Using the Partial Credit Model and Rasch Model to Examine the FOCIS Survey

Xin Xia, University of Virginia, United States
Robert H. Tai, University of Virginia, United States

The IAFOR International Conference on Education in Hawaii 2024
Official Conference Proceedings

Abstract
This study examined the dimensionality and effectiveness of the five categories Likert Scale of the framework for observing and categorizing instructional strategies (FOCIS), a survey that measures students' preference for learning activities in science instructions, developed by Tai et al. in 2012. The data included 6546 students from 3rd to 12th grade including 4 school districts. The results show that the FOCIS survey has 7 dimensions measuring students' preferences. This study only tests the effectiveness of the *Competing* dimension. Compared to the Partial Credit Model (PCM) model and Rasch model, condensing down the categories to dichotomous items fits the data better. The AIC and BIC decreased, and the infit outfit improved on the Rasch model.

Keywords: Item Response Theory, Partial Credit Model, Validity

Introduction

A framework for observing and categorizing instructional strategies (FOCIS) and an instrument to measure students learning activities preferences in science learning was developed by Tai et al. (2021) in 2012 (Figure 1). Educators can understand the types of activities students enjoy in science by using the FOCIS survey. However, this survey has not been tested through the item response theory method yet. The aim of this study is to examine the measurement dimensionality and effectiveness of the five categories Likert Scale. The original students used the confirmatory factor analysis to examine latent variables. In this study, the main focus is on the effectiveness of response categories.

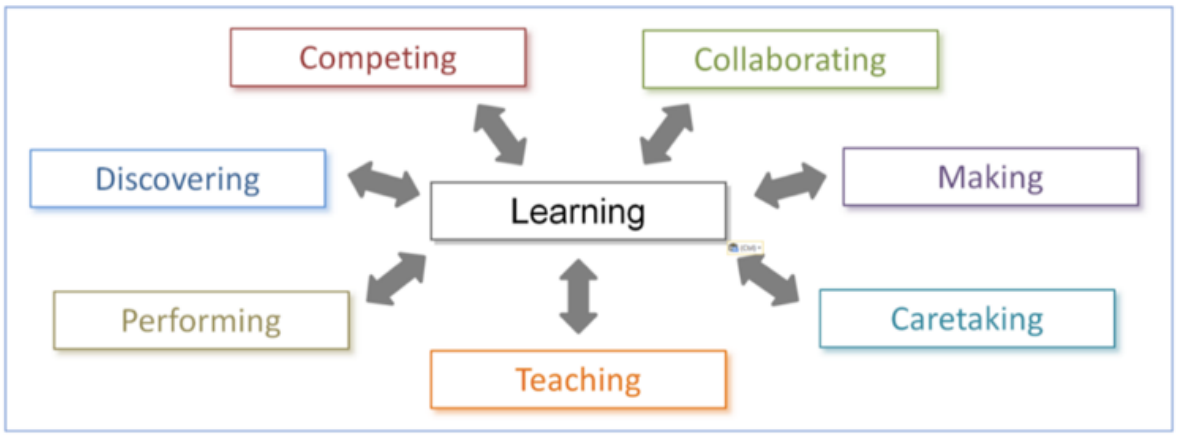

Figure 1. *Framework for Observing and Categorizing Instructional Strategies (FOCIS)*

Perspective/FOCIS Framework

A sufficient among of studies demonstrate that students' interest and attitude toward science are two effective indicators of students' future career expectations and participation in science, technology, engineering, and mathematics (STEM) (Koballa, Jr. & Glynn, 2007; Lent et al., 1997; Luce & His, 2014; Luce & Woodman, 2014; Simpson et al., 1994). Science-focused activities can spark students learning interests in science. To better understand students' science learning activities preferences, Tai et al. developed a conceptual framework for observing and categorizing instructional strategies (FOCIS) and deconstructed activities into seven categories: (a) *Collaborating*, (b) *Competing*, (c) *Making*, (d) *Discovering*, (e) *Performing*, (f) *Caretaking*, and (g) *Teaching/tutoring* (2012).

Based on the definition of the authors, collaborating happens when group members work together on a project or task. This category contains 4 indicators to measure. *Competing* activities were defined as compelling participants to seek to win, which have four indicators. Making activities is the process of constructing an object by applying ideas and materials (four indicators). Discovering activities contain five indicators that measure the performance of participants in learning new things, figuring things out, and problem-solving. Performing is the activities associated with presentations and audiences as outcomes at a specific place and time, meeting a challenge (four indicators). Caretaking (three indicators) is caring for others, animals, and even objects. Last but not least, teaching is helping others to learn (four indicators).

Purpose and Research Questions

The purpose of this research was to conduct a dimensionality analysis to test how many dimensions the FOCIS has. Through running a Partial Credit Model analysis of the dimension of *Competing*, this study enquires whether the Likert-Scale can effectively capture information. Using the FOCIS survey data from 2012, this project will answer the following questions:

1. Does the FOCIS survey measure multiple latent traits (multidimensional)?
2. Is a 5-point Likert Scale effective in measuring the individuals in 4 items of *Competing*?

Design/Procedure

Sample

The FOCIS survey instrument was given to students in the 3rd through 12th grades in 2012. A total of 7382 students took the survey. After removing all missing values, the sample size is 6546. Surveys were handed out across three different school districts in Virginia and two different school districts in Illinois. Twenty-five schools participated in data collection.

Measures

The FOCIS survey is the instrument to measure students' preferences about those types of activities. According to the framework, the survey contains seven dimensions with 28 items. Response categories were provided on a 5-point Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), which were recoded into: 1 = 0, 2 = 1, 3 = 2, 4 = 3, 5=4. Recoding was required for the intended PCM analysis. Focus on the *Competing* dimension, the wording of the question, "I like to focus on my own goals, rather than competing with others", was reversed. Therefore, before analysis, this item was recoded as 5=0, 4=1, 3=2, 4=3, 5=4. The target variables are listed below (Full items attached in Appendix B):

feelcmpt: I like an activity that involves "Being in a competition".
exctcmpt: I get excited when I hear there will be a competition.
mptothr: I enjoy competing against other people.
focusown: I like to focus on my own goals, rather than competing with others.

For demography information, sex was coded 1 for male and 2 for female. Race was coded as white, Latino, black, Asian, and multi-race.

Statistical Analyses

Using RStudio eRm package and ltm package, the focus of this analysis is to re-evaluate the effectiveness of the FOCIS survey in the Item Response Theory lens. Two Rstudio packages were used in this study to conduct the factor analysis (CFA), ltm, and Psych.

Factor analysis considers the possibility that items are related to one or more common factors and treated as parallel observations. The assessment of factor analytic fit tests is based on the prediction of item-covariance uniformity in the variance-covariance matrix (van der Lans et al., 2021). To answer the first research question factor analysis was conducted on the full data set to test the dimensionality of the survey.

Considering this instrument using a five-point Likert scale (strongly disagree, disagree, neutral, agree, and strongly agree), this project would use the partial credit model (PCM), a Rasch-based model, to analyze the polytomous categories response data. Embretson and Reise (2000) illustrated that PCM is perfectly appropriate for analyzing attitude responses on a multi-point scale. The IRT model is able to place the personal traits and item difficulty on the same scale and able to exam whether items in the survey vary in difficulty, compared to the sum scores method used in this article. This study mainly focused on one of the dimension analyses: *Competing*. The model equation is below:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta - \delta ij)\right]}{\sum_{r=0}^{mj}\left[\exp\sum_{j=0}^{x}(\theta - \delta ij)\right]}$$

Where $\sum_{j=0}^{0}(\theta - \delta ij) \equiv 0$

The δ_{ij} ($j = 1, \ldots, m_i$) term is the item step difficulty associated with a category score of j; a higher value of a particular δ_{ij} means the more difficult a particular step is relative to other steps within an item This equation can be interpreted as the probability of an examinee responding in category x on an m_i step item, which is a function of the difference between an examinee's trait level and a category intersection parameter. This study mainly focused on one of the dimension analyses: Competing. (Embretson & Reise, 2000).

Analyses and Findings

Descriptive Statistics

The sample consists of 6546 students in 25 schools. Descriptive statistics for the sample are provided in ***Table 1***. 51.12% of students are male and 48.88% of students are female. The race and ethnic information are as below: 0.57% American Indian; 1.79% Asian; 16.18% Black; 16.98% Hispanic; 0.1% Pacific Islander; 15.16% Multiple races and ethnic group students; and 49.22% White students. As ***Table 2 and Figure 2*** show, category 5 is the most endorsed one in feelcmpt, exctcmpt, cmptothr, while the first category is the most endorsed one in the focusown item.

The frequencies of participants' total scores reflected the distribution of total scores (***Table 3, Figure 3***). Preliminary analysis of the parametric assumptions revealed that *Competing* total scores were normally distributed. The skewness of total scores is -0.497448 (|skewness values| < 1.0). the highest frequency of total raw scores is 13.

Variables	Percent %
Gender	
Male	51.12
Female	48.88
Race	
AmInd	0.57
Asian	1.79
Black	16.18
Hispanic	16.98
Pacific Islander	0.1
Multi-racial	15.16
White	49.22

Table 1. *Descriptive Statistics for Student Gender, Student Race (N =6546)*

	Thresholds					Item Total Score	count
	0	1	2	3	4		
feelcmpt	0.1189	0.102	0.2605	0.214	0.3046	16257	6546
exctcmpt	0.1801	0.1257	0.2178	0.1819	0.2944	14956	6546
cmptothr	0.1743	0.1166	0.1911	0.1856	0.3324	15614	6546
focusown	0.3291	0.13	0.1146	0.2324	0.194	11994	6546

Table 2. *Proportions for Each Level of Response and Item Total Scores (N =6546)*

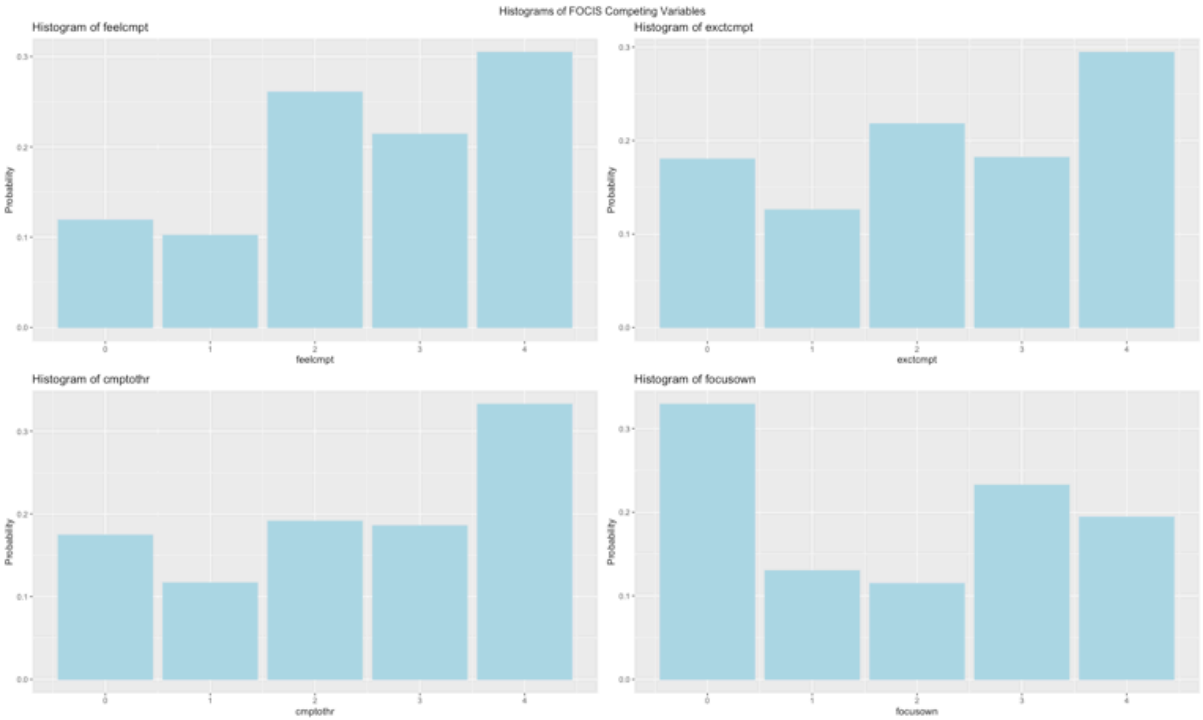

Figure 2. *Histogram of FOCIS Competing variables*

Total Score	Frequency
0	329
1	184
2	194
3	223
4	319
5	237
6	344
7	385
8	411
9	497
10	556
11	514
12	728
13	731
14	410
15	338
16	146

Table 3. *Frequencies of Participant Total Score (N =6546)*

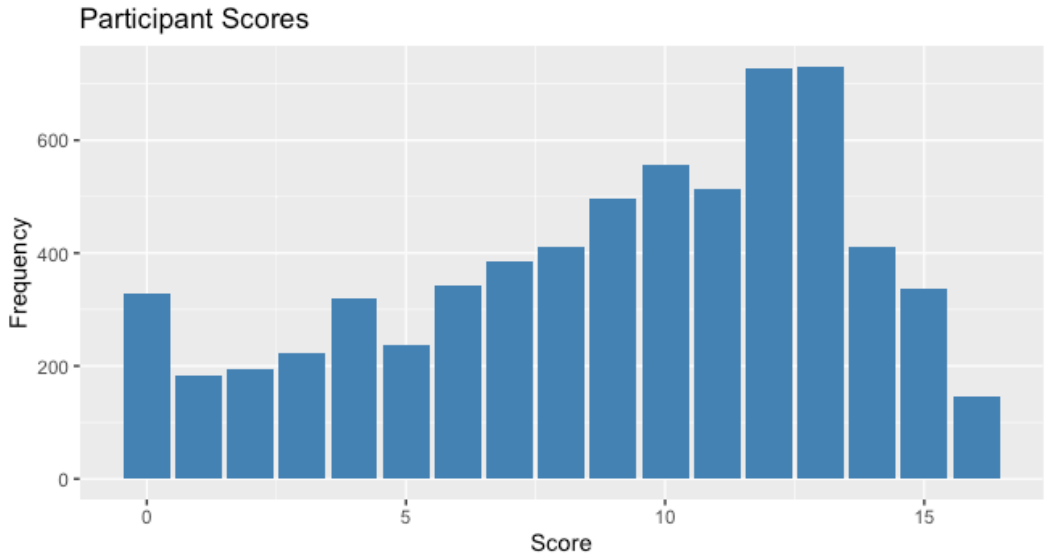

Figure 3. *Histogram of FOCIS Competing personal raw total scores*

Dimensionality Analysis

The screen graph (***Figure 4***) shows that this scale has seven dimensions. The eigenvalues did not drop much after the seven factors and the eigenvalue of the seventh factor was larger than 1. What's more, using polychoric correlations, the pattern of items loading on each factor showed feelcmpt, exctcmpt, cmptother loading on the same factor. The loading of focusown is relevantly lower than other items (Table 4). The Cronbach's alpha is 0.7306, indicating the items have acceptable internal consistency.

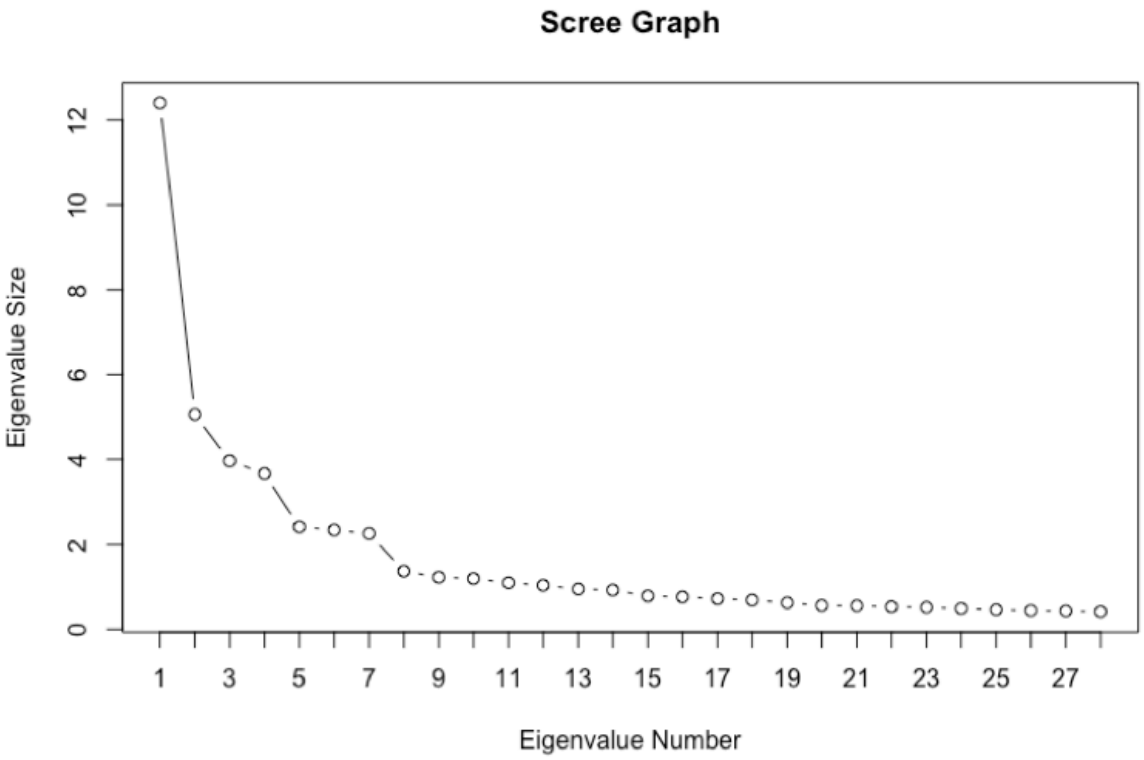

Figure 4. *Scree Graph*

Competing Items	Factor loading						
	1	2	3	4	5	6	7
feelcmpt I like an activity that involves "Being in a competition"	-0.02	0.85	0	-0.01	-0.02	0.01	0.01
exctcmpt I get excited when I hear there will be a competition.	-0.01	0.88	0	0.03	0.02	0.01	0.02
cmptothr I enjoy competing against other people.	0.02	0.89	0	-0.01	-0.01	-0.01	-0.02
focusown I like to focus on my own goals, rather than competing with others.	-0.17	0.14	0.01	0.02	-0.09	-0.06	0.01

Table 4. *Factor loadings*

Infit and Outfit of Competing

As the infit outfit table shows all items in the *Competing* dimension are misfitting. The item feelcmpt, exctcmpt, cmptothr are overfitting. The infit and outfit are smaller than 0.6, which is considered overfitting by liberal standards. The model might absorb noise or random variation in the data rather than the underlying pattern. The infit and outfit of focusown item are larger than 1.4, which indicates this item is underfitting. The model might have too few parameters and fails to capture the underlying pattern (***Table 5***).

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
feelcmpt	3431.256	6070	1	0.565	0.582	-27.432	-28.934	0.786
exctcmpt	3106.681	6070	1	0.512	0.528	-29.464	-34.111	0.836
cmptothr	3025.23	6070	1	0.498	0.528	-28.667	-33.289	0.834
focusown	11775.39	6070	0	1.94	1.764	30.305	37.599	-0.076

Table 5. *Infit and Outfit Outcomes*

Item Difficulty and Thresholds

On average, the mean of the item difficulty is 0. The most difficult item is focusown (beta=1.22). Feelcmpt is the easiest item (beta=-0.946). How all thresholds of the four items are reversed. For instance, the first threshold of item one is -0.575 while the second threshold is -1.489 smaller than threshold one (***Table 6***).

	Location	Threshold 1	Threshold 2	Threshold 3	Threshold 4
feelcmpt	-0.946	-0.575	-1.489	-0.954	-0.766
exctcmpt	-0.0895	-0.092	-0.504	0.074	0.164
cmptothr	-0.18525	-0.089	-0.478	-0.073	-0.101
focusown	1.22075	0.802	1.227	1.024	1.83

Table 6. *Item difficulty and Item Thresholds*

Person Traits and Thetas

The mean person trait (mean=0.23) is higher than the mean item difficulty (mean=0), which suggests that, on average, the individuals in your sample have a higher trait level than the average difficulty of the test items. In other words, the test items might be relatively easier for the respondents in your sample. Examples of person traits are listed in ***Table 7***.

However, it's essential to consider the distribution of item difficulties and person traits, as individual items might still be challenging for some respondents, and some items might be easier for others. Analyzing the distribution and variance of both person traits and item difficulties can provide more insights into the appropriateness of the test items for the sample (***Figure 5***).

In conclusion, the test items may not be optimally targeted for the sample since the average ability level of the individuals is higher than the average difficulty of the items. The test items might be relatively easier for the respondents in this sample. It might be helpful to include more challenging items in the test to better discriminate between individuals with higher ability levels.

Estimate	Standard Deviation	Error	2.50%	97.50%
theta 1	0.96855899	0.5151915	-0.0411977	1.9783157
theta 2	0.96855899	0.5151915	-0.0411977	1.9783157
theta 3	-0.5990954	0.469392	-1.5190868	0.32089609
theta 4	0.73726666	0.4521136	-0.1488597	1.62339302
theta 5	-0.3977457	0.430668	-1.2418395	0.4463481
theta 6	0.73726666	0.4521136	-0.1488597	1.62339302
theta 7	0.96855899	0.5151915	-0.0411977	1.9783157
theta 8	1.29245982	0.635236	0.04742024	2.53749941

Table 7. *Person Traits*

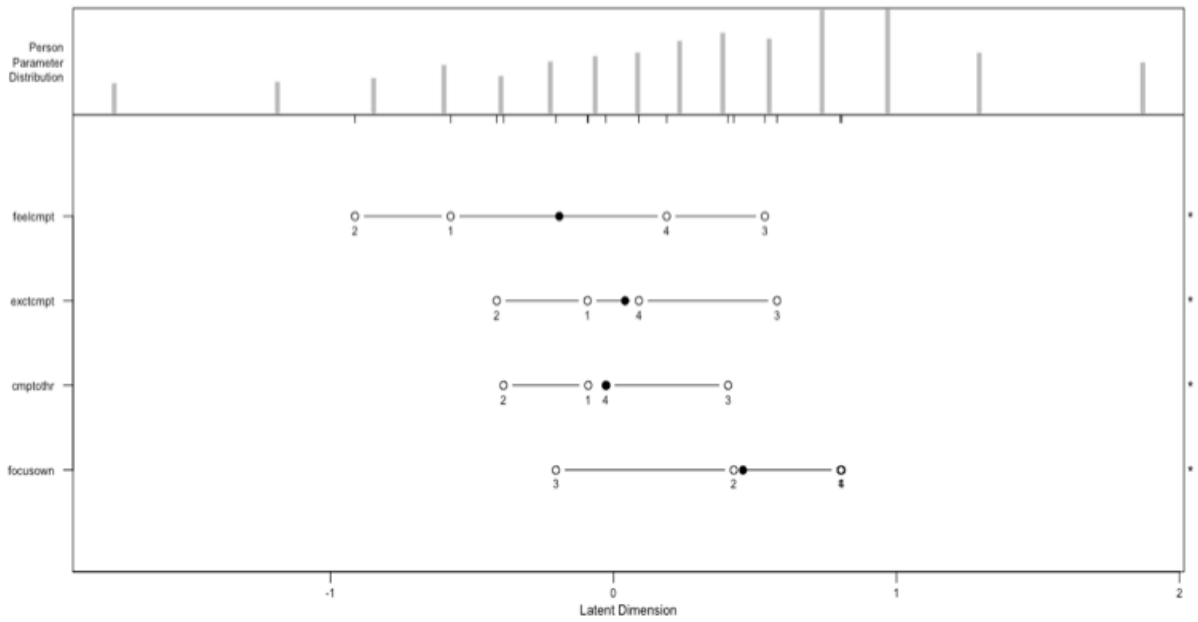

Figure 5. *Person Item Map*

Category Response Curves (CRCs)

As ***Figure 6*** shows, the second category and fourth category are never the most likely categories among all four items. People are not picking the second and fourth item. The third category of feelcmpt, exctcmpt, cmptothr items has a short peak response window while the focusown item does not contribute to this item. Results indicate that the four items might not need 5 categories to capture information. It suggests that maybe the dichotomous category functions better. Since all items only have three or two major points, the results suggest that maybe consider condensing five categories down to dichotomous categories.

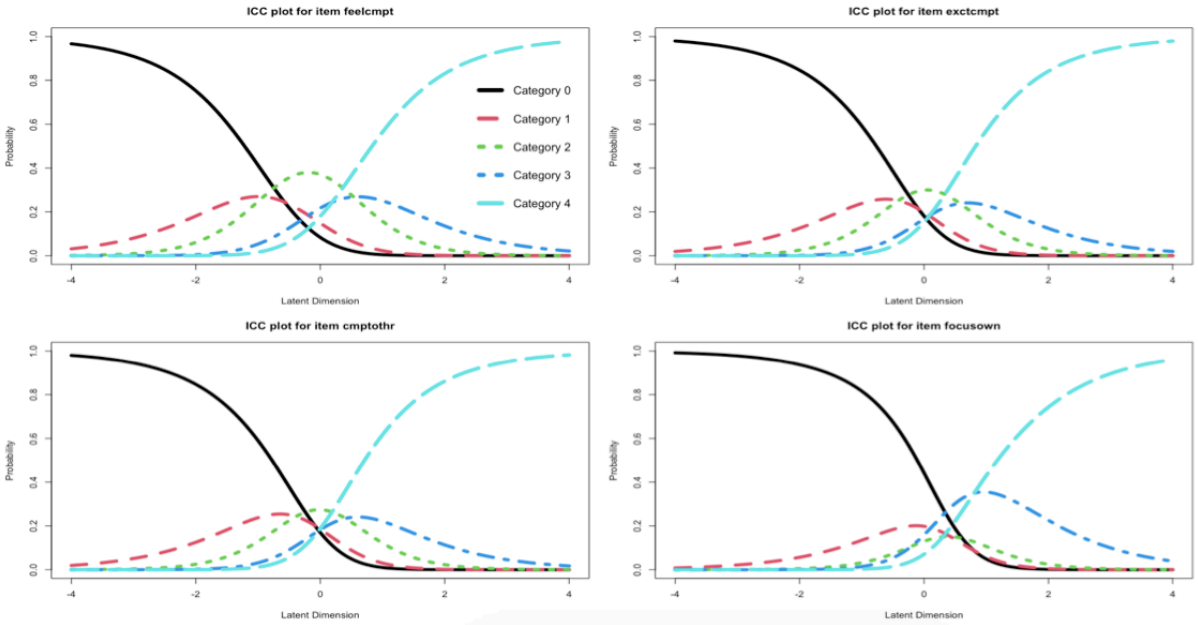

Figure 6. *Category Response Curves*

Item Information

For the first three items (***Figure 7***), the most information this survey got was from items located on zero. The item focusown provided the highest amount of information compared to the other three items. However, most information about the scale is slightly larger than zero. Similarly, on the total information (***Figure 8***), the most information about the scale is slightly bigger than zero.

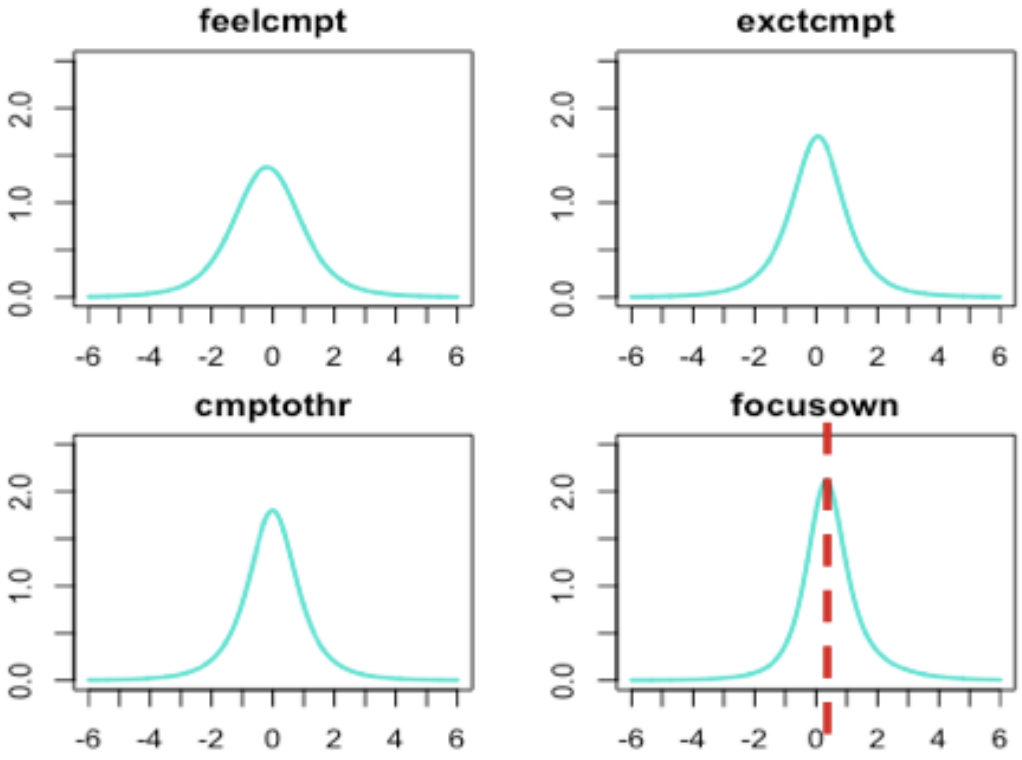

Figure 7. *Item Information*

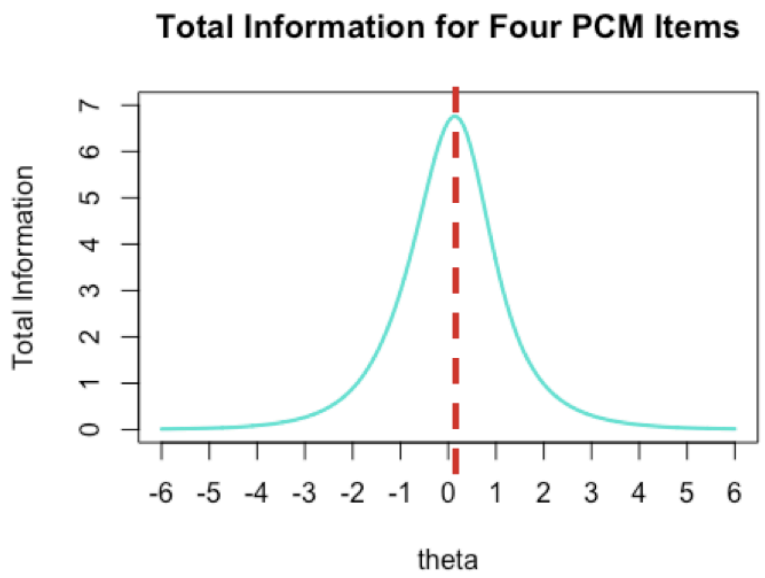

Figure 8. *Item Total Information*

Condense Categories

Based on the results above, this study condensed the categories down to dichotomous categories. According to the frequency table of items endorsement and density plot, we combine categories one to three as one category, coded as 0. Categories four and five combine into one category coded as 1 (as ***Figure 9***). The condensed data was analyzed by Rasch Model through eRm package. To compare the model fit, this study looked at infit and outfit as well as AIC and BIC.

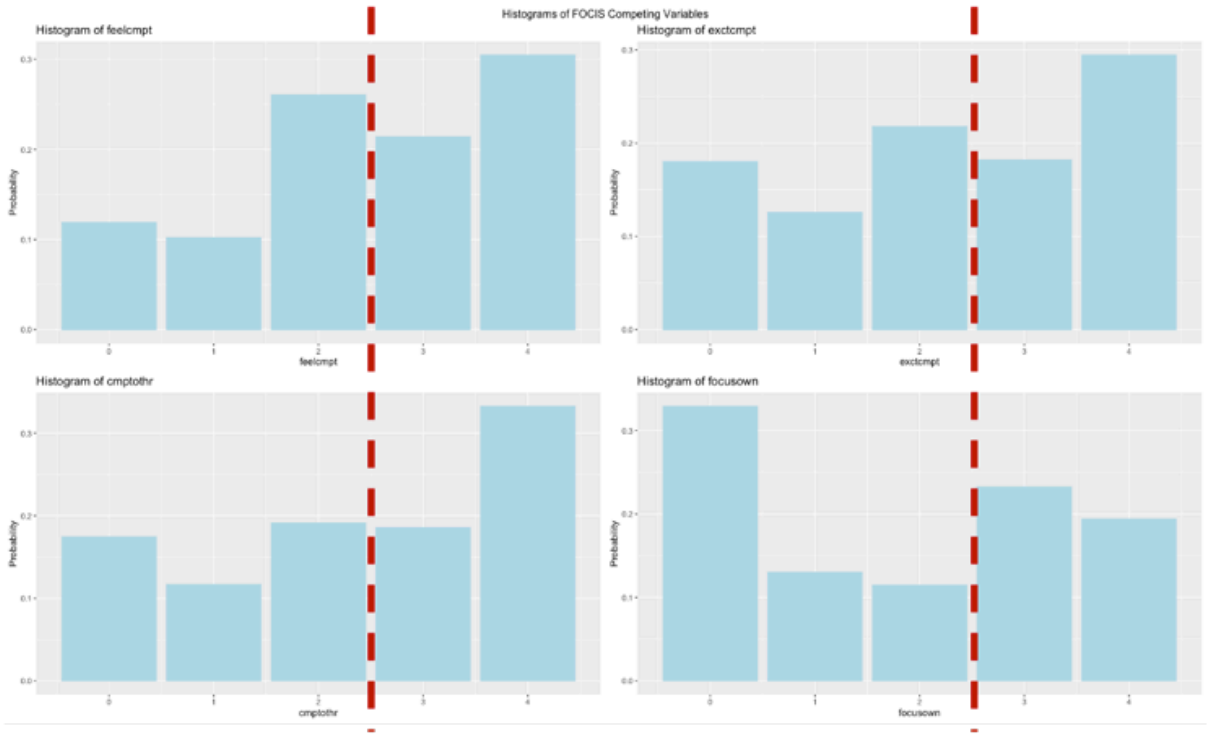

Figure 9. *Condense Categories*

Model Comparison

The infit and outfit of feelcmpt, exctcmpt, cmptothr all increased. However, the infit and outfit of focusown item underfitting worse than the PCM model (***Table 8***).
AIC and BIC described that the Rasch Model fit better. Both values of AIC and BIC decreased dramatically (***Table 9***).

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
PCM								
feelcmpt	3431.256	6070	1	0.565	0.582	-27.432	-28.934	0.786
exctcmpt	3106.681	6070	1	0.512	0.528	-29.464	-34.111	0.836
cmptothr	3025.23	6070	1	0.498	0.528	-28.667	-33.289	0.834
focusown	11775.393	6070	0	1.94	1.764	30.305	37.599	-0.076
Rasch								
feelcmpt	2090.032	3447	1	0.606	0.728	-15.139	-13.787	0.55
exctcmpt	2228.24	3447	1	0.646	0.674	-20.679	-20.399	0.836
cmptothr	2129.38	3447	1	0.618	0.654	-21.413	-21.097	0.808
focusown	6988.273	3447	0	2.027	1.585	33.435	37.791	-1.425

Table 8. *Infit and Outfit Comparing Table*

	value	npar	AIC	BIC	cAIC
PCM					
joint log-lik	-30181.51	30	60423.02	60624.36	60654.36
marginal log-lik	-37758.23	15	75546.46	75648.26	75663.26
conditional log-lik	-19834.34	15	39698.69	39800.48	39815.48
Rasch					
joint log-lik	-7224.279	6	14460.557	14497.431	14503.431
marginal log-lik	-13881.577	3	27769.155	27789.514	27792.514
conditional log-lik	-4356.108	3	8718.216	8738.576	8741.576

Table 9. *Model Fit*

Conclusion

Factor analysis presented the FOCIS survey did have 7 dimensions. The loading on focusown item is not aligned with other items well, The focusown item is the only item in the whole survey that has the reversed phrase question, which might explain the reason for the underfitting of this item. People might be confused about the phase or not pay attention to the pattern change in this item.

Based on the infit, outfit, and CRCs of PCM, the results showed five categories did not capture the information well for this population. Two or three categories of the four items of *Competing* are the never likely answer items with reversals. Compared to the condensed Rasch model, AIC and BIC suggested the model fit better. Additionally, the AIC and BIC of the Rasch model indicate the model fits better than PCM.

Contribution/Significance

The FOCIS instrument could be used to assess the enjoyment of different science activities with 3rd to 12th-grade students and provide information for youth STEM program developers and instructions for evaluating and modifying the program. This analysis indicated that a dichotomous design of *Competing* dimensions can better capture information. It is also worth conducting an analysis of other dimensions in the future.

References

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. L. Erlbaum Associates (pp.105).

Koballa, Jr., T. R., & Glynn, S. M. (2007). Attitudinal and motivational constructs in science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 75-102). Lawrence Erlbaum Associates.

Lent, R. W., Brown, S. D., & Gore Jr., P. A. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. *Journal of Counseling Psychology*, *44*, 307–315. https://doi.org/10.1037/0022-0167.44.3.307

Luce, T. J., & Woodman, A. J. (2014). *Tacitus and the Tacitean Tradition*. Princeton University Press.

Tai, R. H., Ryoo, J. H., Mitchell, C. E., Kong, X., Skeeles-Worley, A., Almarode, J. T., Maltese, A. V., & Dabney, K. P. (2021). Gauging Informal STEM Youth Program Impact: A Conceptual Framework and a Measurement Instrument. *Journal of Youth Development*, *16*(4), 103–133. https://doi.org/10.5195/jyd.2021.981

van der Lans, R. M., Maulana, R., Helms-Lorenz, M., Fernández-García, C. M., Chun, S., de Jager, T., ... & Moorer, P. (2021). Student perceptions of teaching quality in five countries: A Partial Credit Model approach to assess measurement invariance. *SAGE Open*, *11*(3), 21582440211040121.

Contact email: bj5jy@virginia.edu