Using Rasch Modeling to Validate AI Scoring Models

Steven McGee
Yaurie Hwang
Willow Kelleigh

The Learning Partnership

Abstract
This study explores the development and validation of AI scoring models for assessments related to the *Exploring Computer Science* curriculum. Using BERT, AI models were trained to replicate human scorers and validated using Rasch modeling. Results showed that the AI models approximated human scorers, offering opportunities for timely, formative feedback.

Acknowledgment
The authors were supported in part by National Science Foundation grants CNS-2219491, CNS-2122907, CNS-1738776, and CNS-1738572. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

Suggested Citation
McGee, S., Hwang, Y., & Kelleigh, W. (2025, April 26). Using Rasch Modeling to Validate AI Scoring Models [Poster]. National Council for Measurement in Education. Denver, CO. https://doi.org/10.51420/Conf.2025.4

1 Introduction

A key strategy for broadening computer science participation in the Chicago Public Schools (CPS) has been the enactment of a high school computer science graduation requirement (Dettori, et al., 2018). Since 2020, roughly 18,000 students graduate from CPS each year having taken at least one full year of computer science. *Exploring Computer Science* (ECS) is the primary course that students have been taking to fulfill the graduation requirement. The ECS curriculum is composed of activities that engage students in computer science inquiry around meaningful projects (Margolis et al., 2012). The pedagogy of ECS is structured around three interwoven strands: equity, inquiry, and computer science concepts. The accompanying ECS professional development program prepares teachers to implement these inquiry-based activities. It emphasizes building a classroom culture that is culturally responsive and adapting lessons to the backgrounds and interests of the students (Goode et al., 2014).

The Chicago Alliance For Equity in Computer Science (CAFÉCS) was formalized as an RPP in 2017 to ensure that CPS provides schools with sufficient support and accountability so that the *quality* of computer science course experiences is equitable across the district (Henrick et al., 2024). CAFÉCS is a research-practice partnership among the Chicago Public Schools (CPS), DePaul University, Loyola University, The Learning Partnership, and the University of Illinois Chicago. To ensure that all students in Chicago participate in engaging, relevant, and rigorous computing experiences, CAFÉCS addresses problems of practice through research and development that increases opportunities for all students to pursue computing pathways and prepares all students for the future of work. CAFÉCS has contributed to the literature on the positive impact of the ECS course on students' attitudes towards computer science (McGee, et al., 2017; McGee, McGee-Tekula, Duck, McGee, et al., 2018) as well as students' choices about future high school computer science coursework (McGee et. al., 2017, McGee, McGee-Tekula, Duck, Dettori, et al., 2018) and college computer science majors (Jeffrey, 2024). In addition, the results in Chicago have shown that students are achieving equivalent posttest performance of computational thinking in ECS across all races, ethnicities, and genders (McGee et al., 2019; McGee, McGee-Tekula, Duck, McGee, et al., 2018). ECS serves as a foundation for successful student performance in other coursework, including AP CS A (Boda & McGee, 2021) and high school science (Bruno & Polikoff, 2023; Huang, 2024).

To measure student learning, CPS adopted SRI-developed pretest and posttest assessments to measure computational thinking practices aligned with the ECS curriculum (Snow et al., 2017). The assessments include 30 individually scored constructed response and multiple-choice prompts across both pretests and posttests. SRI developed scoring rubrics for all the constructed response prompts with descriptors and example responses for each score level. While the assessments have been beneficial for summative research purposes, their use as formative assessments have been hampered by the time- and labor-intensive efforts involved in scoring thousands of assessments each year.

In this research, we sought to explore the development and validation of AI scoring models to replace the human scorers, thus creating the opportunity to use the assessments for formative purposes. Our primary research question was: Can we develop AI scoring models that approximate the scoring behaviors of human scorers?

2 Methods

This section describes the process of human scoring of the assessments, the development of AI scoring models, and the use of Rasch modeling to validate the AI model for both the pretest and posttest.

2.1 Assessment of Computational Thinking practices

To measure the development of computational thinking practices, we used assessments that were aligned to the computational thinking practices in ECS (Snow et al., 2017). The assessments were developed and field tested by SRI International over two years using Evidence-Centered Design (ECD) (Mislevy & Haertel, 2006), an assessment methodology that is especially advantageous when the knowledge and skills to be measured involve complex, multistep performances. The ECD process involved (1) working with various stakeholders to identify the important computer science skills to measure, (2) mapping those skills to a model of evidence that can support inferences about those skills, and (3) developing tasks that elicit that evidence (Bienkowski et al., 2015). The assessments were field tested with 941 students over two years (Snow et al., 2017).

Separate pretest and posttest forms were created. The pretest contains six tasks that measure students' initial understanding of CS concepts and computational thinking practices. For example, in one task students are asked to develop an algorithm to assign students to after school clubs that maximizes students' preferred choice while keeping the enrollment within the limits for each club. There are a series of subtasks that ask students about specific aspects of the problem and the solution. Across the six tasks there are a total of 19 subtasks that are scored independently. The posttest also contained six tasks, two of which were on the pretest and four of which were different. The two common tasks were used to equate the two forms and allow for measurement of growth from pretest to posttest. SRI developed scoring rubrics with student work examples for each of the tasks. Across all the pretest and posttest tasks, there are a total of 30 question prompts that are each scored individually.

There were three sources of evidence for establishing the validity of the assessments at measuring the computational thinking practices covered in the ECS curriculum. (a) Content validity was established through an expert review of the alignment between the knowledge and skills, the curriculum learning goals, and computational thinking practices. (b) Cognitive think-aloud interviews were conducted with a subset of students participating in the pilot test of the assessments. (c) The reliability of the assessments was moderate to high. The tasks within each assessment are well aligned with each other and with the targeted learning goals. See Snow et al. (2017) for additional details on the pilot study and validation results.

From the 2016-17 school year through fall 2024, teachers administered the SRI-developed ECS pretest at the beginning of the year and the posttest at the end of the year. Over that time span, we collected a total of 32,720 pretests and 10,091 posttests. We hired The Graide Network for scoring the pretests and posttests. Over that time period, The Graide Network recruited and trained a total 76 undergraduate preservice teachers and computer science undergraduate majors to score the performance tasks. The scorers were provided training on each of the rubrics prior to scoring. We used the *Facets* software to conduct Many-Facet Rasch Measurement analysis (MFRM) (Eckes, 2011) to calibrate the scorers and scale the student responses. *Facets* develops a model based on how well the student performed across the range of question prompts with set difficulties considering the severity of the scorer relative to the other scorers. Within MFRM, the goal is not for scorers to arrive at agreement on the scores, but instead to model the variation in how the scorers interpreted the rubrics. If the raters are internally consistent in how they apply the rubric, *Facets* can adjust the students' scores based on the severity or leniency of the scores relative to other scorers. As part of the calibration process, we had overlapping subsets of scorers rate the same students until there were no subsets, and the raters had high reliability as measured by infit and outfit statistics ranging between 0.5 and 1.5. All scorers achieved acceptable fit statistics.

2.2 AI Model Development

We used BERT to develop the AI scoring models. BERT is an open-source machine learning framework for natural language processing. It provides a general-purpose language model that can then be fine-tuned on smaller task-specific datasets, like the ECS assessments. We randomly split the student responses for each questions into a training set (40%), validation set (20%), and a testing set (40%). The text responses were preprocessed by converting the text into machine-readable format. In the training

process, we used a hyperband to search through the possible combinations of model layers that optimizes the results. The BERT model uses the numerical characteristics of the student texts to identify the set of numerical parameters that best predict the human scores. It is important to point out that the human scorers varied in the severity of their application of the rubrics and that the resulting models were optimized across the range of severity of the scorers. After developing the model with the training dataset, it was fine-tuned for generalizability using the validation set. The resulting model was applied to the testing data to determine the accuracy of the model predictions to the human scores. Tables 1 and 2 provide the accuracies of each model in comparison to the human scorers. In pretest task 3, students are asked to develop an argument on whether an alarm clock is a computer or not. They could develop their argument by selecting from a list of the four traits of a computer and justifying why an alarm does or does not have that trait. Students could earn full credit by taking either position and providing a solid justification for that position. All but two of the models achieved 75% accuracy or higher. The two subtasks in posttest task 2 achieved low accuracy. Posttest task 2a asks students to develop a set of web site requirements for building a class web site and posttest task 2b asks students to identify and correct a web site bug. The final models were applied to the whole dataset of digital pretest responses for each subtask. The resulting scores were added to the dataset of scores from the 76 human scorers and the MFRM analysis was conducted. For the purposes of validation, we developed on posttest MRFM with posttest tasks 2a and 2b and one MFRM that excluded those two items.

Task	Sub Task	% Accuracy
Pre 1		93
Pre 3	Not a computer	75
	Process Information	82
	Takes Input	92
	Stores Information	88
	Produces Output	90
Pre 4	a	82
	c	82
	e	94
Pre 5	e	86
Pre 6	a	91
	b	75
	c	95
	d	95

Table 1: Percent accuracy of the pretest AI models in comparison to human scorers.

Task	Sub Task	% Accuracy
Post 1	a	87
	b	89
	c	91
	d	92
Post 2	a	34
	b	17
Post 3	a	96
	b	88
	c	85
	d	94
	e	98
	f	97

Table 2: Percent accuracies of the posttest AI model in comparison to human scorers.

3 Findings

Figure 1 provides a visual representation of the distributions of student ability, item difficulty, and rater severity on the pretest. The first column is the logit metric. The student column contains the estimates of the student abilities, with higher values indicating more ability. Higher logit values for items and raters indicate more difficult items and more severe raters, respectively. Visually we see that the students are spread out in terms of their ability estimates on the ECS pretest (Person reliability of .84). The item labeled Pre3b is the easiest item to receive high ratings on, whereas the item labeled Pre2 (multiple choice) is the hardest. The rater ID shows that raters differ in their severity levels (chi-square of 7774.7 (51), $p < .001$) and hence statistical methods to consider this variability are warranted. The rater ID labeled AI represents the AI model. The overall severity index of the AI was slightly higher than most

of the human scorers, but not the highest severity index. However, the MNSQ infit statistic was 0.96 and outfit statistic was 1.07, which were well within the acceptable range between 0.5 and 1.5.

Figure 2 provides a visual representation of the distributions of student ability, item difficulty, and rater severity on the posttest. The results for inclusion and exclusion of posttest tasks 2a and 2b were similar. We report here the results including posttests tasks 2a and 2b. Visually we see that the students are spread out in terms of their ability estimates on the ECS pretest (Person reliability of .80). The item labeled Post1a is the easiest item to receive high ratings on, whereas the item labeled Post3e is the hardest. The rater ID shows that raters differ in their severity levels (chi-square of 3038.0 (30), $p < .001$) and hence statistical methods to consider this variability are warranted. The rater ID labeled AI represents the AI model. As is the case for the pretest, the overall severity index of the AI models was slightly higher than most of the human scorers, but not the highest severity index. However, the MNSQ infit statistic was 1.14 and outfit statistic was 1.12, which are well within the acceptable range between 0.5 and 1.5.

4 Practical Implications

The findings suggest that the AI model behaves similarly to human scorers and thus can be used to estimate student scores in a timely manner. The pretest scores can be used by the teacher to identify students who will need extra support during the class to improve their overall outcomes on the posttest at the end of the year. The posttest scores can be used to reflect on pretest to posttest growth in computational thinking abilities.

5 References

Bienkowski, M., Snow, E., Rutstein, D., & Grover, S. (2015). Assessment Design Patterns for Computational Thinking Practices in Secondary Computer Science: A First Look. Technical Report. SRI International, Menlo Park, CA.

Boda, P. A. & McGee, S. (2021). Broadening participation and success in AP CS A: Predictive modeling from three years of data. In Proc. ACM Tech. Sym. Comp. Sci. Ed. (SIG CSE'21). ACM, Virtual Event, USA, 6 pages. https://doi.org/10.1145/3408877.3432421

Bruno, P. & Polikoff, M. (2023, July). The Effect of Computer Science Course-Taking on Science Outcomes in Chicago Public Schools [report]. Chicago, IL: The Learning Partnership. https://doi.org/10.51420/report.2023.1

Dettori, L., Greenberg, R. I., McGee, S., Reed, D., Wilkerson, B., & Yanek, D. (2018, February). CS as a graduation requirement: Catalyst for systemic change. Proceedings of SIGCSE '18, Baltimore, MD, USA, 406-407. https://doi.org/10.1145/3159450.3159646

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Peter Lang.

Goode, J., Margolis, J., & Chapman, G. (2014). Curriculum is not enough: The educational theory and research foundation of the exploring computer science professional development model. *In the Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 493-498. Retrieved from http://www.exploringcs.org/wp-content/uploads/2010/09/Curriculum-is-not-Enough.pdf

Henrick, E., Schmidt, D., McGee, S., Rasmussen, A. M., Dettori, L., Greenberg, R. I., Reed, D, & Yanek, D. (2024). Assessing the Impact of an RPP on a Large Urban School District: The Case of CAFÉCS. *Peabody Journal of Education*, *99*(3), 380–394. https://doi.org/10.1080/0161956X.2024.2357040

Huang, R. (2024, December). The Role of Skyline Curriculum Implementation in Moderating the Relationship Between ECS Course-Taking and High School Students' Science Performance in

Chicago Public Schools [Report]. Chicago, IL: The Learning Partnership. https://doi.org/10.51420/Report.2024.9

Jeffrey, W. (2024). A Computer Science Pipeline? Examining the Link between High School Coursework and Bachelor's Degree Attainment [report]. Chicago, IL: The Learning Partnership. https://doi.org/10.51420/Report.2024.8

Margolis, J., Ryoo, J. J., Sandoval, C. D. M., Lee, C., Goode, J., & Chapman, G. (2012). Beyond access: broadening participation in high school computer science. *ACM Inroads, 3*(4), 72-78. doi:10.1145/2381083.2381102

Mislevy, R.J. & Haertel, G.D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice* 25, 4, 6–20.

McGee, S., Greenberg, R. I., McGee-Tekula, R., Duck, J., Rasmussen, A. M., Dettori, L. & Reed, D. F. (2019, February). An Examination of the Correlation of Exploring Computer Science Course Performance and the Development of Programming Expertise. Proceedings of SIGCSE '19, Minneapolis, MN.

McGee, S., McGee-Tekula, R., Duck, J., Dettori, L., Greenberg, R.I., Reed, D.F., Wilkerson, B., Yanek, D., & Rasmussen, A.M. (2018, April). Does Exploring Computer Science Increase Computer Science Enrollment? Paper presented at the American Education Research Association annual meeting, New York.

McGee, S., McGee-Tekula, R., Duck, J., McGee, C., Dettori, L., Greenberg, R.I., Snow, E., Rutstein, D., Reed, D., Wilkerson, B., Yanek, D., Rasmussen, A., & Brylow, D. (2018, February). Equal Outcomes 4 All: A study of student learning in Exploring Computer Science. Proceedings of SIGCSE '18, Baltimore, MD, USA, 50-55. doi: 10.1145/3159450.3159529

McGee, S., McGee-Tekula, R., Duck, J., White, T., Greenberg, R. I., Dettori, L., Reed, D. F., Wilkerson, B., Yanek, D., Rasmussen, A.M., & Chapman, G. (2017). Does a Taste of Computing Increase Computer Science Enrollment? Computing in Science & Engineering, 19(3), 8-18.

Snow, E., Rutstein, D., Bienkowski, M., & Xu, Y. (2017). *Principled Assessment of Student Learning in High School Computer Science.* Paper presented at the International Conference on Computer Science Education Research, Tacoma, WA USA.

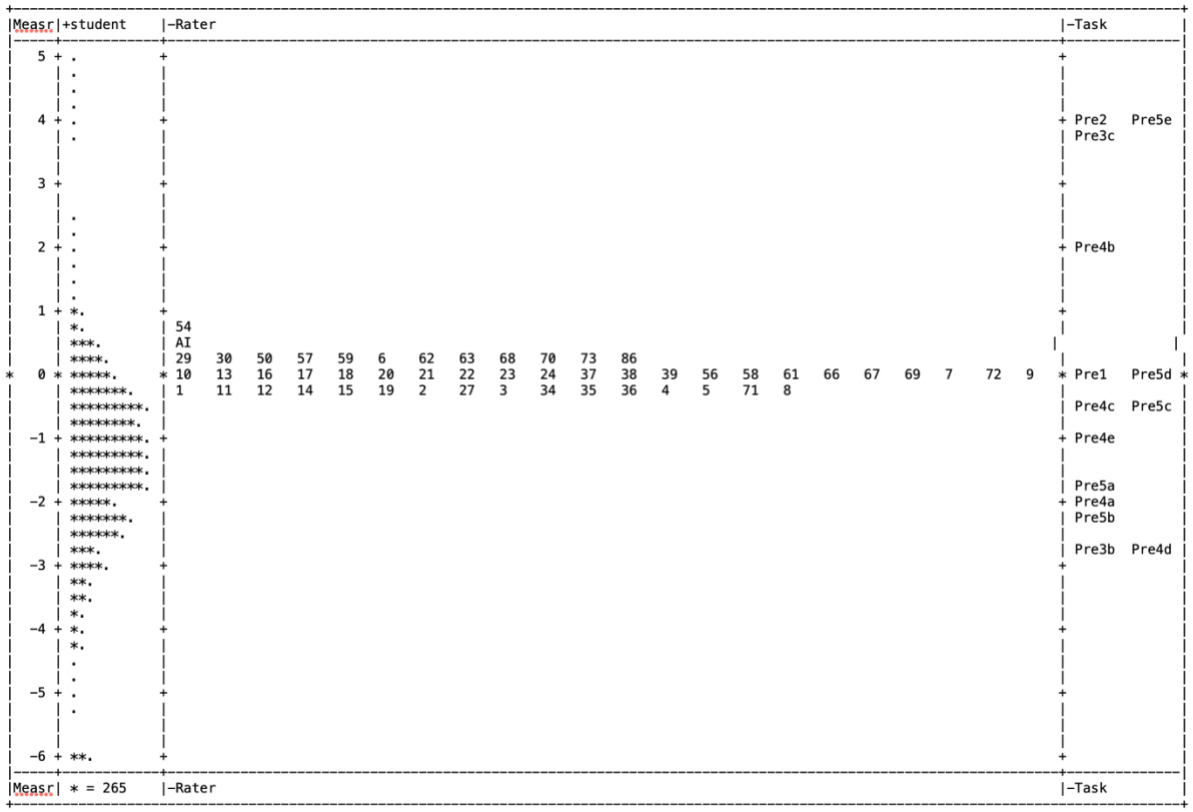

Figure 1: Wright Map of Rasch model output for the ECS pretest

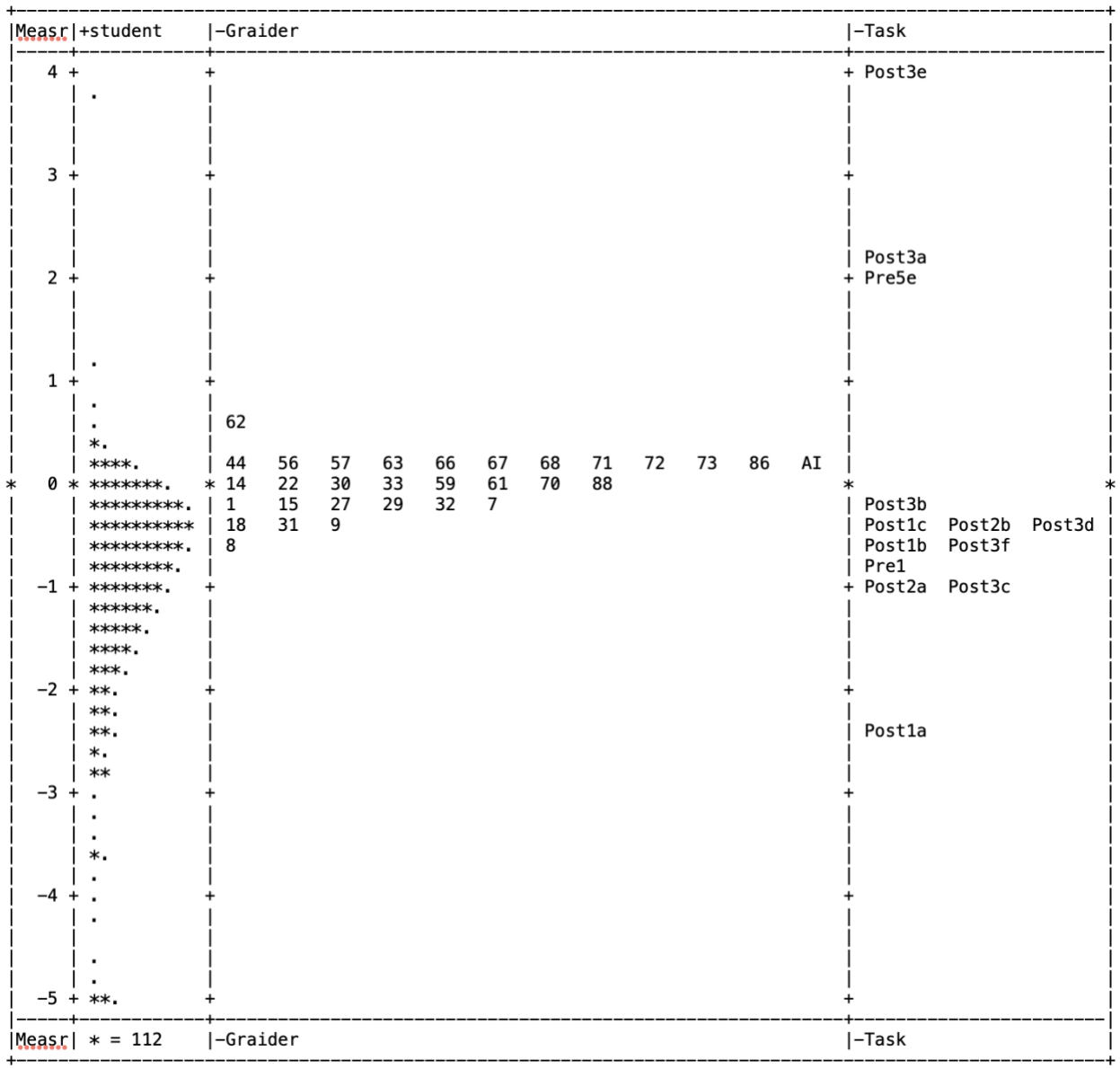

Figure 2: Wright Map of Rasch model output for the ECS posttest