Impact of Ability Range Restriction on Item Characteristics in Multistage Adaptive Testing

Kyoungwon Lee Bishop[1], Hacer Karamese[1], Xin (Grace) Li[1], Yoon Ah Song[2]

[1]WIDA at the University of Wisconsin-Madison

[2]Center for Applied Linguistics

April 2023

Author Note

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL. Please do not cite without permission from the corresponding author.

Correspondence concerning this manuscript should be addressed to Kyoungwon Lee Bishop at kyoungwon.bishop@wisc.edu

Abstract

The purpose of this study was to compare the performance of two routing methods, maximum Fisher information (MFI) and Access, and four field test (FT) sampling methods in the context of Access Online. Simulations were performed to compare the routing methods and sampling methods. Sample size per FT folder for calibration was also manipulated. The feasibility and efficiency of the manipulated conditions were evaluated in terms of theta and item recovery, and item fit. The results did not provide a consistent pattern with regards routing method. However, FT Condition 3 (i.e., random assignment) consistently outperformed the other FT sampling conditions. Finally, a sample size of 3000 might be sufficient to achieve steady FT parameter estimates and item fit.

Keywords: Multistage adaptive test, field testing, routing, sample size

Impact of Ability Range Restriction on Item Characteristics in Multistage Adaptive Testing

Purpose

This study answers the following research questions in the context of Access Online:

1. How does the MFI routing compare with the current Access routing rule?
2. How does the routing method affect the administration of FT items?
3. How does examinee proficiency distribution affect *b*-parameters?
4. How does calibration sample size affect *b*-parameters?

Method

Routing Method

Two routing methods were considered for module selection; the current Access routing rule based on pre-determined θ scores and conditional standard errors of measurement (CSEM) [1] and MFI. The routing rule might impact measurement precision due to the administration of different modules at a given stage. This might result in differences in examinee distributions per folder at a given stage, which may have a practical implication on FT administration and calibration.

Field Test Assignment

This study considered three FT folder assignment conditions. Condition 1 mirrored the current Access FT assignment rule [2], in which the FT folders were assigned based on examinees' estimated tier level at a particular stage. While Tier B examinees were administered one of tier A, B, or C folders, tier A and C examinees will not be given tier C and A folders, respectively. In Condition 2, examinees received FT folders matching their tier estimates. In contrast, examinees were given FT folders independent from their tier estimates and FT tier under Condition 3 (i.e., random assignment). In other words, examinees had equal chances of getting any of FT folders. Tables 2 and 3 summarize the number of FT folders by tier and stage for listening and reading domains, respectively. The stage column denotes the stage examinees' tiers estimated to administer the corresponding

FT folder under Conditions 1 and 2. The stage information for Entry folders is NA since they are randomly administered to students under all three conditions.

Calibration Sample

Five different sample size conditions were investigated: 500, 1000, 2000, 3000, 4000, and 5000. While calibration samples were randomly drawn from the examinees assigned to a particular folder under the full-restricted and random FT assignment condition, two sample tier ratio conditions were considered under FT Condition 1. Condition 1.1 mirrored the current sample ratio and implemented the pre-determined percentages for sampling. For example, if a Tier A folder is given at Stage 3, the calibration sample should include an equal number of tier A and B students. In contrast, Condition 1.2 considered the observed tier proportions, and the calibration sample mirrored the observed tier proportions rather than the pre-determined proportions.

Calibration

Current Access implements fixed parameter calibration (FPC) procedure for FT via *Winsteps* (Linacre & Wright, 2000). In FPC, the operational parameter (OP) and FT items are combined, and FT item parameters are estimated while all of the OP item parameters are "fixed" during calibration. In this study, the FPC using joint maximum likelihood (JML) was implemented via *TAM* package (Robitzsch et al., 2022). The relevant literature (e.g., Nicklin & Vitta, 2022; Robitzsch et al., 2022) indicated that the *TAM* package performs consistently with *Winsteps*. The authors of this study verified that difficulty and infit estimates from *TAM* package are comparable with *Winsteps*, but *TAM* package produced lower outfit estimates up to .20.

Data

Data considered in this study were from Access Online 503, including Listening and Reading domains. OP item parameters were used to assemble operational multistage adaptive testing (MST) panels. Figures 1 through 10 show test characteristic functions (TCFs) and test information functions (TIFs) for the MST panels used in this study. The

FT items administered during the operational administration were used to simulate data with realistic properties. Table 2 and 3 provide Tier and Stage information for the FT folders.

Final $\hat{\theta}$ from Access Online 503 were treated as true θ scores for response pattern generation to reflect realistic proficiency distributions, which differ across domains and clusters. Figure 11 and Table 1 show proficiency distributions.

Simulation Steps

The *mstR* package (Magis, Yan, & von Davier, 2018) was used for data generation, MST administration, and ability estimation using the aforementioned routing methods. FPC was carried out using the *TAM* package (Robitzsch et al., 2022). The following simulation steps were followed for each condition:

1. Generate 0/1 item responses using the pre-calibrated OP item parameter estimates and the true examinee ability (θ) for the entry folders
2. Compute the examinee's provisional $\hat{\theta}_{MLE}$ using the responses to entry folders
3. Apply the appropriate routing rule (MFI and Access) to route the examinee to a module in the next stage and generate 0/1 item responses for the routed module
4. Compute the examinee's provisional $\hat{\theta}_{MLE}$ using the responses to all administered modules
5. Repeat Steps 3 and 4 for each stage
6. Compute the examinee's final $\hat{\theta}_{MLE}$ using the responses to all administered modules [3]
7. Generate 0/1 item responses using the generating FT item parameters and the true examinee ability (θ) for each item in the FT folders [4, 5]
8. Apply the appropriate FT assignment rule (Conditions 1, 2, and 3) to assign examinees to one of the FT folders
9. Randomly draw samples (500, 1000, 3000, and 5000 per FT folder) from the population for calibration under each FT sampling method (Condition 1.1, 1.2, 2, and 3)

10. Calibrate the FT items using FPC
11. Repeat steps 9 and 10 100 items

Evaluation Criteria

The simulation results were evaluated with respect to the recovery of theta estimates under the different routing conditions and the recovery of b-parameters under the different routing, FT sampling, and sample size conditions.

Theta Estimates

This section describes the criterion considered to address the first two research questions.

Recovery. Standard error (SE), bias, and root mean square error (RMSE) conditioning on θ were calculated to compare the routing methods. They are defined as

$$SE(\theta_k) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[\hat{\theta}_j - \left(\frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j\right)\right]^2}, \tag{1}$$

$$bias(\theta_k) = \left(\frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j\right) - \theta_j, \tag{2}$$

$$RMSE(\theta_k) = \sqrt{bias(\theta_k)^2 + SE(\theta_k)^2}, \tag{3}$$

where θ_k denotes rounded θ to one decimal place for examinee j and N is the number of examinees at θ_k. The rounding approach was used to reduce the jagged pattern on the conditional plots due to many unique θs. The mean standard error (MSE), mean absolute bias (MAbias), and mean average root mean square error (MRMSE) across θ_k were calculated as follows:

$$MSE = \frac{1}{K}\sum_{k=1}^{K} SE(\theta_k), \tag{4}$$

$$MAbias = \frac{1}{K}\sum_{k=1}^{K}\sqrt{bias(\theta_k)^2}, \tag{5}$$

$$MRMSE = \frac{1}{K}\sum_{k=1}^{K}RMSE(\theta_k), \tag{6}$$

where K denotes the total number of rounded true theta points (θ_k) within a cluster and domain.

Module Exposure. Percentage of examinees routed to each tier was computed to understand module exposure rates under each routing method.

Correlation. The Pearson correlation coefficient was computed to measure the strength of the relationship between true and estimated theta scores.

Parameter Estimates

This section describes the criterion considered to address the last two research questions.

Recovery. SE, bias, and RMSE at the item level were calculated to investigate the impact of FT sample distribution on individual items. They are defined as

$$SE(b_i) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{b}_i - \left(\frac{1}{R}\sum_{r=1}^{R}\hat{b}_i\right)\right]^2}, \tag{7}$$

$$bias(b_i) = \left(\frac{1}{R}\sum_{r=1}^{R}\hat{b}_i\right) - b_i, \tag{8}$$

$$RMSE(b_i) = \sqrt{bias(b_i)^2 + SE(b_i)^2}, \tag{9}$$

where b and $\hat{b}$ denote generating and estimated difficulty parameters for item i and R is the number of replication, which is set to 100.

The MSE, MAbias, and MRMSE across FT items within a panel or within a folder

were calculated as follows:

$$MSE = \frac{1}{I}\sum_{i=1}^{I} SE(b_i), \tag{10}$$

$$MAbias = \frac{1}{I}\sum_{i=1}^{I} \sqrt{bias(b_i)^2}, \tag{11}$$

$$MRMSE = \frac{1}{I}\sum_{i=1}^{I} RMSE(b_i), \tag{12}$$

where I is the total number of items within a panel or folder per cluster and domain. MAbias was used to avoid the cancellation of positive and negative values across different items.

Infit and Outfit Statistics. The second criterion compared infit and outfit statistics for estimated item parameters to assess the individual item fit across the studied conditions. As a general rule of thumb, items are flagged as misfitting when infit and outfit values are less than .5 or greater than 1.5 (de Ayala, 2013). The statistics were averaged across replications per each item using the following equation:

$$\hat{\alpha}_i = \frac{1}{R}\sum_{r=1}^{R} \hat{\alpha}_i, \tag{13}$$

where α denotes to estimated infit or outfit statistic for item i and R is the number of replications. Mean infit and outfit statistics across items per cluster and domain were calculated as $\bar{\hat{\alpha}}_i = \frac{1}{I}\sum_{i=1}^{I} \hat{\alpha}_i$.

Correlation. The correlation between item response (0/1) and $\hat{\theta}$ were computed for each FT item to check the item quality. Items with negative correlations are not acceptable. In practice, FT items with a correlation of .2 or below are eliminated.

Results

Theta Estimates

Recovery. Figures 12 and 13 compare the two routing methods with respect to the recovery of θ for listening and reading, respectively. The horizontal dashed lines in bias plots serve as a zero line (i.e., no bias). The zigzag pattern on some plots can be explained by the uneven number of examinees at a given theta point (see Figure 11). As shown in Figure 12 the routing methods perform similarly at the higher and upper ends of the θ scale for listening. While the MFI routing method (blue line) produced smaller SE, bias, and RMSE at some of the θ points under Clusters 23 and 912, the Access routing method resulted in lower SE, bias, and RMSE at some other theta points under Clusters 1, 45, 68, and 912. The routing methods followed a similar trend for the reading domain and produced similar results at the tails of the θ scale (see Figure 13). However, the difference in the middle was more visible regarding SE and RMSE and MFI produced lower values, particularly for Clusters 1, 23, 45, and 68.

Table 4 includes MSE, MAbias, and MRMSE by routing, cluster, and domain. The green cells indicate the best-performing (i.e., lowest value) routing method for each cluster and domain. For Listening, the difference between MFI and the Access routing method is small, but the latter resulted in lower measurement error for Clusters 45 and 68. In contrast, MFI worked better for Cluster 23. The pattern was not consistent for Clusters 1 and 912. For Reading, MFI produced the lower SE and RMSE, but higher MAbias except for Cluster 912, where the opposite pattern was observed.

Correlation. The correlation between θ and $\hat{\theta}$ are given in Table 5. For Listening, MFI yielded a higher correlation for Clusters 23 and 912, whereas the Access yielded a higher correlation for Clusters 1, 45, and 68. For Reading, MFI produced a higher correlation for all clusters but Cluster 912.

Module Exposure. Tables 6 and 7 provide the percentage of students routed to each folder under different routing conditions for Listening and Reading, respectively.

Some folders were not administered to any students under MFI (see purple cells) because the corresponding TIFs were not maximized at any theta points (see Figures 2 and 4, 6, 7, 8, 9). Taken together, the module exposure rates had an impact on the FT process (e.g., FT administration across tiers, sample size, sample proportion) when the folders were not administered randomly. For example, the Tier B Listening folders at Stage 3 were not administered to any Cluster 23 students since there were no Tier B students at Stage 3 under MFI. Therefore, while Condition 1.1 did not include Tier B students, Condition 2 did not include these folders. Another example is the Tier B Reading folders administered based on tier estimate at Stage 9. These folders were administered to less than 300 Tier C students, so sample size requirements were not met for the conditions greater than 500 per folder under Condition 1.1.

Parameter Estimates

As described above, some clusters and domains did not meet the sample size requirement when FT condition was not Condition 3, which is the reason for not applicable (NA) values in some tables. Additionally, the simulation results followed an extremely similar pattern for some study conditions. Therefore, selected results were reported in this section.

Recovery. Table 8 depicts MSE, MAbias, and MRMSE for *b*-parameter estimates aggregated across all of the Listening FT items administered to Cluster 912 examinees by sample size, routing, and FT sampling conditions. The green and orange cells indicate the lowest and highest values under each sample size condition.

As shown in Table 8, Condition 2 consistently yielded lower MSE, whereas Condition 3 yielded the highest. However, the difference between FT conditions was small when the sample size was fixed. Similarly, the two routing methods produced comparable values, implying that the calibration precision is steady across FT sampling and routing conditions. It is clear that the magnitudes of MSE decrease as sample size increases. The descending rates were higher as the sample size increased from 500 to 3000 and lower as

the sample size increased from 3000 to 5000.

Regarding MAbias, Conditions 2 and 3 produced the highest and lowest values, respectively. When comparisons between Access and MFI routing were made, the latter produced lower MAbias only under Condition 2. The sample size did not produce a notable pattern in MAbias, which is consistent with the relevant literature (e.g., Cai, 2018). The MRMSE values, which is a function of both MSE and MAbias, are dominated by the latter. Therefore, the pattern is consistent with the pattern of MAbias values.

When Table 9 was created for other clusters and domains, a similar pattern was observed for MSE, and MRMSE, in which the difference in MSE values was small across routing conditions, increase in the sample size results in a decrease in MSE, MRMSE values follow the pattern of MAbias, Condition 3 consistently produced the lowest MAbias, and Condition 2 yielded in the highest MAbias. However, the routing condition yielding the lowest and highest MAbias seemed to differ across domains and clusters. As seen in Table 9, MFI yielded the lower MAbias compared to Access for Clusters 1 and 23 of Listening and Clusters 1 and 912 of Reading.

As described before, the FT items vary in their target tier and (see Tables 2 and 3) are administered based on tier estimates at varying stages except under Condition 3. Therefore, MSE, MAbias, and MRMSE statistics were computer per folder to understand whether the pattern differs by folder. Table 10 presents the statistics for the entry folder, which was administered to examinees independent of their tier estimates. In other words, all FT conditions in column two are equivalent (i.e., random) for this folder. Therefore, the differences between FT conditions can be explained by the random error, and different response patterns can explain differences between routing methods for OP items as a result of different routing decisions. Thus, this table facilitates a meaningful comparison of differences among the routing methods. Consistent with Table 8, Access and MFI routing are comparable with respect to MSE. However, the former produced lower MAbias for this folder. As mentioned previously, MRMSE values followed the pattern of MAbias.

When the folder-level comparison was made for other folders, the pattern of MAbias seemed to differ by folder tier and the stage examinee tiers estimates. Therefore, MAbias for folders administered to Cluster 912 examinees are given in Tables 11 and 12 for further discussion. Recall that Stage information is irrelevant for Condition 3. The findings only for $N = 3000$ were given since the best-performing (i.e., lowest MAbias) condition was similar across different sample size conditions. As highlighted in Table 11, Condition 3 yielded in lowest MAbias for all Tier B folders except for Folder 5. Furthermore, the Access routing outperformed MFI for four folders. While Condition 1.2 yielded the lowest MAbias for two Tier C folders, Condition 3 yielded the lowest MAbias for another two Tier C folders (see Table 12). Note that when more than one folder at the same tier level is administered at a given stage, the folders are randomly assigned to examinees. Therefore, these folders were expected to perform the best under the same condition. However, this was not observed in some cases, such as Tier B folders in Stage 5 and Tier C folders in Stage 3. Since the groups are randomly equivalent, the difference can be attributed to the generating item parameters in each folder.

To further investigate the variation at the folder level across domains and clusters, the number of folders producing the lowest MAbias in each routing and FT conditions was calculated (see Table 13) under the $N = 3000$ condition. For folders where $N = 3000$ was not met (e.g., Listening Cluster 23), results from the next largest sample size were used. One clear pattern is that Condition 3 consistently worked the best for Tier A folders. The combination of various factors, including but not limited to actual folder difficulties, misrouting, sample distribution at a given stage, and the accuracy of tier estimates at a given stage, might have contributed to the variation in Tier B and C folders.

InFit and OutFit Statistics. Figure 14 plots the infit and outfit statistics for Listening Cluster 912 under $N = 3000$. While the values seem to differ across studied conditions, they remain within the acceptable range of .5 and 1.5. Tables 14 and 15 include averaged values across FT under $N = 3000$. Consistent with the plots, the values seemed

to differ by routing and FT sampling condition. However, the differences are usually ignorable, and the fit statistics were acceptable under all studied conditions.

Correlation. Table 16 depicts the mean correlation across clusters and domains under $N = 3000$. Condition 3 produced the largest average item correlation for all domains and clusters except for Cluster 45 of Listening. A consistent pattern was not observed with regard to the routing method, but the two routing methods yielded comparable results.

When correlations were investigated at the item level, it was observed that a total of five Reading items (CAL IDs of 20339, 20171, 20130, 20392, 20393) had .2 or less correlation under Condition 2.

Discussion

The student population, MST panels, and FT folders represented the operational data. Still, they are not comprehensive, and the results might look different when other sets of operational data are used. Therefore, the generalizability of findings is bounded by the data and study conditions considered in this study.

Theta Estimates

This study was unable to provide a clear answer concerning the most appropriate routing method for Access Online. The lack of a definitive pattern in the routing methods could be attributed to inconsistent differences in folder-level TIFs and TCFs (i.e., Tier A-B or B–C) across stages at a given cluster and domain. The routing methods must be explored using panels appropriate for both MFI and Access routing. Given that the selection of routing method can impact FT,it is crucial to investigate the appropriateness of the routing method in the context of Access Online.

Neither Access nor MFI routing did a good job of controlling module exposure rates, which again can be explained by the inconsistent TIFs and TCFs across stages. Although MFI yielded practical issues when TIFs were overlapped by impacting the FT administration and calibration, this may not be a major concern in practice considering that the TIFs will not overlap when the panels are built appropriately for MFI.

Furthermore, examinees might have been assigned to different tiers given a fixed $\hat{\theta}$ under MFI because TIFs do not consistently target the same ability range across the stages. This should be monitored when constructing panels so each tier matches its target ability range across stages and examinees are consistently administered the content aligned with their ability.

Parameter Estimates

Of the four FT sampling conditions, Condition 3 performed the best in recovering generating *b*-parameters and did not cause any technical problems (e.g., sample size) during the simulation. The challenge with Condition 3 is that examinees might be able to recognize which folder contains FT items when FT folders noticeably mismatch with their abilities (e.g., Tier A students are given Tier C folders). This mismatch might also impact examinees test-taking behavior and decrease examinee engagement. Further research should investigate how the mismatch between examinee ability and FT folder tier affects the results. Furthermore, it remains unclear why Condition 3 did not consistently produce the lowest MAbias for Tier B and C folders. Future research might expand this study by manipulating study conditions, such as *b*-parameters.

The findings of the present study indicate that 3000 examinees per folder are sufficient to achieve steady parameter estimates, and the sample size is little or no influence on the bias and fit statistics.

Table 1
Number of examinees by cluster and domain

Cluster	Listening	Reading
1	213,507	213,289
23	429,667	429,181
45	385,704	385,126
68	422,954	422,109
912	420,514	418,396

Table 2
Listening: FT Folders by cluster, tier, and stage

Cluster	ID	Tier	Stage
	3	Entry	NA
	4	B	3
1	1, 2	A	4
	5, 6	B	6
	7, 8	C	8
	5	Entry	NA
	6, 7	B	3
23	1, 2	A	4
	8, 9	C	5
	3, 4	A	6
	5	Entry	NA
	1, 2	A	4
45	6	B	5
	3	A	6
	4	B	6
	3	Entry	NA
	1, 2	A	4
68	4	B	4
	6, 7	C	4
	5, 8	C	8
	1	Entry	NA
	2, 3	B	3
912	4, 5	B	5
	6, 7	C	3
	8, 9	C	5

Table 3
Reading: FT Folders by cluster, tier, and stage

Cluster	ID	Tier	Stage
1	5	Entry	NA
	6, 7	B	3
	1, 2	A	4
	3, 4	A	6
	8	B	9
	9, 10	C	9
23	1	Entry	NA
	2, 3	B	3
	9, 10	C	3
	11, 12	C	4
	4, 5	B	6
	7, 8	C	7
	6	B	10
45	3	Entry	NA
	1	A	7
	6,7	B	6
	2	A	6
	15, 16	C	5
	11, 12	C	8
	13, 14	C	10
	8, 9	C	7
	10	C	4
	5	B	10
	4	B	8
68	6	Entry	NA
	3, 4	B	3
	11, 12	C	6
	2	B	7
	1	A	8
	8, 9	C	8
	7	B	9
	14, 15	C	9
	5	B	10
	10, 13	C	10
912	1	Entry	NA
	3, 4	B	4
	11, 12	C	4
	2	A	5
	7, 8	B	5
	10	C	7
	6, 9	B	9
	5	B	10
	13, 14	C	10

Table 4
Mean recovery indices for $\hat{\theta}$

Cluster	Routing	Listening			Reading		
		MSE	MABias	MRMSE	MSE	MABias	MRMSE
1	Access	0.9000	0.2392	0.9402	0.8580	0.2200	0.9176
	MFI	0.9168	0.2377	0.9564	0.8419	0.2288	0.9025
23	Access	0.9621	0.2720	1.0213	0.9936	0.2763	1.0552
	MFI	0.9543	0.2572	1.0113	0.9578	0.2862	1.0215
45	Access	0.8509	0.1600	0.8643	0.9957	0.3045	1.0686
	MFI	0.8589	0.1705	0.8752	0.9608	0.3330	1.0396
68	Access	1.0024	0.2588	1.0477	0.9078	0.1998	0.9322
	MFI	1.0029	0.2822	1.0522	0.8707	0.2011	0.8961
912	Access	0.9833	0.2337	1.0237	0.8410	0.2031	0.8695
	MFI	0.9726	0.2527	1.0162	0.8507	0.1970	0.8815

Table 5
Correlation between θ and $\hat{\theta}$

Cluster	Listening		Reading	
	Access	MFI	Access	MFI
1	0.8612	0.8595	0.8347	0.8456
23	0.8810	0.8868	0.8264	0.8378
45	0.8469	0.8449	0.8555	0.8603
68	0.8444	0.8427	0.8751	0.8887
912	0.8525	0.8584	0.8862	0.8819

Table 6
Listening: Percentage of examinees routed to each folder

Cluster	Tier	Routing	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8
1	A	Access	14.75	20.95	20.18	22.24		
		MFI	7.31	14.34	14.81	22.25		
	B	Access	24.25	29.01	30.89	25.81	35.00	33.95
		MFI	31.69	25.56	17.06	15.50	23.90	38.91
	C	Access	61.00	50.04	48.93	51.94	45.18	46.23
		MFI	61.00	60.10	68.13	62.24	56.41	41.40
23	A	Access	18.17	19.66	22.72	21.45		
		MFI	25.23	13.94	18.22	31.73		
	B	Access	32.32	26.77	26.05	32.50	35.90	36.08
		MFI	0.00	20.91	14.14	9.05	18.12	10.54
	C	Access	49.50	53.57	51.23	46.05	43.32	43.13
		MFI	74.77	65.15	67.64	59.22	60.69	68.27
45	A	Access	3.90	4.03	3.91	3.79		
		MFI	9.82	11.12	6.55	5.32		
	B	Access	14.97	15.09	19.26	18.95	19.48	20.23
		MFI	32.65	14.35	21.34	20.40	23.95	11.68
	C	Access	81.13	80.88	76.84	77.27	76.92	76.17
		MFI	57.53	74.53	72.10	74.28	71.96	84.24
68	A	Access	6.79	7.60	7.82	8.20		
		MFI	20.66	21.06	27.52	13.35		
	B	Access	23.82	28.55	32.47	34.40	35.45	37.38
		MFI	0.00	23.62	23.91	38.66	26.71	31.25
	C	Access	69.39	63.85	59.72	57.39	57.25	55.33
		MFI	79.34	55.32	48.57	47.99	65.45	60.91
912	A	Access	17.80	19.42	18.95	20.75		
		MFI	28.57	46.06	28.19	22.88		
	B	Access	29.22	43.37	47.14	46.99	49.74	53.55
		MFI	18.45	3.56	11.58	30.42	31.63	44.90
	C	Access	52.98	37.22	33.92	32.26	32.51	28.70
		MFI	52.98	50.39	60.23	46.70	49.11	35.84

Table 7
Reading: Percentage of examinees routed to each folder

Cluster	Tier	Routing	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Stage 9	Stage 10
1	A	Access	25.29	25.90	24.47	29.90	27.35	29.52		
		MFI	66.36	23.92	49.18	48.09	28.22	30.05		
	B	Access	41.07	42.66	40.52	39.57	43.05	42.02	49.01	49.66
		MFI	0.00	43.39	35.93	33.84	47.06	53.89	67.43	65.73
	C	Access	33.64	31.44	35.00	30.54	29.60	28.46	28.39	27.75
		MFI	33.64	32.69	14.89	18.07	24.72	16.05	11.09	12.79
23	A	Access	3.53	5.36	5.55	8.69	6.79	8.59		
		MFI	51.39	23.85	48.43	41.14	44.52	45.23		
	B	Access	17.60	19.60	26.16	21.39	24.86	25.38	17.17	19.25
		MFI	0.00	40.75	4.82	30.03	27.63	33.96	66.81	56.43
	C	Access	78.87	75.04	68.29	69.92	68.35	66.02	64.99	62.91
		MFI	48.61	35.40	46.75	28.83	27.85	20.81	20.45	30.84
45	A	Access	8.28	7.89	9.38	9.64	10.33	10.66		
		MFI	32.75	32.76	56.62	31.47	35.18	32.63		
	B	Access	16.29	21.89	23.12	22.35	22.33	22.44	20.10	20.30
		MFI	20.13	21.45	10.26	35.25	31.70	35.47	59.21	52.01
	C	Access	75.43	70.22	67.50	68.01	67.34	66.90	63.87	63.67
		MFI	47.11	45.79	33.12	33.28	33.11	31.90	25.32	32.52
68	A	Access	34.21	35.18	35.95	36.73	39.77	39.16		
		MFI	46.37	46.56	56.43	41.18	34.98	42.83		
	B	Access	31.12	37.07	37.98	40.64	39.07	40.33	47.78	48.22
		MFI	18.95	30.54	0.00	40.66	36.70	35.80	49.01	50.12
	C	Access	34.67	27.75	26.06	22.63	21.16	20.52	19.56	19.12
		MFI	34.67	22.89	43.57	18.16	28.32	21.37	15.92	14.81
912	A	Access	28.40	29.97	31.44	32.71	34.72	35.18		
		MFI	14.52	33.29	40.74	42.40	43.77	40.30		
	B	Access	27.87	31.04	32.30	36.87	34.47	34.45	41.31	7.88
		MFI	29.97	29.10	21.62	30.45	32.36	30.40	65.65	50.73
	C	Access	43.73	38.99	36.26	30.42	30.81	30.37	29.73	63.17
		MFI	55.52	37.61	37.65	27.15	23.87	29.30	4.51	19.42

Table 8
Listening and Cluster 912: b-parameter recovery for the 24 Listening FT items

N	FT	MSE		MABias		MRMSE	
		Access	MFI	Access	MFI	Access	MFI
500	Condition 1.1	0.1172	0.1129	0.1173	0.1224	0.1688	0.1715
	Condition 1.2	0.1163	0.1166	0.1078	0.1208	0.1645	0.1792
	Condition 2	0.1057	0.1061	0.1445	0.0919	0.1875	0.1445
	Condition 3	0.1158	0.1162	0.0722	0.1165	0.1399	0.1658
1000	Condition 1.1	0.0773	0.0764	0.1199	0.1221	0.1444	0.1476
	Condition 1.2	0.0801	0.0802	0.1095	0.1181	0.1406	0.1526
	Condition 2	0.0749	0.0738	0.1404	0.0897	0.1649	0.1200
	Condition 3	0.0823	0.0842	0.0773	0.1136	0.1157	0.1428
2000	Condition 1.1	0.0554	0.0545	0.1227	0.1244	0.1362	0.1380
	Condition 1.2	0.0588	0.0573	0.1097	0.1186	0.1274	0.1401
	Condition 2	0.0536	0.0535	0.1443	0.0927	0.1576	0.1099
	Condition 3	0.0609	0.0582	0.0747	0.1123	0.0999	0.1272
3000	Condition 1.1	0.0459	0.0433	0.1225	0.1230	0.1319	0.1322
	Condition 1.2	0.0477	0.0463	0.1100	0.1184	0.1231	0.1341
	Condition 2	0.0423	0.0432	0.1422	0.0931	0.1510	0.1050
	Condition 3	0.0492	0.0492	0.0753	0.1123	0.0929	0.1235
4000	Condition 1.1	0.0405	0.0371	0.1230	0.1243	0.1305	0.1312
	Condition 1.2	0.0416	0.0393	0.1097	0.1181	0.1199	0.1308
	Condition 2	0.0367	0.0364	0.1421	0.0913	0.1491	0.1004
	Condition 3	0.0422	0.0403	0.0769	0.1119	0.0902	0.1194
5000	Condition 1.1	0.0344	0.0320	0.1241	0.1219	0.1293	0.1272
	Condition 1.2	0.0363	0.0345	0.1106	0.1187	0.1188	0.1288
	Condition 2	0.0315	0.0337	0.1427	0.0909	0.1479	0.0989
	Condition 3	0.0373	0.0363	0.0773	0.1118	0.0885	0.1181

Table 9
Routing condition yielding the lowest MAbias for FT sampling

Cluster	Listening	Reading
1	MFI	MFI
23	MFI	Access
45	Access	Access
68	Access	Access
912	Access	MFI

Table 10
Listening and Cluster 912: b-parameter recovery for entry folder (ID: 1)

N	FT	MSE		MABias		MRMSE	
		Access	MFI	Access	MFI	Access	MFI
500	Condition 1.1	0.1265	0.1272	0.1009	0.1715	0.1614	0.2134
	Condition 1.2	0.1308	0.1266	0.0922	0.1731	0.1595	0.2141
	Condition 2	0.1196	0.1296	0.1041	0.1316	0.1590	0.1843
	Condition 3	0.1202	0.1144	0.1111	0.1534	0.1637	0.1913
1000	Condition 1.1	0.0831	0.0898	0.1007	0.1598	0.1304	0.1831
	Condition 1.2	0.0862	0.0920	0.1084	0.1613	0.1383	0.1855
	Condition 2	0.0838	0.0774	0.0969	0.1360	0.1281	0.1566
	Condition 3	0.0848	0.0942	0.1016	0.1604	0.1321	0.1863
2000	Condition 1.1	0.0616	0.0621	0.1034	0.1641	0.1204	0.1755
	Condition 1.2	0.0643	0.0619	0.1020	0.1713	0.1206	0.1821
	Condition 2	0.0616	0.0642	0.0985	0.1440	0.1164	0.1576
	Condition 3	0.0590	0.0610	0.1085	0.1493	0.1234	0.1612
3000	Condition 1.1	0.0493	0.0473	0.0919	0.1655	0.1043	0.1722
	Condition 1.2	0.0483	0.0512	0.0998	0.1679	0.1109	0.1754
	Condition 2	0.0505	0.0514	0.0978	0.1447	0.1104	0.1534
	Condition 3	0.0491	0.0501	0.1025	0.1533	0.1137	0.1613
4000	Condition 1.1	0.0462	0.0434	0.0996	0.1722	0.1099	0.1776
	Condition 1.2	0.0428	0.0429	0.0979	0.1685	0.1069	0.1739
	Condition 2	0.0424	0.0409	0.0982	0.1373	0.1070	0.1433
	Condition 3	0.0438	0.0401	0.1011	0.1474	0.1101	0.1527
5000	Condition 1.1	0.0357	0.0339	0.1017	0.1703	0.1078	0.1736
	Condition 1.2	0.0354	0.0358	0.1033	0.1670	0.1092	0.1708
	Condition 2	0.0384	0.0357	0.0961	0.1365	0.1034	0.1411
	Condition 3	0.0345	0.0382	0.1055	0.1459	0.1110	0.1508

Table 11
Listening and Cluster 912: MAbias for Tier B folders

N	FT	Stage 3				Stage 5			
		ID : 2		ID : 3		ID : 4		ID : 5	
		Access	MFI	Access	MFI	Access	MFI	Access	MFI
3000	Condition 1.1	0.1677	0.1855	0.1522	0.1585	0.1483	0.1668	0.1420	0.1640
	Condition 1.2	0.1529	0.1955	0.1339	0.1790	0.1589	0.2022	0.1630	0.2067
	Condition 2	0.1176	0.1415	0.1233	0.1500	0.0916	0.0915	0.0927	0.0463
	Condition 3	0.0819	0.1357	0.0757	0.1239	0.0787	0.1237	0.0979	0.1168

Table 12
Listening and Cluster 912: MAbias for Tier C folders

N	FT	Stage 3				Stage 5			
		ID : 6		ID : 7		ID : 8		ID : 9	
		Access	MFI	Access	MFI	Access	MFI	Access	MFI
3000	Condition 1.1	0.0427	0.0791	0.0858	0.0644	0.1198	0.0664	0.1522	0.0565
	Condition 1.2	0.0146	0.0251	0.0442	0.0135	0.1002	0.0196	0.1221	0.0562
	Condition 2	0.0641	0.0585	0.0764	0.0494	0.2955	0.0638	0.3208	0.0922
	Condition 3	0.0576	0.0814	0.1005	0.0946	0.0199	0.0940	0.0626	0.0877

Table 13
Number of folders yielding the lowest MAbias ($N = 3000$)

FT	Tier A		Tier B		Tier C	
	Access	MFI	Access	MFI	Access	MFI
Condition 1.1	1	0	3	6	5	2
Condition 1.2	0	0	4	4	7	12
Condition 2	0	0	2	6	1	0
Condition 3	8	10	8	3	9	6

Note. For folders in which $N = 3000$ were not met, the results from the largest available sample size condition were used.

Table 14
Listening: Mean infit and outfit statistics (N = 3000)

Cluster	FT	Infit		Outfit	
		Access	MFI	Access	MFI
1	Condition 1.1	1.0187	1.0237	1.0169	1.0246
	Condition 1.2	1.0486	1.0500	1.0636	1.0665
	Condition 2	0.9955	0.9924	0.9792	0.9705
	Condition 3	1.0694	1.0728	1.0974	1.1045
23	Condition 1.1	1.0135	1.0281	1.0140	1.0161
	Condition 1.2	1.0211	1.0153	1.0253	1.0150
	Condition 2	NA	0.9730	NA	0.9420
	Condition 3	1.0574	1.0521	1.0832	1.0804
45	Condition 1.1	0.9950	0.9984	0.9519	0.9583
	Condition 1.2	1.0003	1.0083	0.9708	0.9790
	Condition 2	0.9736	0.9789	0.9255	0.9296
	Condition 3	1.0331	1.0415	1.0170	1.0288
68	Condition 1.1	1.0674	1.0604	1.0693	1.0603
	Condition 1.2	1.0699	1.0676	1.0765	1.0784
	Condition 2	1.0440	1.0368	1.0435	1.0368
	Condition 3	1.0772	1.0876	1.0741	1.0923
912	Condition 1.1	1.0460	1.0281	1.0594	1.0353
	Condition 1.2	1.0563	1.0564	1.0749	1.0820
	Condition 2	1.0111	1.0118	1.0229	1.0240
	Condition 3	1.0604	1.0661	1.0785	1.0877

Table 15
Reading: Mean infit and outfit statistics (N = 3000)

Cluster	FT	Infit		Outfit	
		Access	MFI	Access	MFI
1	Condition 1.1	1.0643	1.0543	1.0763	1.0634
	Condition 1.2	1.0656	1.0525	1.0804	1.0625
	Condition 2	1.0337	1.0206	1.0367	1.0291
	Condition 3	1.0895	1.0656	1.1042	1.0709
23	Condition 1.1	1.1443	1.0639	1.1899	1.0900
	Condition 1.2	1.1233	1.0671	1.1741	1.0948
	Condition 2	1.0843	1.0353	1.1127	1.0569
	Condition 3	1.1297	1.0869	1.1935	1.1253
45	Condition 1.1	1.1108	1.0797	1.1281	1.1193
	Condition 1.2	1.1202	1.0953	1.1700	1.1408
	Condition 2	1.0654	1.0529	1.0844	1.0926
	Condition 3	1.1274	1.1158	1.1762	1.1654
68	Condition 1.1	1.0480	1.0551	1.0556	1.0695
	Condition 1.2	1.0487	1.0518	1.0446	1.0542
	Condition 2	1.0263	1.0289	1.0348	1.0434
	Condition 3	1.0727	1.0768	1.0483	1.0567
912	Condition 1.1	NA	NA	NA	NA
	Condition 1.2	1.0739	1.0851	1.0879	1.1130
	Condition 2	1.0243	1.0397	1.0260	1.0597
	Condition 3	1.0840	1.1014	1.0819	1.1123

Table 16
Mean correlation between FT items and $\hat{\theta}$

Cluster	FT	Listening		Reading	
		Access	MFI	Access	MFI
1	Condition 1.1	0.4815	0.4735	0.3798	0.3960
	Condition 1.2	0.4932	0.4858	0.3915	0.3913
	Condition 2	0.3621	0.3526	0.3112	0.3205
	Condition 3	0.5308	0.5310	0.4319	0.4323
23	Condition 1.1	0.4911	0.4929	0.4383	0.4236
	Condition 1.2	0.4846	0.4679	0.4535	0.4227
	Condition 2	NA	0.4021	0.3674	0.3483
	Condition 3	0.5258	0.5260	0.4563	0.4556
45	Condition 1.1	0.5286	0.5139	0.4315	0.4089
	Condition 1.2	0.4515	0.4558	0.4348	0.4208
	Condition 2	0.3889	0.4195	0.3410	0.3310
	Condition 3	0.4736	0.4724	0.4603	0.4598
68	Condition 1.1	0.4534	0.4262	0.4292	0.4183
	Condition 1.2	0.4459	0.4225	0.4282	0.4178
	Condition 2	0.3723	0.3623	0.3544	0.3378
	Condition 3	0.4614	0.4613	0.4642	0.4649
912	Condition 1.1	0.4599	0.4338	NA	NA
	Condition 1.2	0.4740	0.4606	0.4540	0.4435
	Condition 2	0.3727	0.3684	0.3497	0.3529
	Condition 3	0.4846	0.4838	0.4835	0.4827

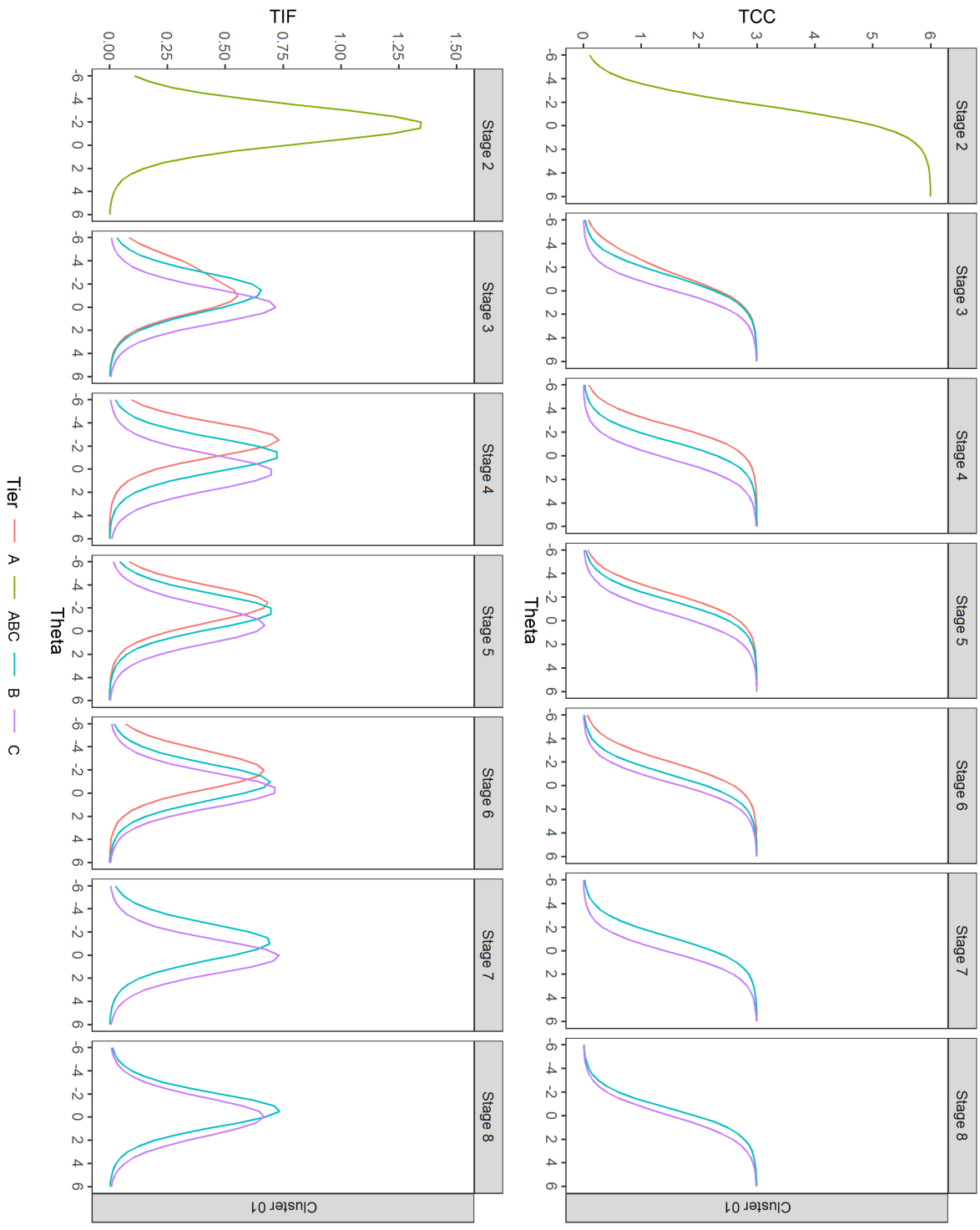

Figure 1. Listening: TCC and TIFs for Cluster 1 Folders

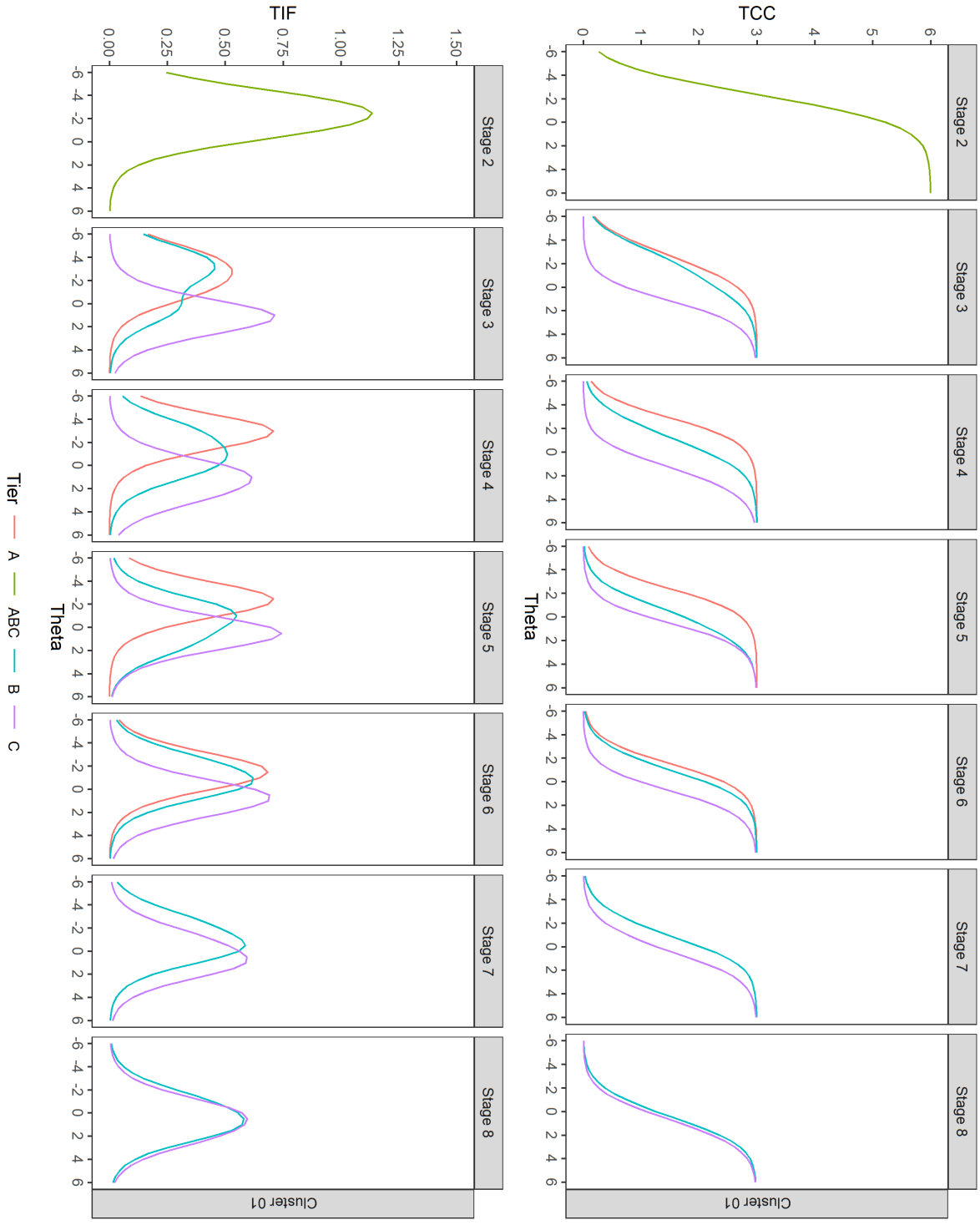

Figure 2. Listening: TCC and TIFs for Cluster 23 Folders

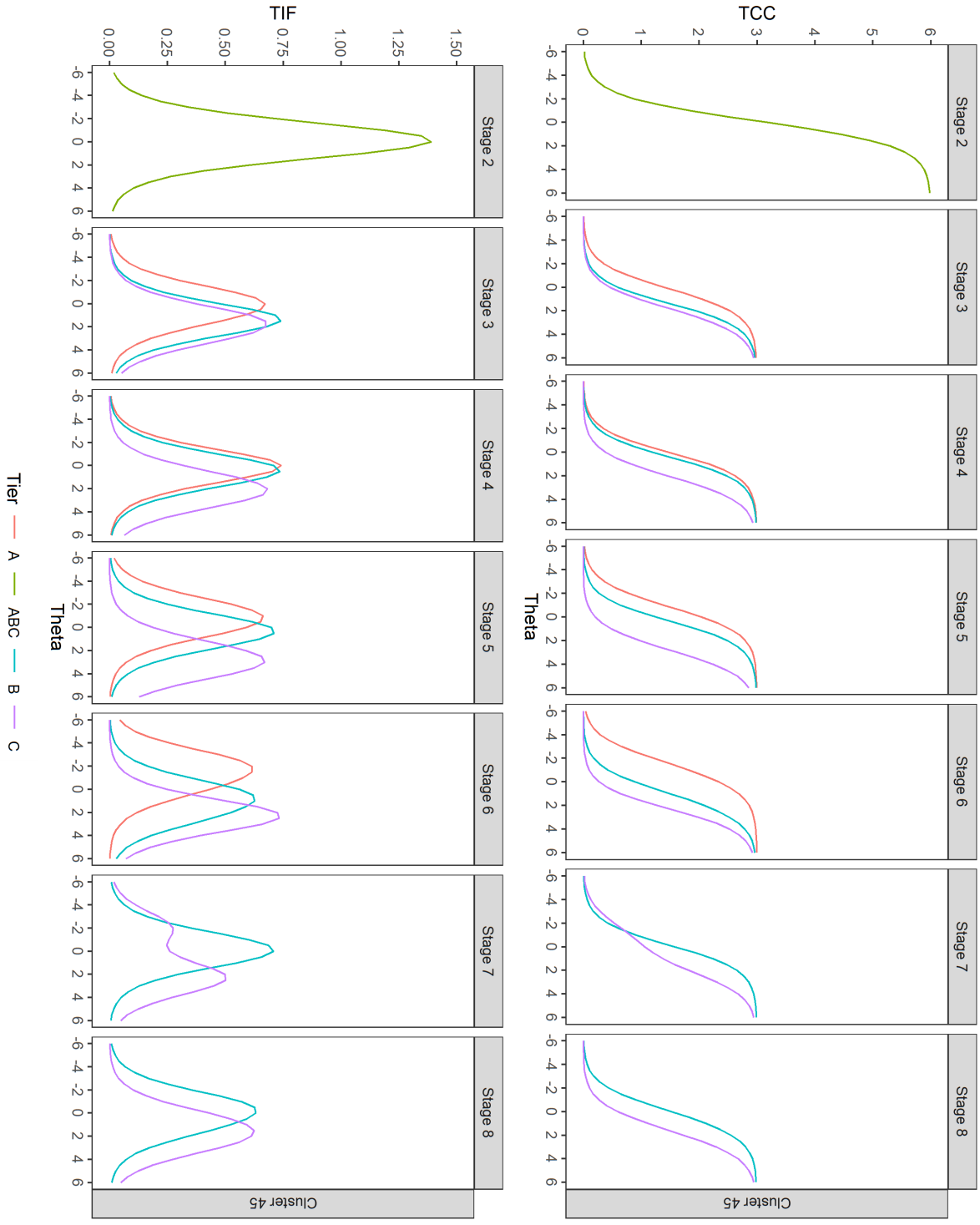

Figure 3. Listening: TCC and TIFs for Cluster 45 Folders

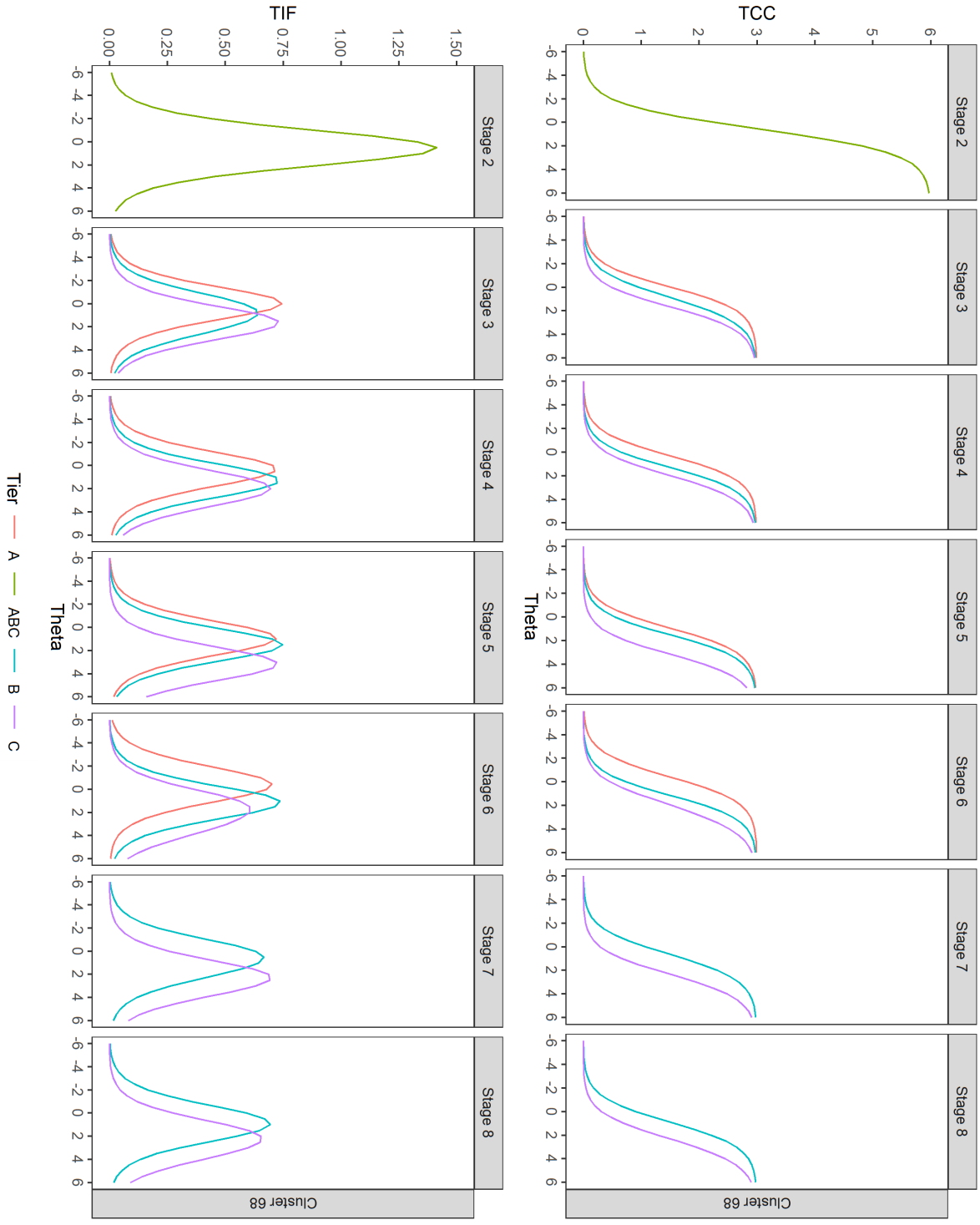

Figure 4. Listening: TCC and TIFs for Cluster 68 Folders

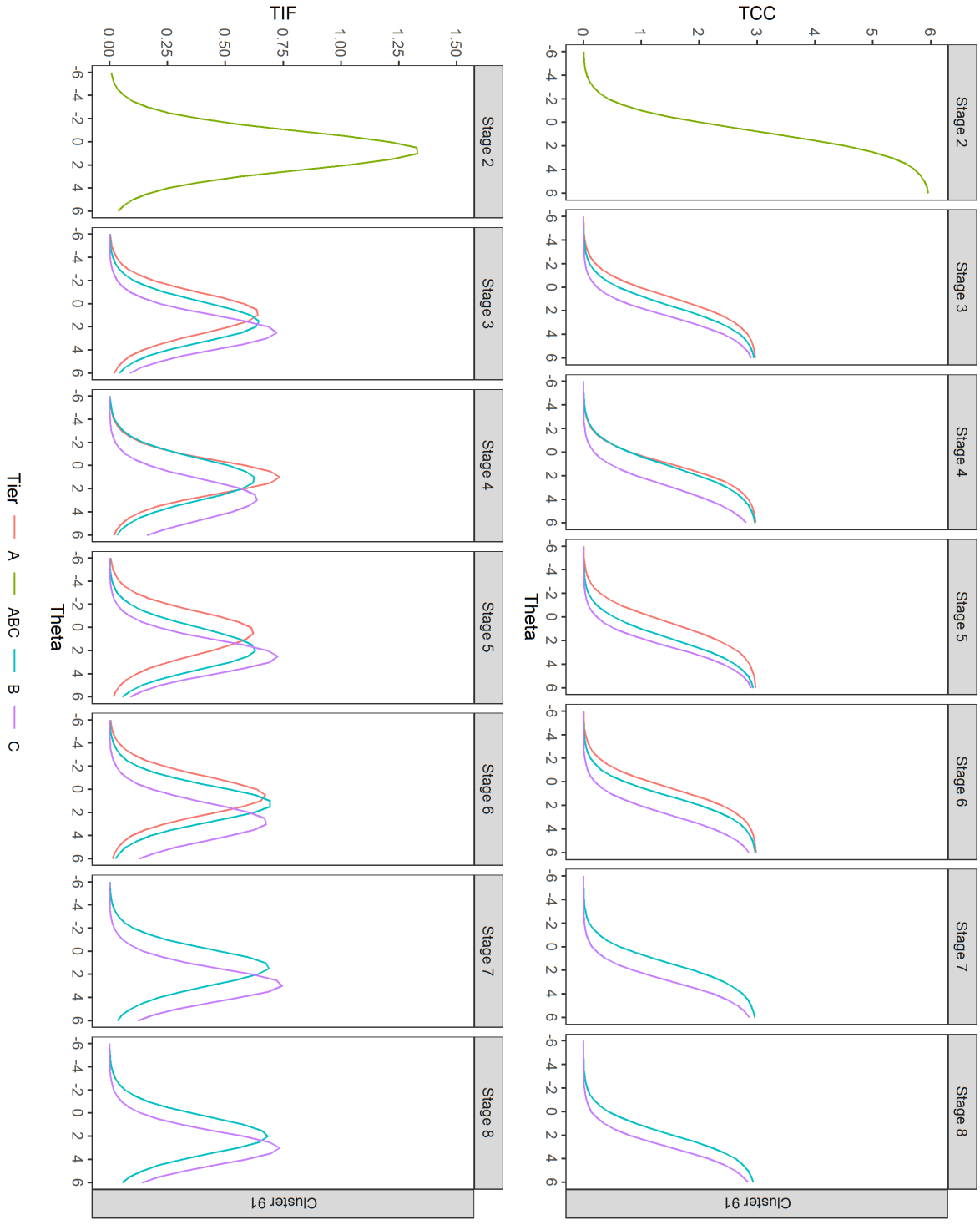

Figure 5. Listening: TCC and TIFs for Cluster 912 Folders

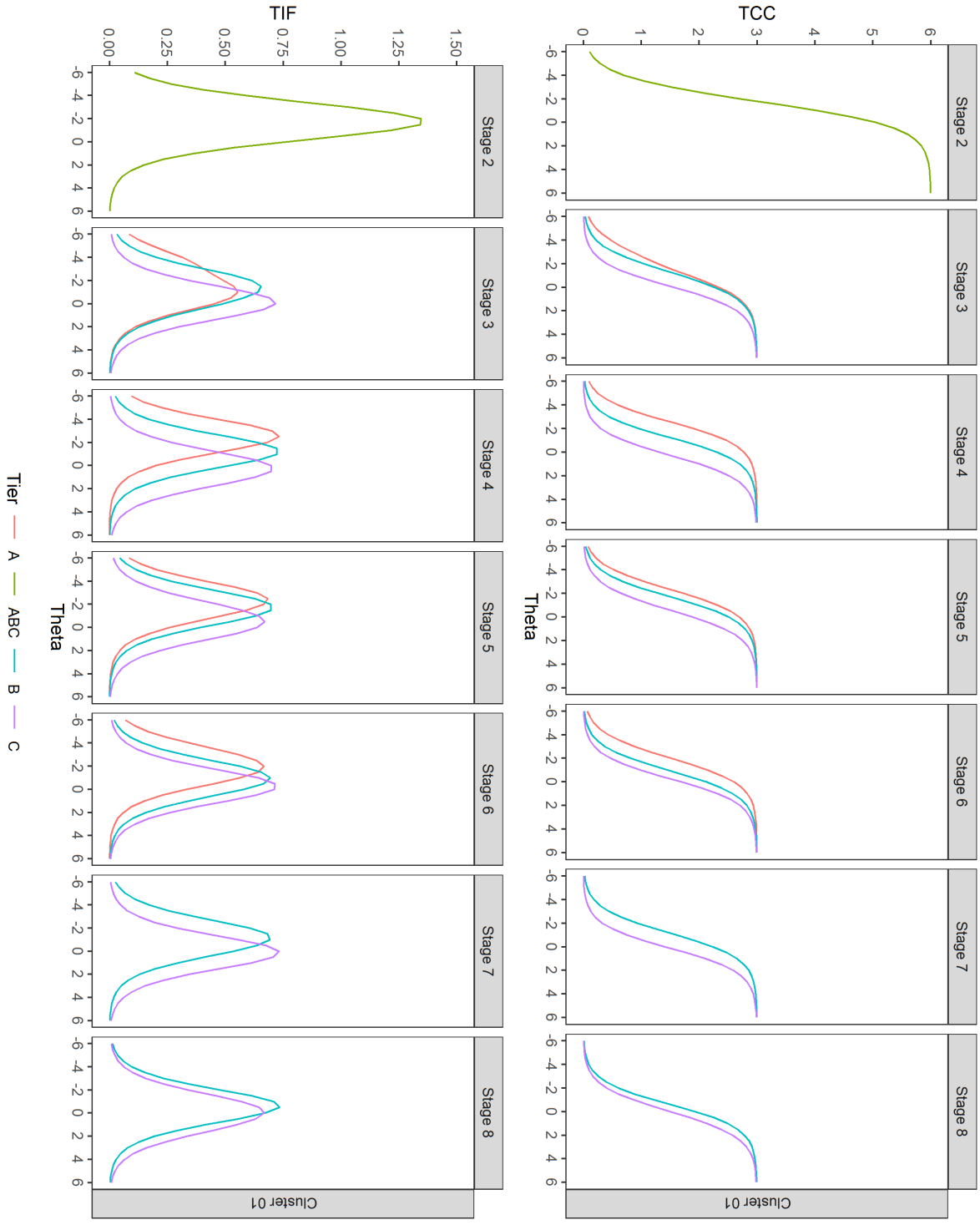

Figure 6. Reading TCC and TIFs for Cluster 1 Folders

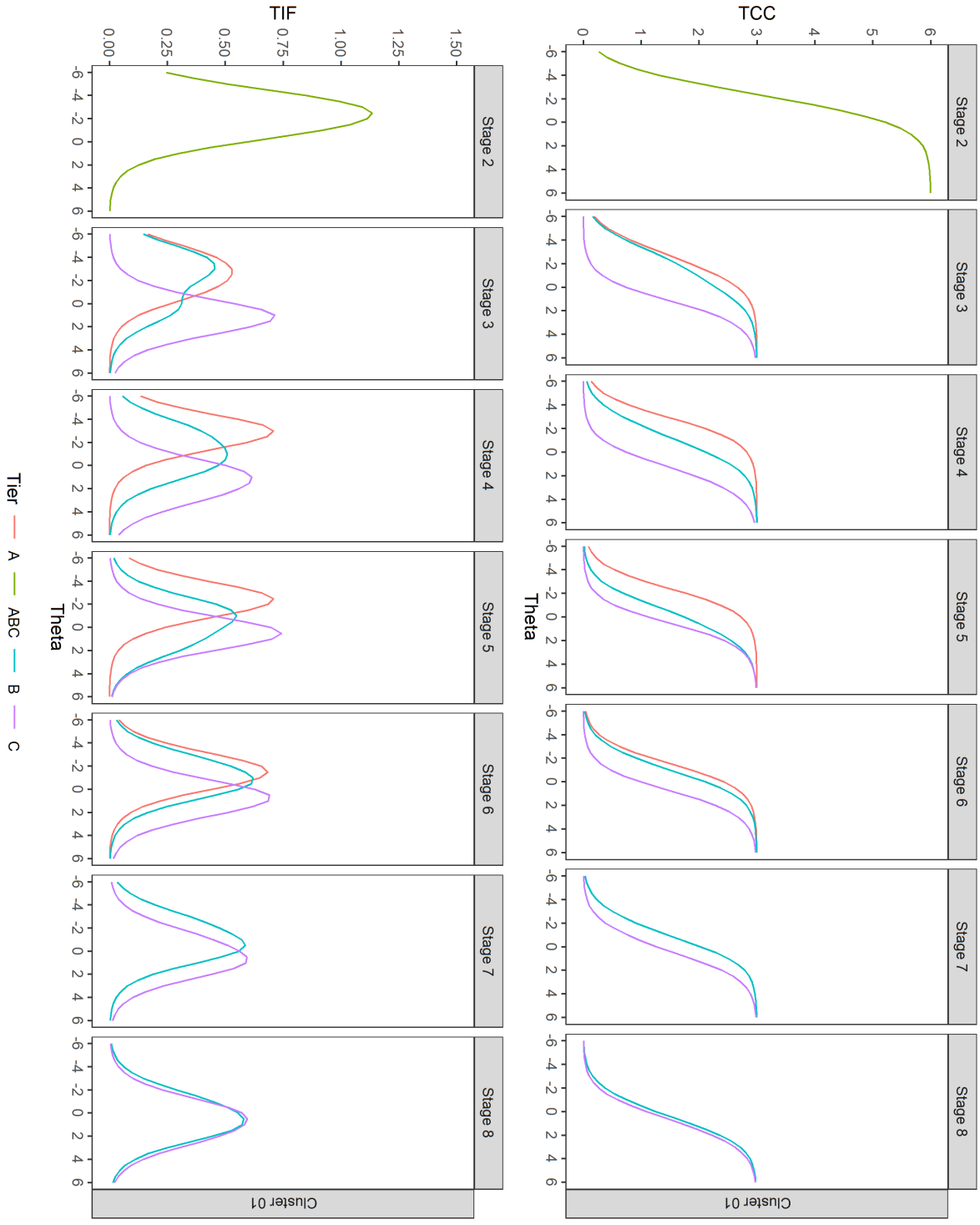

Figure 7. Reading TCC and TIFs for Cluster 23 Folders

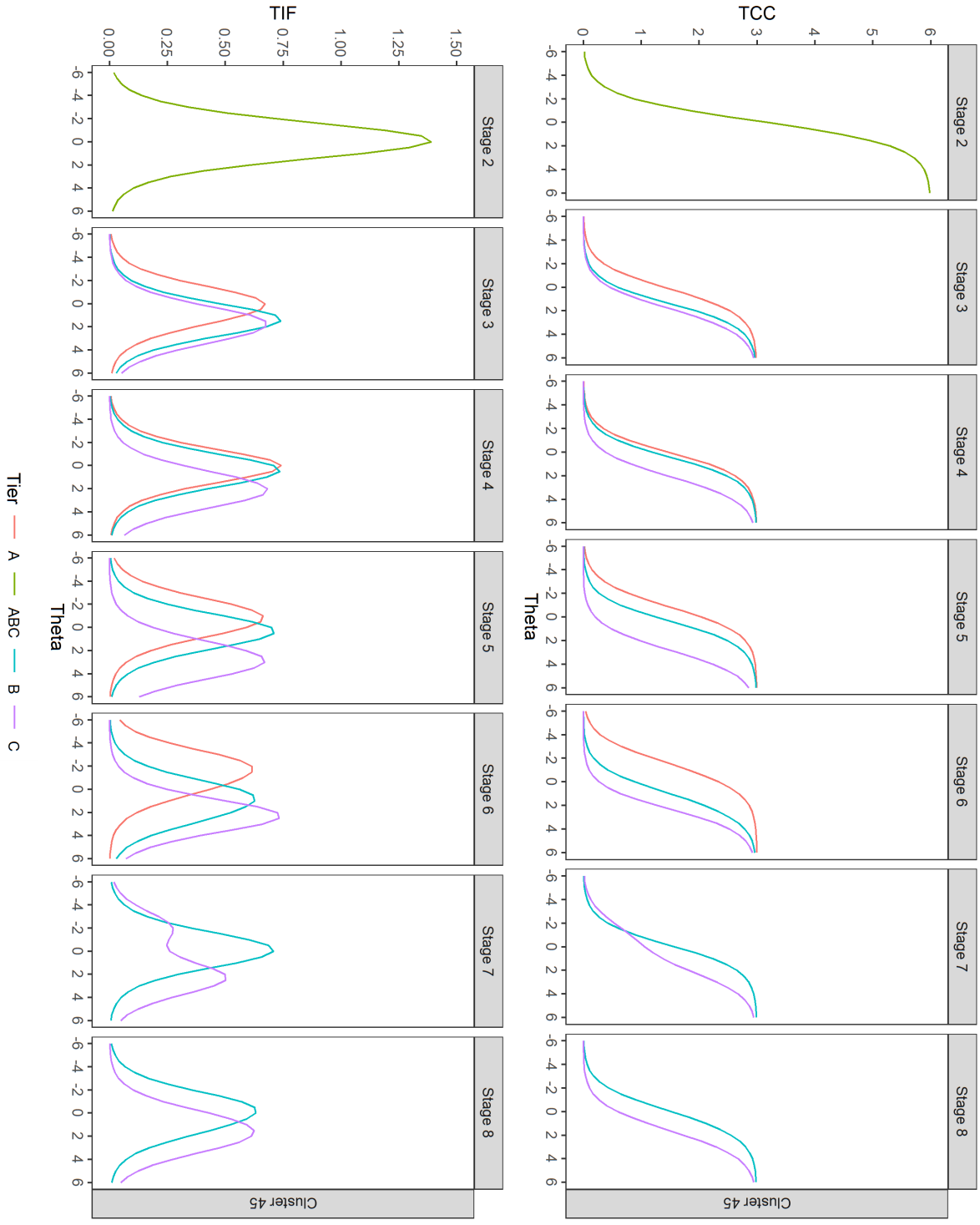

Figure 8. Reading TCC and TIFs for Cluster 45 Folders

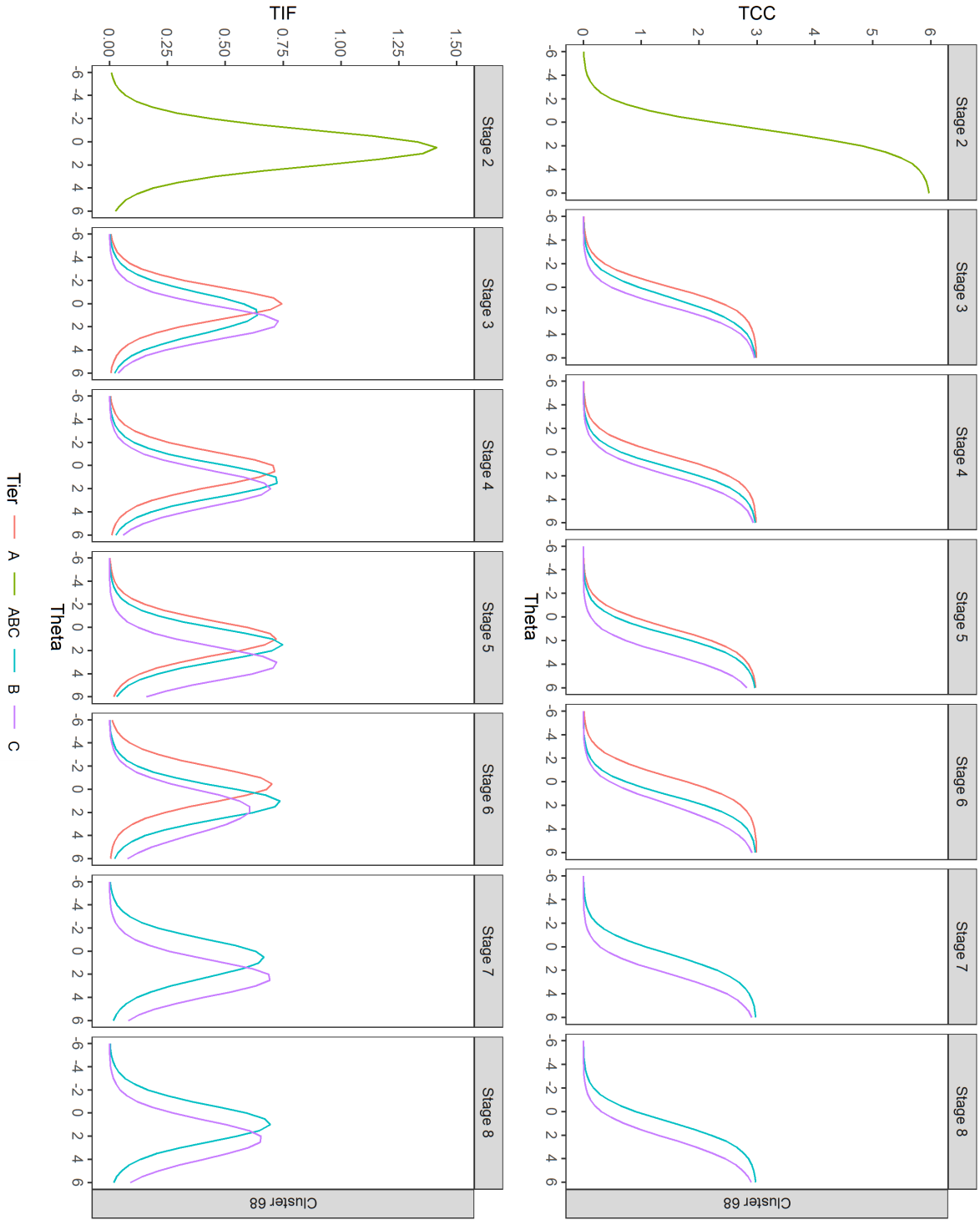

Figure 9. Reading TCC and TIFs for Cluster 68 Folders

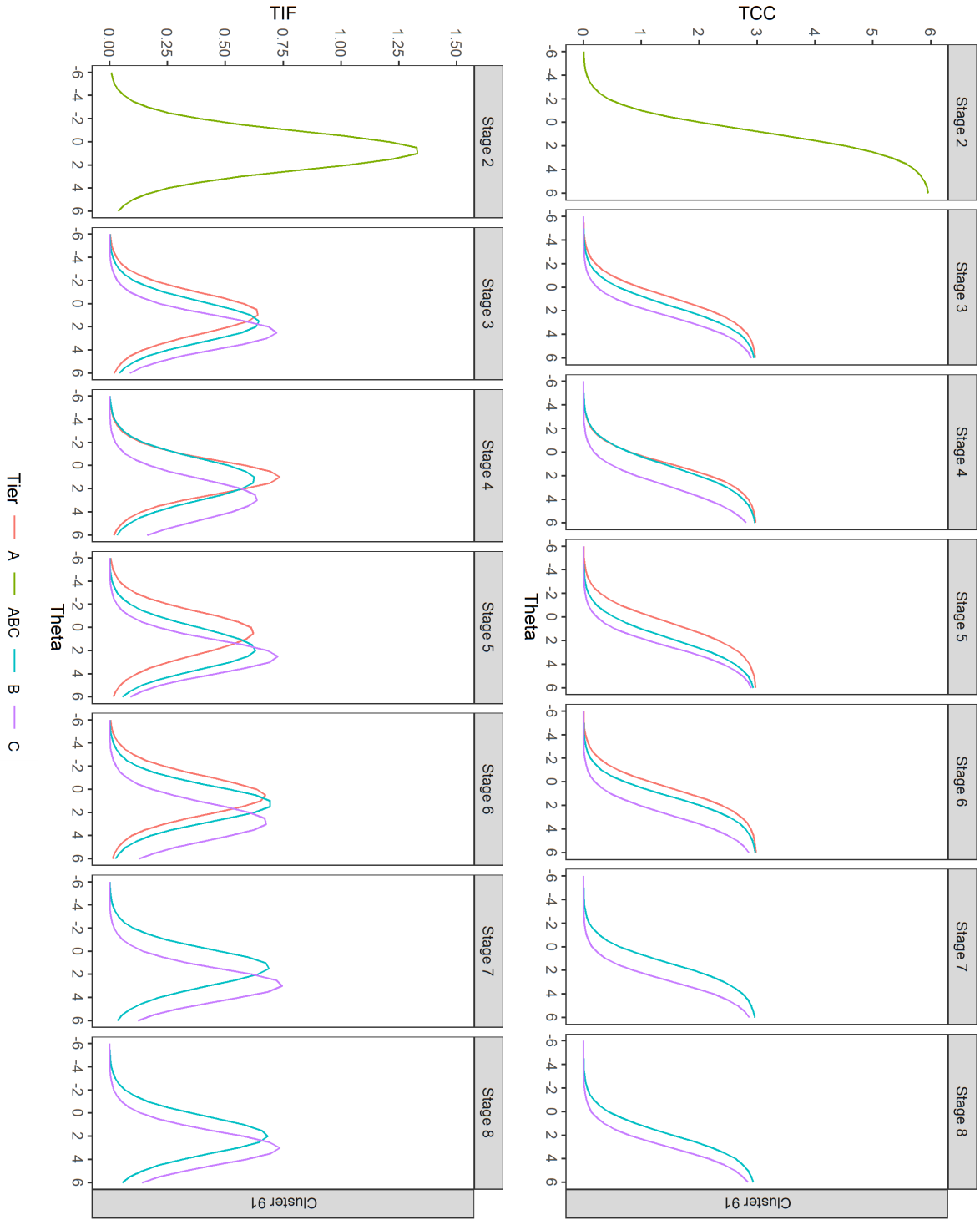

Figure 10. Reading TCC and TIFs for Cluster 912 Folders

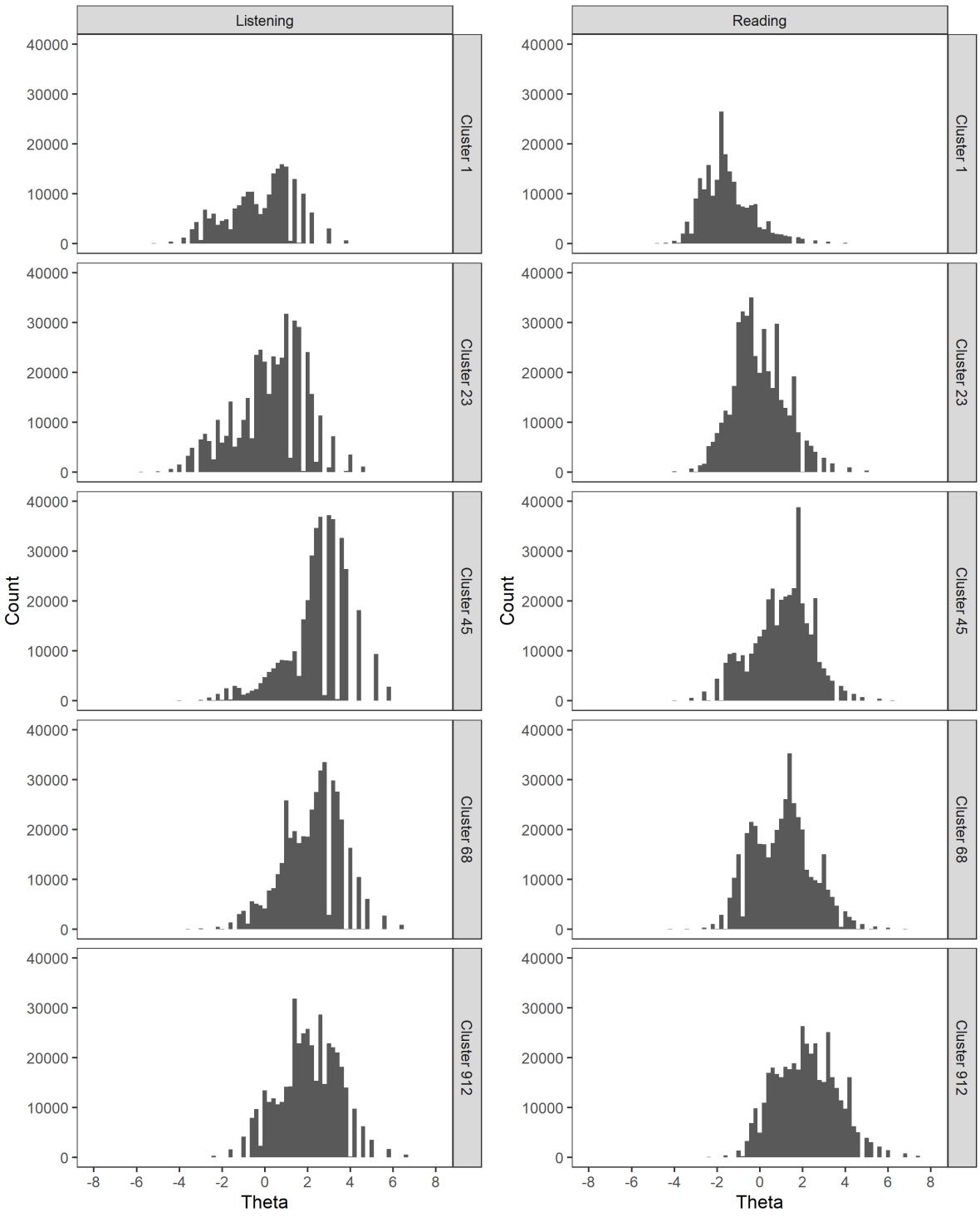

Figure 11. Generating proficiency distribution

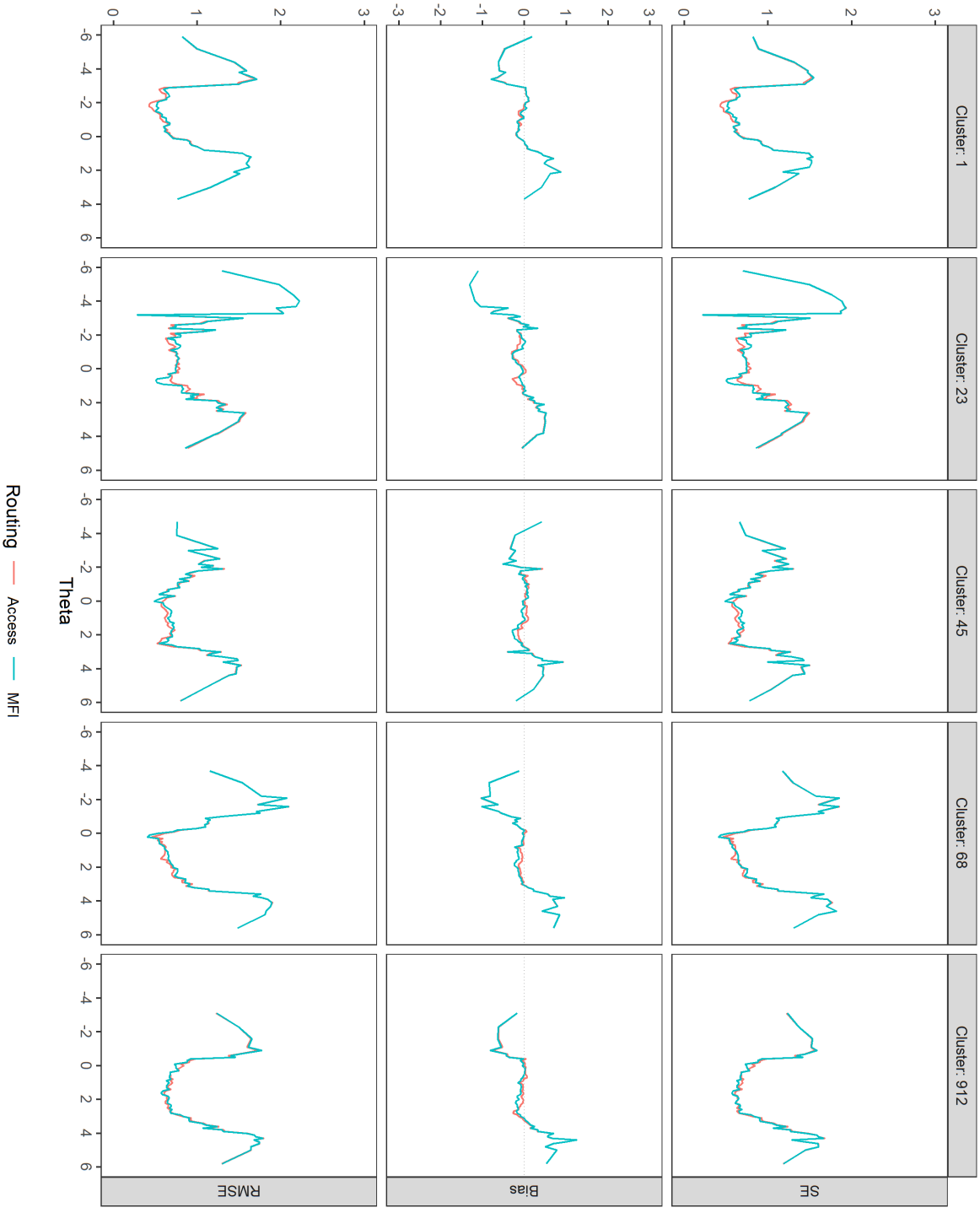

Figure 12. Listening: Recovery of individual theta scores

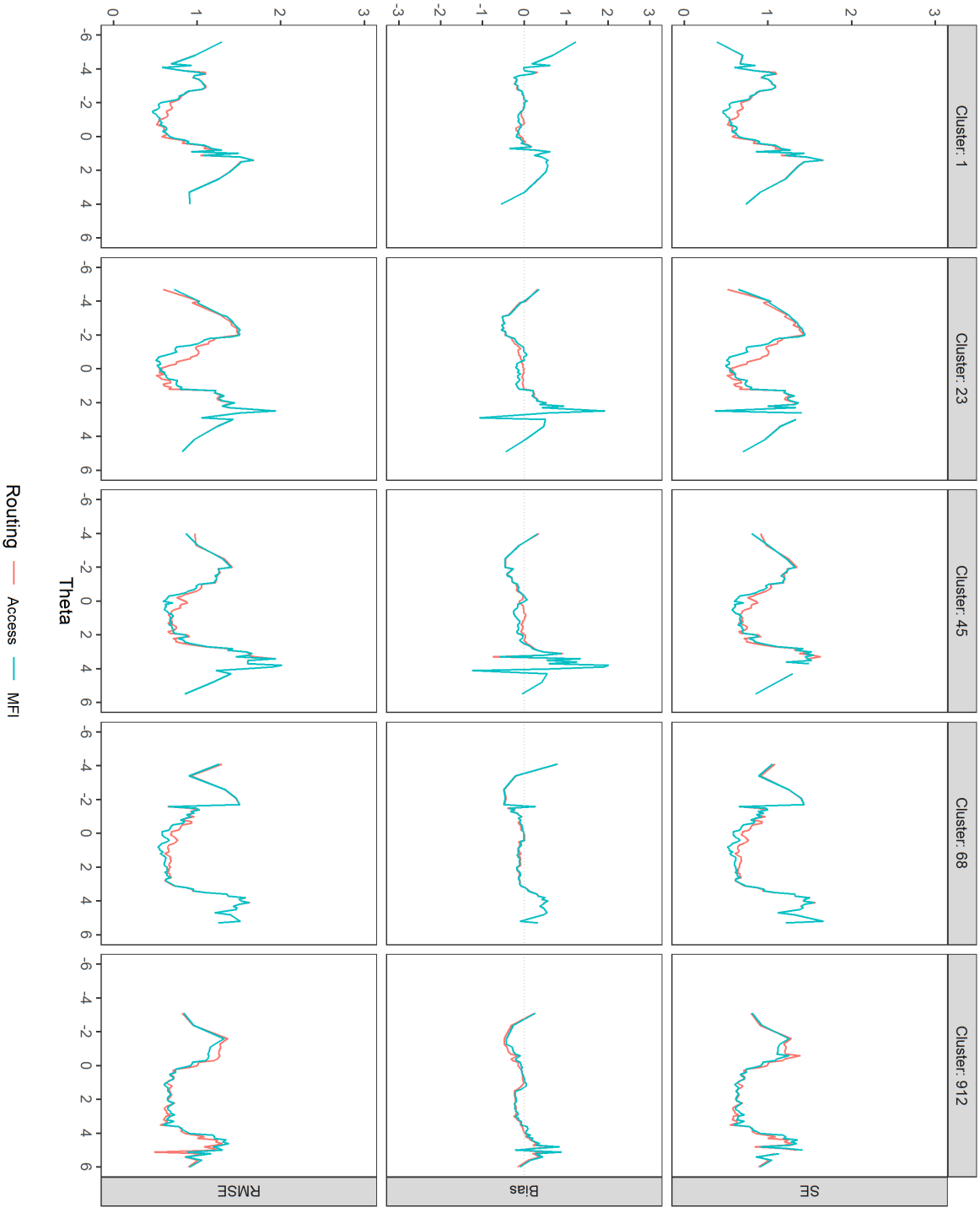

Figure 13. Reading: Recovery of individual theta scores

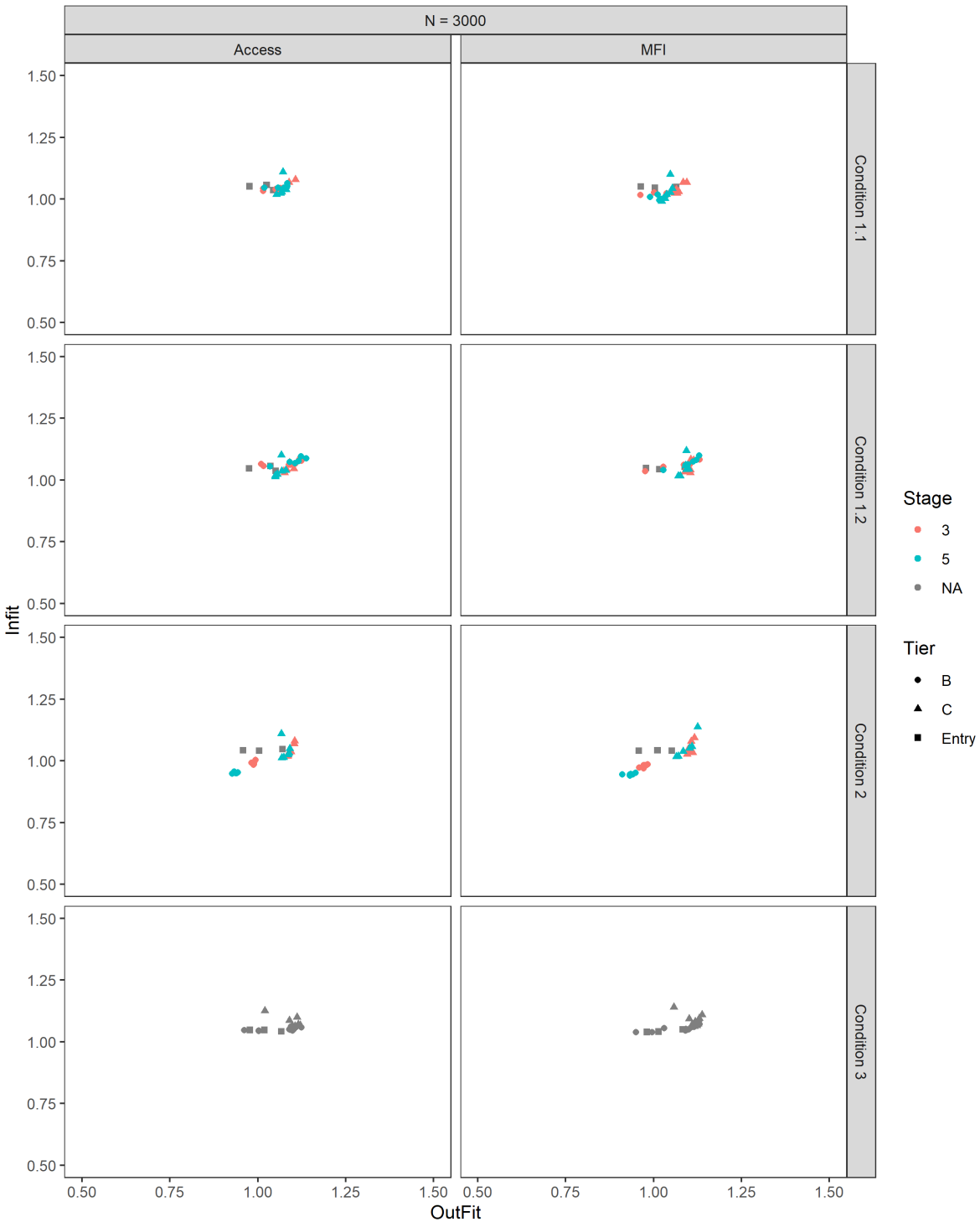

Figure 14. Listening: Infit and Outfit statistics for Cluster 912

Notes

[1]As described in FunctionalRules503ONLINE_FINAL 4.20.21.docx

[2]As described in S601 Functional Rules for Field Testing for Online ACCESS for ELLs_APPROVED_12_22_2021.docx

[3]If $\hat{\theta}_{MLE}$ is lower than Tier A cut score at Stage 6 or Stage 8, examinees are not assigned one of the modules at the last two stages of Listening or Reading panels, respectively.

[4]Although examinees' responses to each FT items were simulated in Step 7, only those responses actually administered were included in the analysis. The entire response generation was primarily used to have the same response pattern under different study conditions to avoid estimation errors attributable to the various response patterns under different conditions.

[5]In the operational setting, FT folders are administered to examinees randomly either before or after the targeted stage to control context effect.

References

Cai, L. S. (2018). *An investigation of item calibration approaches in multistage testing* (Unpublished doctoral dissertation). The University of Nebraska-Lincoln, Lincoln, NE.

de Ayala, R. J. (2013). *The theory and practice of item response theory*. The Guilford Press, New York.

Linacre, J. M., & Wright, B. D. (2000). Winsteps. *URL: http://www. winsteps. com/index. htm [accessed 2013-06-27][WebCite Cache]*.

Magis, D., Yan, D., & von Davier, A. A. (2018). mstR: Procedures to generate patterns under multistage testing [Computer software manual]. Retrieved 2021-11-03, from `https://github.com/cran/mstR`

Nicklin, C., & Vitta, J. P. (2022). Assessing rasch measurement estimation methods across R packages with yes/no vocabulary test data. *Language Testing*, *39*(4), 513–540.

Robitzsch, A., Kiefer, T., Wu, M., Robitzsch, M. A., Adams, W., Rcpp, L., & LSAmitR, R. E. (2022). Package 'TAM'. *Test Analysis Modules–Version*, 3–5.