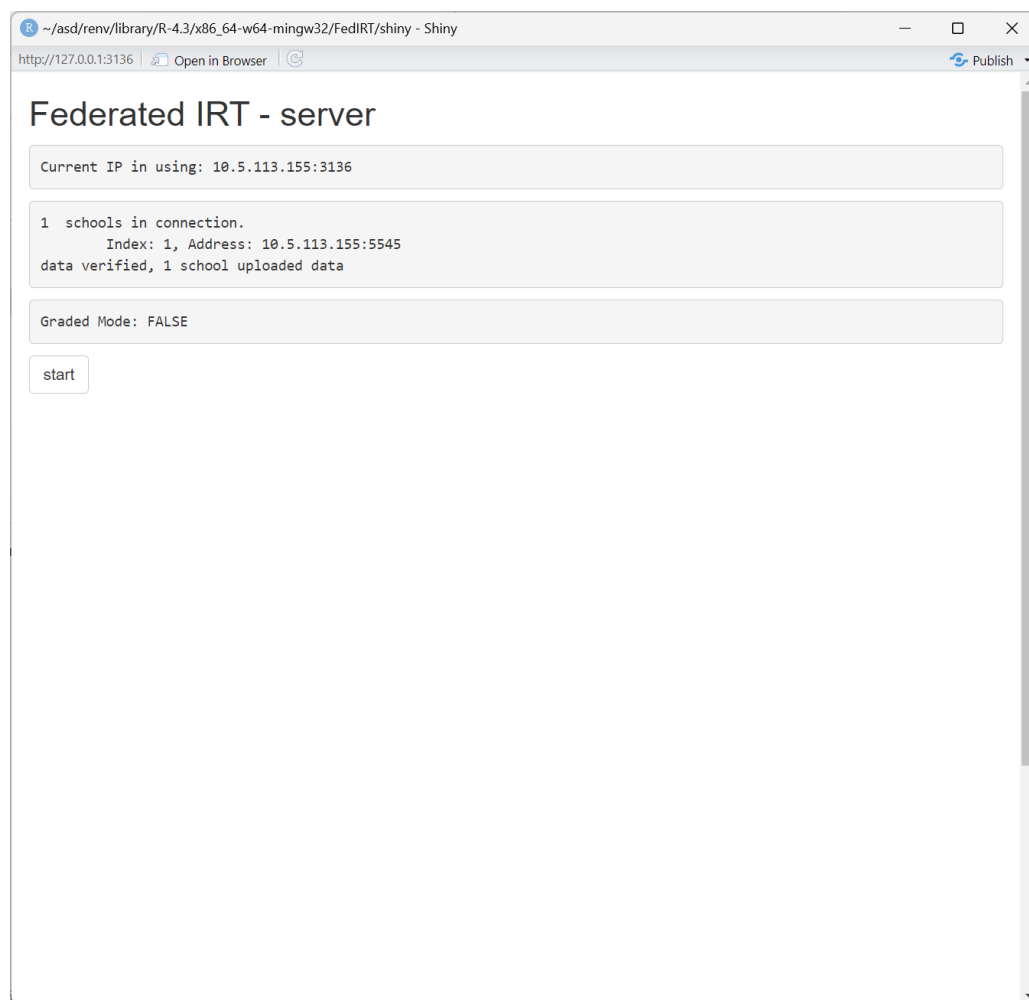


Figure 6: Server interface when 1 school uploaded dataset.

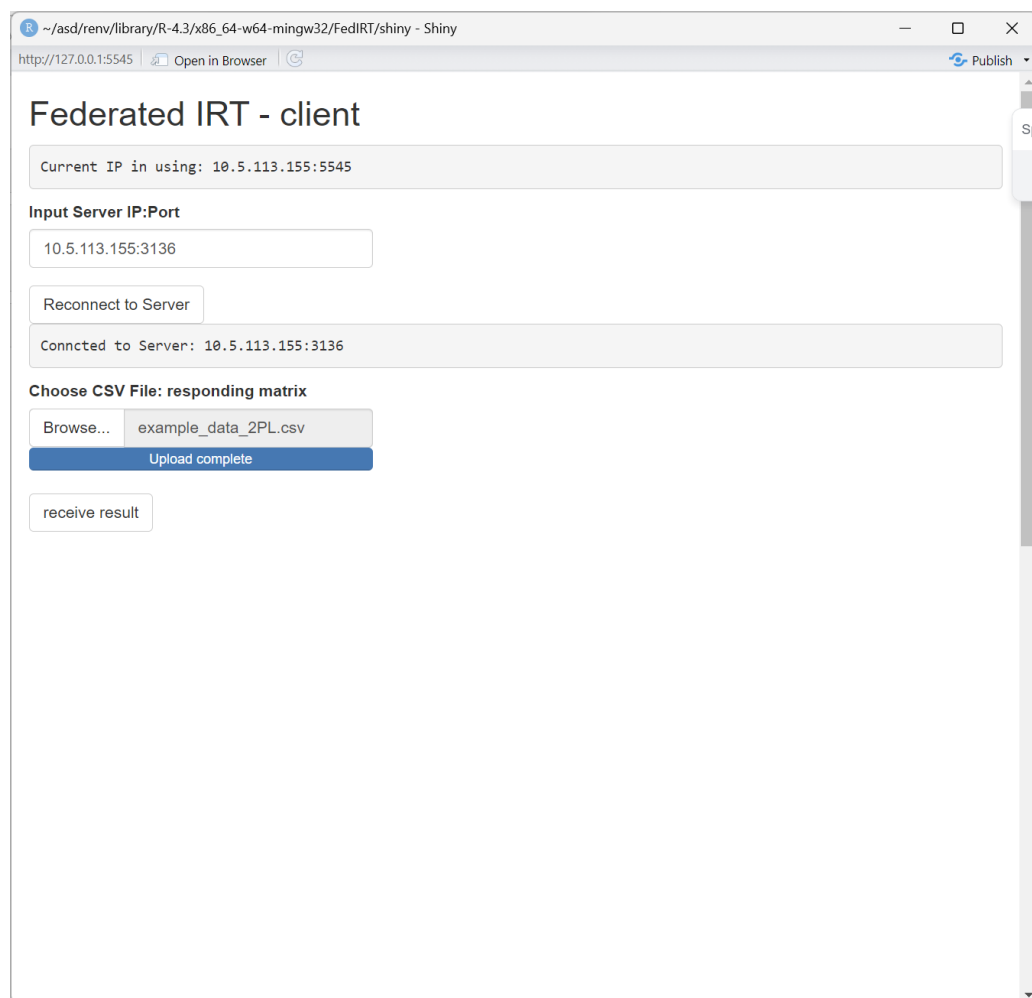


Figure 7: Client interface when a dataset is chosen without errors.

94 After all the clients upload their data, the server should click “start” to begin the federated
 95 estimates process and after the model converges, the client should click “receive result”. The
 96 server will display all item parameters and the client will display all item parameters and
 97 individual ability estimates.

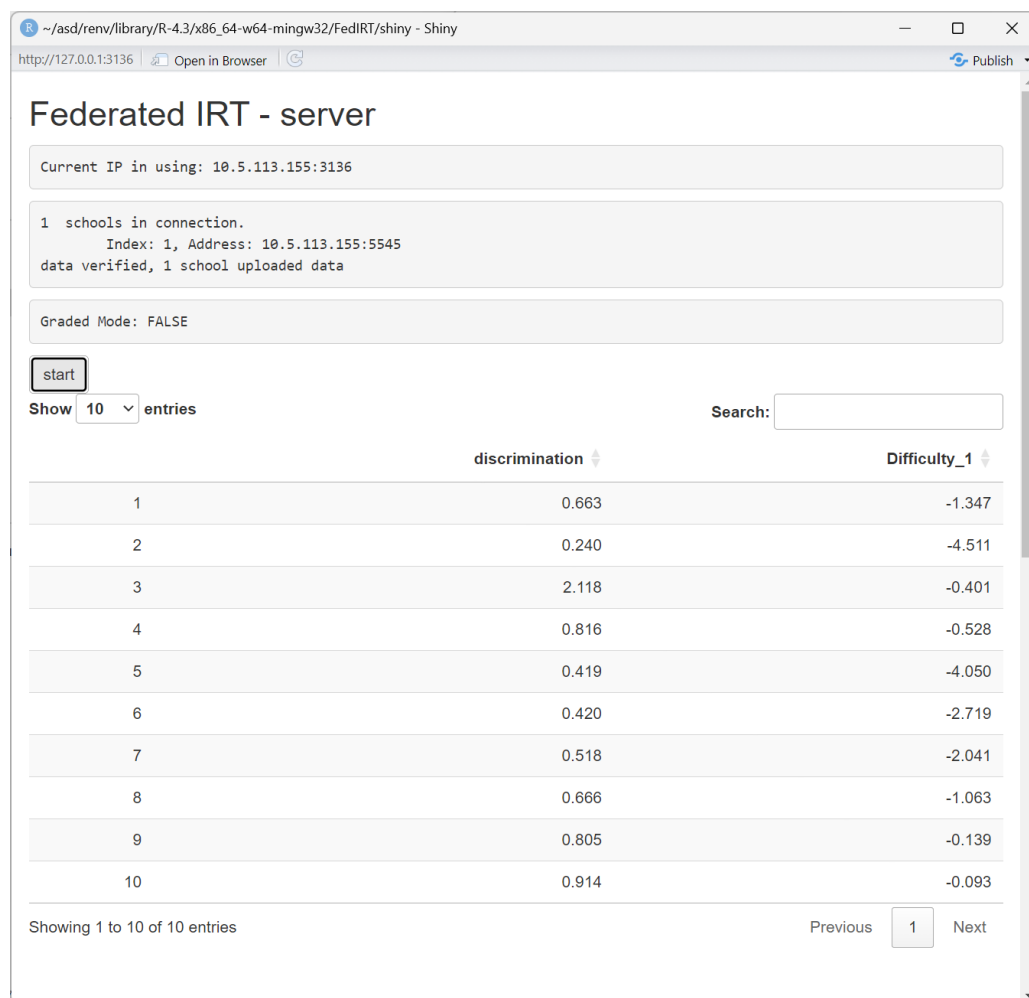


Figure 8: Server interface when finished iteration.

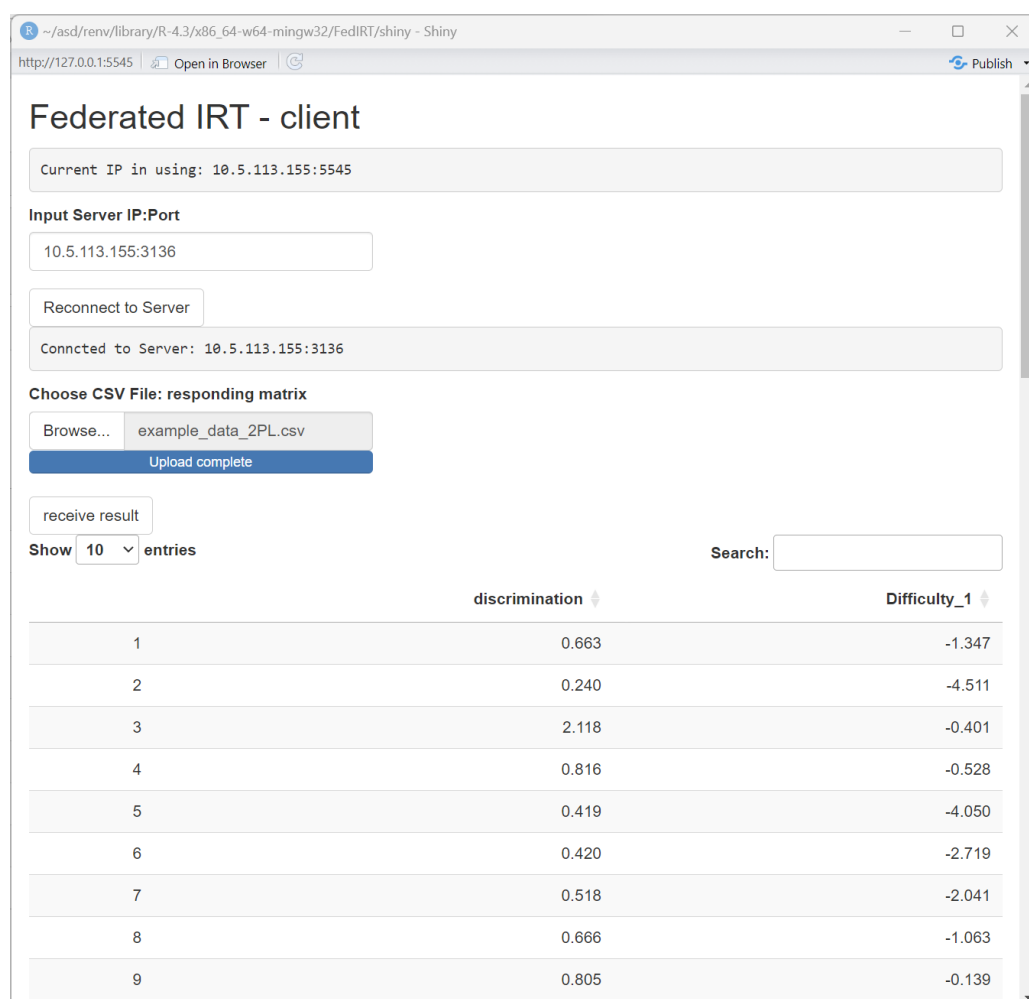


Figure 9: Client interface when result received.

The clients will also display bar plots of the ability estimates.

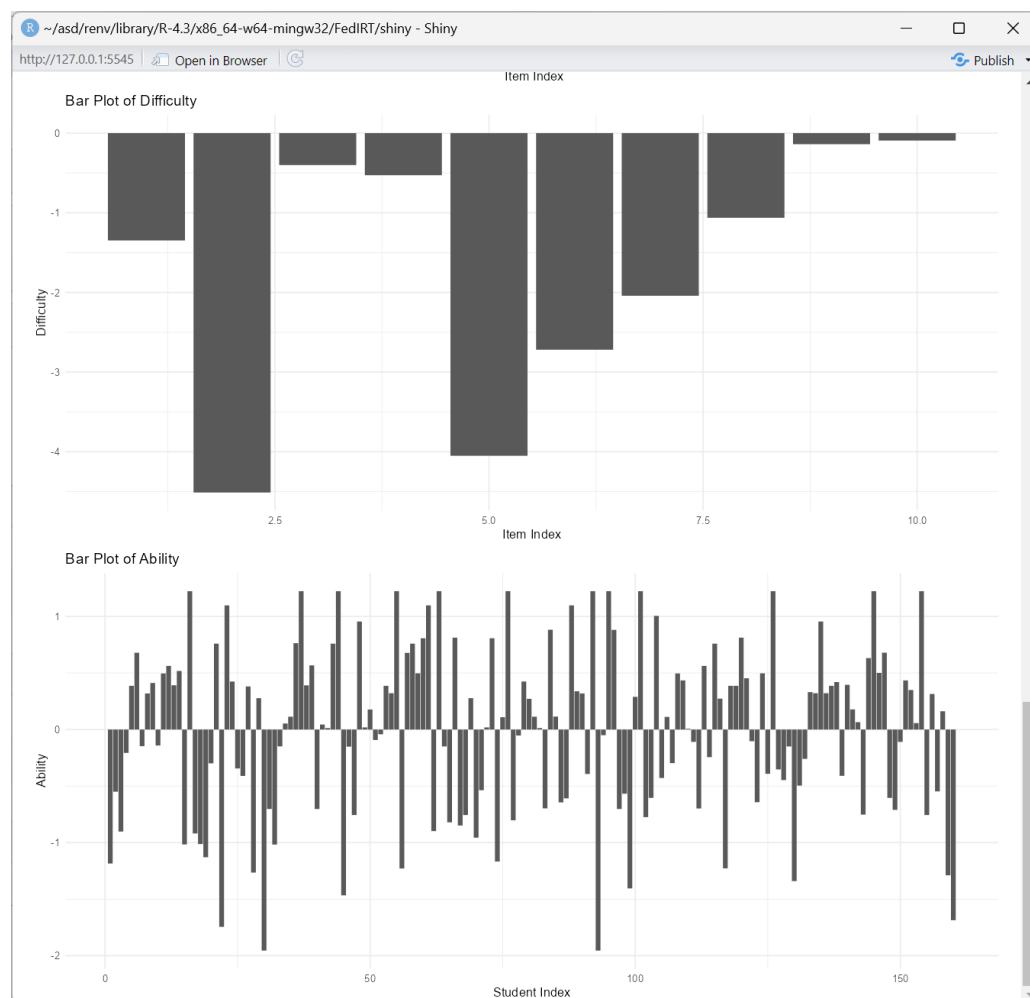


Figure 10: Client interface to display bar plots of discrimination, difficulty and individual ability.

References

- 99 Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters:
100 Application of an EM algorithm. *Psychometrika*, 46(4), 443–459.
- 101 Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r
102 environment. *Journal of Statistical Software*, 48, 1–29.
- 103 Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- 104 Lemons, M. Q. (2014). *Predictive modeling of uniform differential item functioning preservation
105 likelihoods after applying disclosure avoidance techniques to protect privacy* [PhD thesis].
106 Virginia Polytechnic Institute; State University.
- 107 McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-
108 efficient learning of deep networks from decentralized data. *Artificial Intelligence and
109 Statistics*, 1273–1282.
- 110 Rizopoulos, D. (2007). Ltm: An r package for latent variable modeling and item response
111 analysis. *Journal of Statistical Software*, 17, 1–25.
- 112 Zhou, B., & Ji, F. (2024+). Federated item response models. In *Psychometrika (In submission)*.
- 113

- ¹¹⁴ Zhou, B., & Ji, F. (2023). Federated psychometrics: A distributed, privacy-preserving, and
¹¹⁵ efficient IRT estimation algorithm. *APHD Research Gala*.

DRAFT