Interdisciplinary Journal of Environmental and Science Education
2022, 18(4), e2297
ISSN 2633-6537 (Online)
https://www.ijese.com

Research Article

Bayesian Versus Frequentist Estimation for Item Response Theory Models of Interdisciplinary Science Assessment

Hyesun You [1*]

[1] University of Iowa, Iowa City, IA, USA
***Corresponding Author:** hyesun-you@uiowa.edu

Citation: You, H. (2022). Bayesian Versus Frequentist Estimation for Item Response Theory Models of Interdisciplinary Science Assessment. *Interdisciplinary Journal of Environmental and Science Education, 18*(4), e2297. https://doi.org/10.21601/ijese/12299

ARTICLE INFO

Received: 29 Mar. 2022

Accepted: 20 Jun. 2022

ABSTRACT

Along with the trend emphasizing ID learning, ID assessments to measure students' ID understanding have been developed by several scholars. The interdisciplinary science assessment for carbon cycling (ISACC) was developed to assess ID understanding among high school and college students in integrating knowledge from different science disciplines to explain a scientific phenomenon, global carbon cycling. The ISACC's construct validity was checked using traditional item response theory (IRT) models in 2021. The current study was motivated by the desire to reveal the difference in IRT analysis results of the ISACC using a Bayesian approach in comparison with the results using the traditional approach. The Bayesian approach has several strengths over the traditional IRT. The results of the study imply the need for additional research for the development and validation of interdisciplinary science assessments through strong psychometric properties.

Keywords: interdisciplinary understanding, carbon cycling, assessment, item response theory, Bayesian approach

INTRODUCTION

Natural phenomena and related social scientific issues are intrinsically interdisciplinary (ID). A variety of academic disciplines in natural science are fundamental for students to explain a phenomenon or related issues along with their reasoning for different phenomena. The current science education system mainly focuses on a discipline-based curriculum, but since the 1980s, efforts have begun to find the balance between specialization and integration. During the 1990s, scholars paid close attention to designing and managing interdisciplinary curricular and associated research projects and reporting the practical and theoretical consequences of relationships between particular disciplines (Klein, 1990). In this trend, national documents and standards of many countries for science education (e.g., the next generation science standards and framework) have demonstrated the value and essentiality of ID approaches to learning and led to substantial support for ID learning.

The definitions and characteristics of "interdisciplinary understanding" have been shown in the literature. Spelt et al. (2009) indicated the importance of individual disciplines which can be integrated for ID understanding. Klein (1990) highlighted that ID understanding is required as an extra step of linking the identified ID knowledge, which is a different aspect from multidisciplinary thinking that does not necessitate the integration of knowledge. Boix Mansilla and Duraisingh (2007) described ID understanding, as follows:

> "The capacity to integrate knowledge and modes of thinking in two or more disciplines or established areas of expertise to produce a cognitive advancement–such as explaining a phenomenon, solving a problem, or creating a product–in ways that would have been impossible or unlikely through single disciplinary means" (p. 219).

Reiska et al. (2018) operationalized ID understanding as

> "students' overall ability to connect knowledge from different fields and their ability to integrate disciplines" (p. 2),

which is conceptualized based on the definitions from Boix Mansilla and Duraisingh (2007) and Klein (1990).

During the past decades, even though there has been a growing emphasis on ID learning and teaching in both secondary- and college-level education, assessments to measure students' ID understanding have not actively been developed. The ID assessment could provide a powerful tool to discover and create links between relevant science subjects but only six ID assessments in the secondary and college level (Reiska et al., 2018; Schaal et al., 2010; Shen et al., 2014; Tripp et al., 2020; Yang et al., 2017; You et al., 2021) were located by

a recent comprehensive literature search. You et al. (2021) developed the interdisciplinary science assessment for carbon cycling (ISACC) to assess the performance of high school and college students in integrating knowledge from different science disciplines to explain a scientific phenomenon, global carbon cycling. The ISACC comprises 19 items, including 11 multiple-choice (MC) items and eight constructed response (CR) items, covering nine core concepts of carbon cycling. The ISACC was administered to 454 students in grades 9-16. The students' data were analyzed to examine construct validity using two item response theory (IRT) models: a two-parameter logistic model (2PLM) for MC questions and a generalized partial credit model (GPCM) for CR questions. The authors validated the assessment using the frequentist approach. The current study was motivated by the desire to elucidate the difference in IRT analysis results of the ISACC using a Bayesian approach to compare the results using the traditional approach. However, the Bayesian approach is scarcely used in science education. The goal of this study is to examine if Bayesian IRT models and traditional IRT methods produce similar parameters on MC and CR items. The guiding research question is, as follows:

RQ: How do Bayesian IRT models compare and contrast to traditional IRT models when the ISACC items are analyzed? Do Bayesian methods and traditional IRT methods produce similar parameter estimates on the items?

Interdisciplinary Science Assessments and Their Data Analysis

Only six ID assessments were identified by a recent, comprehensive literature search using seven databases, including ERIC, Education Source, PsycINFO, Academic Search Complete, Education Research Complete (from EBSCO), Web of Science, ProQuest (for dissertations and theses), and Google Scholar. Search words used included "(Science or STEM) and (interdisciplinary or multidisciplinary or cross-disciplinary or integrated) and assessment." Reiska et al. (2018) examined the change of high school students' ID understanding throughout two years of schooling and explored differences in ID understanding among students from different schools. The authors used a concept mapping method to discover ID understanding and employed an automatic analysis applying a numeric interdisciplinary quality index (IQI) to assess multiple concept maps. Schaal et al. (2010) also used concept maps to assess ninth graders' ID knowledge regarding mammalian hibernation strategies both pre-and post-test. The authors showed improvement in the students' interconnected science concepts with biological and physical perspectives in a post-test compared to a pre-test, which implies some potential of promoting learners' ID abilities through ID instruction. Shen et al. (2014) developed an interdisciplinary assessment to assess college students' ID of osmosis, which involves knowledge from multiple science disciplines. The ID assessment included 15 disciplinary items and 25 interdisciplinary ones. Students' responses were analyzed using the Rasch model (Rasch, 1960, 1980) and Rasch partial credit model (PCM; Masters, 1982). Shen et al. (2014) reported some good psychometric properties, but to have a more reliable and valid tool, other aspects of construct validity are needed, such as infit, dimensionality (for Rasch model's assumption), and differential item functioning to provide information on whether items functioned differently across genders and races. Moreover, as the assessment was implemented as a homework assignment, this may undermine the construct validity. Tripp et al. (2020) developed essay prompts in which students ponder real-world issues that inherently require ID understanding and examined how a previously developed ID rubric captures students' ID understanding in the writing activities for the purpose of validation. The results revealed that the writing assessment did not fully capture students' ID understanding, but instead, their perception regarding interdisciplinary science supported the robustness of the interdisciplinary science framework (IDSF) developed previously by the authors. They argued that the IDSF could be a better model to guide instructors on factors to consider when developing ID curricula and assessments. Yang et al. (2017) designed an interdisciplinary assessment by selecting 20 items targeting crosscutting concepts (CCs). The item format of the selected items was MC or two-tiered MC. A total of 801 students from grades 4 to 8 in five urban schools participated in the assessment. The evidence of reliability and validity has been established using Rasch measurement. The item reliability was 0.98, and person reliability was slightly over 0.60, which does not reach the acceptable value of 0.70. Furthermore, this study reported low person separation of 1.15 (acceptable value is 2), which reflects low power of the items in distinguishing between high and low performers. Even though the assessment had a good fit based on infit and outfit statistics of the items, Yang et al. (2017) reported the misalignment between the average item difficulty and the mean of student ability. There were major gaps at the top of the Wright map, where few students matched the high-difficulty items and at the bottom where no items matched low-ability students. This result indicated the need of items with different difficulty levels for a further revised version of the assessment that could have more desirable construct validity. Last You et al. (2021) developed an assessment to measure students' interdisciplinary understanding of global carbon cycling in the construct-modeling framework (Wilson, 2005). The ISACC was designed for high school and college students. The assessment requires testing students' integration knowledge of physics, chemistry, biology, and earth science. To measure a broad level of disciplinary and ID understanding, the test included both item types: disciplinary and ID items. The disciplinary items require knowledge of only a single science discipline, whereas the ID items require knowledge of more than two science disciplines. Table 1 shows the examples of D and ID items. The details of the ISACC are described in the instrument section in **Table 1**.

Frequentist Versus Bayesian Methods

Recognizing the difference between a Bayesian approach and a frequentist approach is essential for arguing why one procedure might be preferred over the other. Bayesian methods differ from the frequentist approach in three main ways: conceptions of probability, parameters (random vs. fixed), and use of prior information. Frequentist statistics usually allow researchers to perform a hypothesis test and formulate a null and an alternative hypothesis. The null hypothesis assumes no difference between specified populations. A p-value is the probability of obtaining results

Table 1. ISACC items and the discipline descriptions

Items	Discipline(s)	Items	Discipline(s)
I1MC	Earth science	I11MC	Biology, earth science, physics
D2CR	Biology	I12CR	Earth science, physics
I3MC	**Biology, earth science**	I13MC	Biology, earth science
D4CR	Biology	I14CR	Biology, chemistry, earth science
D5MC	Biology	**I15CR**	**Biology, earth science**
D6MC	Biology	D16MC	Earth science
I7CR	Biology, chemistry	D17MC	Chemistry
I8CR	Biology, chemistry, earth science	I18MC	Biology, physics
I9MC	Earth science, physics	I19CR	Biology, earth science
I10MC	Biology, earth science		

Note. D: Single disciplinary; I: Interdisciplinary; MC: Multiple choice; CR: Constructed response;

I3MC: The most popular timber product grown in the United States today is pinus taeda known as loblolly pine. The pine trees' average height and circumference are reported to have been increasing since the 1960s. The level of CO_2 in the atmosphere has also been increasing rapidly since 1950. It has been proposed that the increased burning of fossil fuels might explain the increased growth for this species. Among the following statements, which one do you agree with regarding this proposed explanation?

A. This explanation would be difficult to test because of all the other factors, such as temperature and light level, which might have affected the growth of the trees.
B. This explanation makes sense because an increase in atmospheric CO2, one of the inputs in photosynthesis, will always lead to increased glucose production and more plant growth.
C. This explanation cannot be right because increased CO2 causes global warming, which is detrimental to plant growth.
D. This explanation cannot be right because the burning of fossil fuels releases sulfur dioxide (SO_2), which causes acid rain and kills plants.
E. None of the above

I15CR: The geological carbon cycle is complicated, with many different pieces playing their roles. Usually, a change in one part of the cycle causes compensating changes in other parts, but sometimes the system takes a long time to get back into balance. For example, it takes a long time for the oceans to increase their uptake of carbon dioxide, so they might not be able to compensate for CO_2 increase in the air, resulting in an imbalance. How could deforestation lead to an imbalance in carbon dioxide levels?

at least as extreme as the one in the data, assuming that the null hypothesis is true. In frequentist approaches, parameters are treated as fixed, reflecting fixed features of the population, and the data are treated as a random, varying from sample to sample. The maximum likelihood (ML) estimation widely used in the frequentist methods yields the highest probability of the parameter values of the data observed.

Bayesian statistical methods bring a different philosophy from traditional statistical inference. Bayes's theorem is a statement of conditional probability. The conditional probability is expressed as the degree of uncertainty by treating parameters as random rather than fixed (Kruschke, 2015). In other words, Bayesian inference uses probabilities that are conditional on data to express beliefs about unknown quantities. Bayesian statistics start with a prior distribution. The prior distribution is the "initial" expected distribution of the parameter. The use of prior distributions represents a powerful mechanism for controlling confounding. If prior information is incorporated, model parameters can be updated in Bayesian inference. Once a prior distribution is chosen, the likelihood of the data being given a specific value of a parameter is computed and multiplied by the prior. This yields the probability of the parameter given the data or the posterior distribution. Many researchers agree that it is appealing to consider data and information from previous studies in analyzing current data. Furthermore, another advantage to the Bayesian approach is that the posterior can be continually updated in multiple steps, using the previous posterior distribution as the prior in a further refined procedure. Prior distributions can be noninformative or informative. An informative prior distribution leads to shifts of parameter estimates toward the mean of the prior distribution.

In the Bayesian approach, the likelihood of the data being given a specific value of a parameter is combined with prior information to create a posterior distribution. The posterior distribution is proportional to the product of the prior and the likelihood function, with the likelihood of receiving more weight as the sample size increases. It is assumed that the estimation incorporating accurate prior information regarding parameters could outperform the ML estimation. Gelman and Rubin (1995) indicated that the posterior distribution is located at a point of compromise between the prior distribution and the data, which provide a closed true representation. The posterior distribution is obtained via simulation using Markov chain Monte Carlo (MCMC) methods (Kruschke, 2015). A wealth of Monte Carlo simulation studies and recent Bayesian methodology studies noted the potential benefits of Bayesian methods over ML methods when small samples are considered (e.g., McNeish & Stapleton, 2016). Previous research in IRT revealed the benefits of Bayesian procedures, arguing that Bayesian estimation typically generates item parameters in a more accurate and consistent way than maximum likelihood procedures (e.g., Gao & Chen, 2005; Hsieh et al., 2010; Lord, 1986).

METHOD

Instrument

The ISACC was the first assessment developed to measure the performance of high school and college students in integrating knowledge of a carbon cycling phenomenon. Several previous studies have validated the ISACC through qualitative and quantitative processes (You et al., 2018, 2021,

Table 2. 2PLM item parameters estimates, logit: $a\theta+c$ or $a(\theta-b)$

Item	Discrimination (*a*)	Difficulty (*b*)
I1MC	0.55	-0.73
I3MC	0.46	-0.76
D5MC	1.08	-2.74
D6MC	0.59	-1.00
I9MC	0.65	0.31
I10MC	0.93	0.99
I11MC	0.74	0.86
I13MC	0.91	-0.40
D16MC	0.80	-0.46
D17MC	0.90	-0.40
I18MC	0.50	1.53

Table 3. GPC model item parameter estimates, logit: $a[k(\theta-b)+\Sigma dk]$

Item	Discrimination (*a*)	Location parameter (*b*)	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
D2CR	0.63	0.67	0	0.81	0.44	0.15	0.14	0.81	0.72	-	-
D4CR	0.86	0.14	0	1.85	1.79	1.15	0.34	1.86	2.58	-	-
I7CR	0.36	0.26	0	5.26	5.58	0.20	1.90	0.76	1.66	2.69	2.28
I8CR	0.41	1.74	0	0.16	1.89	0.90	1.23	0.04	0.19	3.41	0.24
I12CR	0.34	0.85	0	4.35	3.91	1.63	6.71	2.03	2.27	0.43	4.83
I14CR	0.30	0.77	0	1.59	2.54	0.37	1.48	2.36	7.95	1.14	3.92
I15CR	0.53	1.13	0	2.92	7.23	0.64	1.39	8.77	2.25	1.24	1.06
I19CR	0.35	1.47	0	3.35	2.52	0.58	2.00	2.03	1.16	0.24	0.52

Note. *a*: Slope parameter; *b*: Item location characterizing overall difficulty; *d*: Threshold parameters

2022). The ISACC assessment items were finalized by the selection of the core concepts of carbon cycling and experts' review for content validity and pilot testing. The final ISACC comprises 11 MC items and eight CR items covering nine core concepts of carbon cycling. For CR items, scoring rubrics were developed to grade. Two IRT models, a 2PLM and a GPCM, were used to validate the ISACC items. In the results, all items were unidimensional, having one carbon cycling construct and the local independence assumption were met. All items except for D5MC showed a good fit to the models and satisfactory psychometric properties.

The item difficulties ranged from -2.63 to 1.48 logits across the MC items. I18MC was the most difficult item, while D5MC was the easiest item. Items I1MC, I3MC, D5MC, D6MC, I13MC, D16MC, and D17MC with negative item difficulties were relatively easy items, and items D2CR, D4CR, I7CR, I9MC, I10MC, I11MC, I12MC, and I14CR had relatively medium difficulty, around 0.5 in the middle of the ability scale. Items I8CR, I15CR, I18MC, and I19CR were hard items. Even at the highest ability level shown (+3), the probability of a correct response was only 0.8 for the difficult items (**Table 2** and **Table 3**).

Discrimination parameters describe how well an item differentiates between people's abilities below the item location and those having abilities above the item location. Discrimination ranged from 0.32 (I14CR) to 1.15 (D5MC). I14CR has low discrimination, and the probability of getting I14CR correct for low performers is almost the same as it is in high performers. D5MC has a very low difficulty level (b=-2.57) and does not provide useful information because 92.7% of the participants were correct on the item. However, D5MC has the highest discrimination (a=1.08), where the probability of a correct answer changes greatly as a person's ability increases. Item parameters in the GPCM are calculated from the value of the slope parameter and the spread of the thresholds (Embretson & Reise, 2013), such that higher values for information have steeper slopes, and the between-category threshold parameters for an item are distributed evenly. D2CR shows the highest value for information across all items, which implies that the item estimates discrimination more accurately than other items (**Table 3**).

Participants and Data

This current study used the same data set as the ISACC. Four hundred fifty-four high school and college students participated, and all students' test responses were collected by a web-based assessment system. Of 454 students, 41.9% were male and 58.1% were female. The racial diversity of the students was: White (39.0%), Asian (28.9%), Hispanic or Latino (23.3%), African American (5.1%), Native Hawaiian or other Pacific Islander (0.4%), and other (3.3%).

Models: Two-Parameter Logistic Model and Generalized Partial Credit Model

The selection of IRT models can be determined based on the number of scored responses and sample size. In this study, the responses are a combination of dichotomous items and polytomous items, and the sample size is 454 individuals. According to recommendations for the reasonable models for parameter estimation (Embretson & Reise, 2013), this study used a mixed-format IRT model: a two-parameter logistic model (2PLM) (Birnbaum, 1968) and a generalized partial credit model (GPCM) (Muraki, 1992). The 2PLM is used for dichotomous score items, and the GPCM is used for CR items with two or more score categories. The two models predict the probability of a correct response to an item based on ability and two item parameters, difficulty and discrimination. The item difficulty parameter (*b*) describes how difficult the item is, whereas the item discrimination parameter (*a*) determines how well an item identifies examinees with different levels of

the latent trait (Embretson & Reise, 2013). Difficult items have large, positive theta values, whereas easy items have large, negative theta values (Reise & Waller, 2002). The theoretical range of the discrimination values is −∞ to +∞; however, items with negative discrimination values are considered problematic. The negative values indicate that examinees with a high level of ability are less likely to answer items correctly.

The item response function of the 2-PLM is defined as

$$P_i(\theta)=\frac{\exp\left[a_i(\theta-b_i)\right]}{1+\exp\left[a_i(\theta-b_i)\right]},$$

where a_i is the discrimination parameter for item i, b_i is the difficulty, and θ is the ability of a person.

Under the GPCM, the probability that responds in category x for item i with m_i+1 categories is expressed as

$$P_{ix}(\theta)=\frac{\exp\left[\sum_{k=0}^{x} a_i(\theta-b_{ik})\right.}{\sum_{h=0}^{mi}\exp\left[\sum_{k=0}^{x} a_i(\theta-b_{ik})\right]},$$

where a_i is the discrimination parameter for item I, and b_{ik} is the step difficulty parameter.

Data Analysis

The current study used Stan software for estimation of the model parameters under a Bayesian approach. Stan uses the no-U-turn sampler (NUTS) algorithm, an extension of the Hamiltonian Monte Carlo (HMC) method (Hoffman & Gelman, 2014), which is faster than other algorithms such as the Gipps sampler and the Metropolis algorithm. Also, Stan has efficient and powerful posterior parameters (Nishio et al., 2020). We will use rstan, an R package that interfaces with Stan in the R computing environment. Basically, a Stan program comprises three basic building blocks: data, parameter, and model blocks.

In the data block of Stan, latent ability and item parameters (i.e., difficulty, discrimination) were included. Hyperparameters in the priors for the two models were specified in the data block. A normal distribution with an unknown mean (mu_beta_di) and unknown standard deviation (sigma_beta_di) was specified as a prior for the item difficulty of the MC items and with an unknown mean (mu_beta_pi) and standard deviation (sigma_beta_pi). The priors for the discrimination parameter for both MC and CR items were specified with a lognormal distribution of including a mean of zero and unknown standard deviation (alpha_di and alpha_pi). For GPCM models, the number of categories was defined as an integer in this data block. The elements in the response matrix range from 0 to the number of categories.

The next component, parameter block, includes model parameters and their hyperparameters, such as latent person ability and item parameters. For example, for the 2PL code, alpha_di represents discrimination of MC items and beta_di is the difficulty of individual MC items. Unknown standard deviations (sigma_beta & sigma_alpha) and the unknown mean of item difficult (mu_beta) are hyperparameters.

In the model block, priors and models are specified. The existing parameters based on the knowledge of likelihood could leverage at least some aspects of building a prior rather than building a completely subjective prior distribution with no knowledge of the likelihood. For Bayesian analysis, the priors have been inferred based on the parameter results in a previously published ISACC paper (You et al., 2021). We placed a normal distribution prior (0, 1) on theta. The hyperprior for the unknown mean (mu_beta) is specified with normal distribution having a mean of 0 and a standard deviation of 5. Another hyperprior for unknown standard deviation (sigma_beta) is specified with a Cauchy distribution with (0, 5). A lognormal distribution with a mean of zero and unknown standard deviation (sigma_alpha) for the discrimination parameter is specified. If a prior is not specified, a uniform prior will be automatically applied by Stan.

The number of chains in the MCMC method was four, and the number of iterations was 500. In the Stan program, posterior distributions of parameters of interest are generated where point estimates such as mean and median, standard deviations, and 95% credible intervals are included. Codes are available in **Appendix A**.

RESULTS

In this study, results of ISACC study previously reported were evaluated using a Bayesian IRT. In Bayesian approach, prior distributions of item parameter and their likelihood from current data are combined into posterior distributions.

Model Convergence

The model convergence was investigated using the Gelman-Rubin convergence diagnostic statistics (Gelman & Rubin, 1992) of Rhat. Additionally, the effective sample size (ESS)–an estimate of the number of independent samples from the HMC posterior distribution–can be used to evaluate convergence (Gelman et al., 2015). A larger ESS implies a lower possibility of autocorrelation. The Rhat values of each parameter that are close to 1 indicate a high chance that the multiple chains have converged to the same distribution. The ESSs ranged from 143 to 1,950. The study validated that 500 iterations were sufficient for reaching a convergence. **Figure 1** shows trace plots for selected parameters. The x-axis of the graphs displays the variability in sampling estimates of the standard error, and the y-axis of each graph represents the value of the item parameters obtained for each sampling chain. Based on the results above, it is assumed that the MCMC algorithm appropriately estimated parameters for the proposed models.

Parameter Estimates

Item parameter results make it possible to validate the individual items to improve the assessment tool. Items analyzed with Bayesian methods yielded similar item difficulty and discrimination parameters to traditional IRT models. The models provide two parameters, the item difficulty and discrimination. **Table 4** shows the parameters computed with Bayesian approaches. The mean of item discrimination ranged from 0.31 to 1.13, which is close to the item discrimination obtained from the traditional IRT approach (0.30 to 1.08). The mean of MC item difficulty was between -2.71 and 1.57. This result showed similarity with the item difficulty range from the conventional approach (-2.63 to 1.48 logits). As 95% credible intervals did not include 0, these estimation results for item discrimination and difficulty show that all items except for one (I9MC) were significantly valuable for measuring students' interdisciplinary understanding of global carbon cycling.

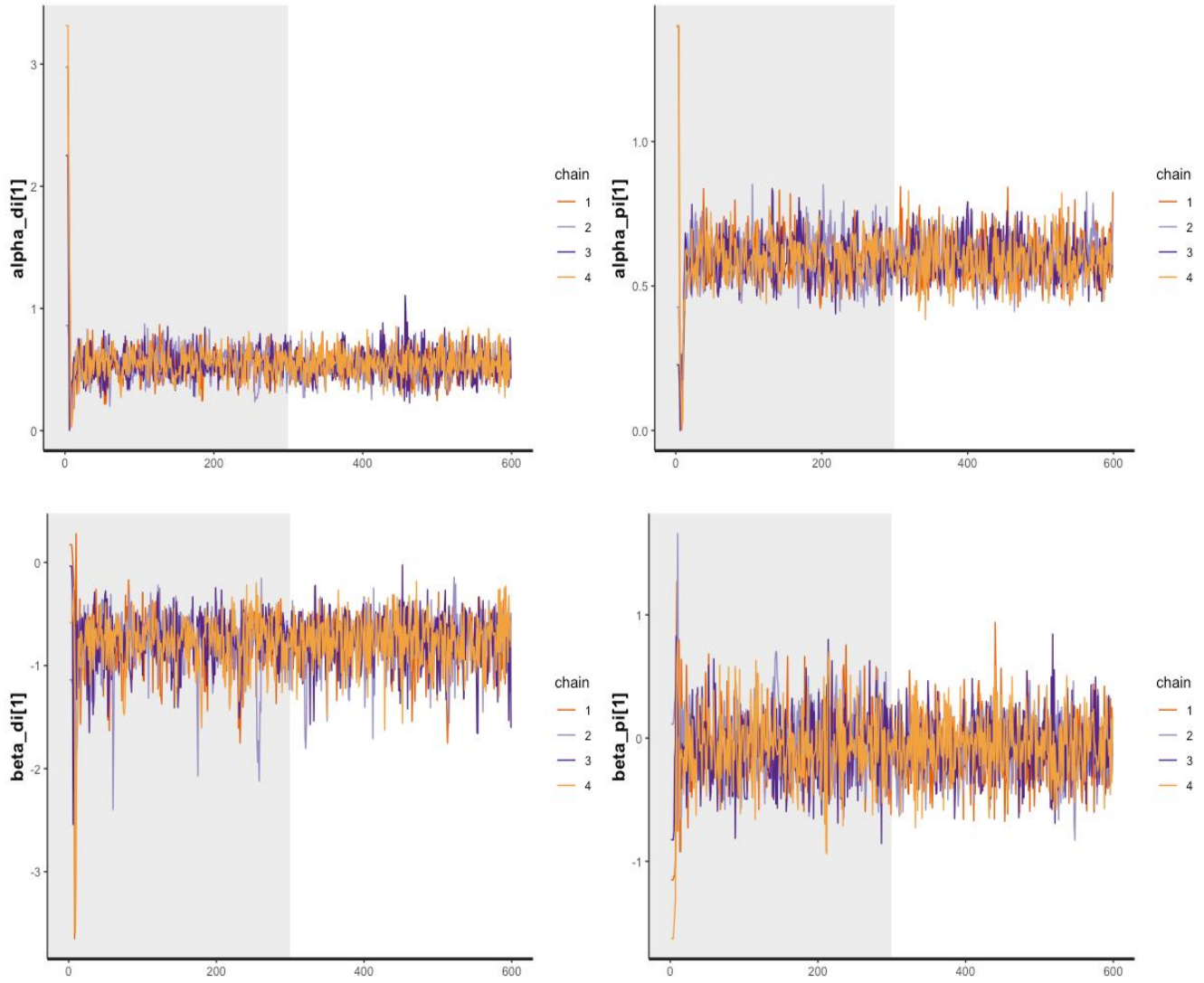

Figure 1. Trace plots for the selected parameters

Table 4. Parameter estimates with Bayesian approach

Item parameter	Item	Mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha_di[1]	D1MC	0.50	0.01	0.13	0.23	0.43	0.51	0.58	0.74	234	1.01
alpha_di[2]	I3MC	0.41	0.00	0.12	0.18	0.34	0.41	0.49	0.65	730	1.00
alpha_di[3]	D5MC	1.13	0.01	0.20	0.76	0.99	1.12	1.27	1.55	956	1.00
alpha_di[4]	D6MC	0.55	0.00	0.12	0.33	0.47	0.55	0.63	0.80	711	1.00
alpha_di[5]	I9MC	0.61	0.00	0.13	0.36	0.52	0.61	0.69	0.88	823	1.00
alpha_di[6]	I10MC	0.91	0.01	0.15	0.64	0.81	0.91	1.01	1.21	867	1.00
alpha_di[7]	11MC	0.70	0.00	0.14	0.45	0.61	0.70	0.79	0.98	881	1.00
alpha_di[8]	13MC	0.88	0.00	0.14	0.63	0.78	0.88	0.96	1.17	1113	1.00
alpha_di[9]	16MC	0.76	0.00	0.13	0.51	0.67	0.76	0.85	1.01	874	1.00
alpha_di[10]	17MC	0.87	0.00	0.13	0.62	0.78	0.87	0.96	1.15	710	1.00
alpha_di[11]	I18MC	0.50	0.00	0.11	0.30	0.43	0.50	0.58	0.73	864	1.00
alpha_pi[1]	D2CR	0.59	0.00	0.07	0.46	0.54	0.59	0.64	0.75	511	1.01
alpha_pi[2]	D4CR	0.81	0.00	0.10	0.64	0.75	0.81	0.87	1.02	563	1.00
alpha_pi[3]	I7CR	0.36	0.00	0.04	0.29	0.34	0.36	0.39	0.45	483	1.00
alpha_pi[4]	I8CR	0.39	0.00	0.05	0.30	0.35	0.38	0.42	0.49	653	1.00
alpha_pi[5]	I12CR	0.35	0.00	0.04	0.28	0.33	0.35	0.38	0.43	591	1.00
alpha_pi[6]	I14CR	0.31	0.00	0.04	0.24	0.28	0.31	0.33	0.39	529	1.00
alpha_pi[7]	I15CR	0.57	0.00	0.06	0.46	0.53	0.57	0.61	0.69	601	1.00
alpha_pi[8]	I19CR	0.34	0.00	0.04	0.26	0.31	0.33	0.37	0.43	515	1.00
beta_di [1]	D1MC	-0.86	0.04	0.43	-1.88	-0.97	-0.78	-0.62	-0.36	143	1.02
beta_di [2]	I3MC	-0.89	0.02	0.39	-1.88	-1.06	-0.83	-0.64	-0.30	476	1.00
beta_di [3]	D5MC	-2.71	0.02	0.40	-3.67	-2.92	-2.67	-2.43	-2.11	653	1.00

Table 4 (continued). Parameter estimates with Bayesian approach

Item parameter	Item	Mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta_di [4]	D6MC	-1.10	0.01	0.32	-1.81	-1.26	-1.06	-0.88	-0.63	540	1.00
beta_di [5]	I9MC	0.32	0.01	0.18	-0.01	0.20	0.32	0.43	0.73	1094	1.00
beta_di [6]	I10MC	1.02	0.01	0.17	0.73	0.90	1.00	1.13	1.39	734	1.00
beta_di [7]	11MC	0.90	0.01	0.22	0.55	0.75	0.87	1.02	1.43	533	1.00
beta_di [8]	13MC	-0.42	0.00	0.14	-0.70	-0.50	-0.40	-0.32	-0.16	755	1.00
beta_di [9]	16MC	-0.49	0.01	0.17	-0.85	-0.58	-0.48	-0.38	-0.18	770	1.00
beta_di [10]	17MC	-0.42	0.01	0.14	-0.74	-0.50	-0.41	-0.33	-0.17	607	1.00
beta_di [11]	I18MC	1.57	0.01	0.38	0.97	1.29	1.53	1.77	2.49	685	1.00
beta_pi [1]	D2CR	-0.07	0.01	0.25	-0.54	-0.24	-0.08	0.09	0.44	930	1.00
beta_pi [2]	D2CR	0.26	0.01	0.28	-0.27	0.08	0.25	0.44	0.86	1129	1.00
beta_pi [3]	D2CR	0.55	0.01	0.32	-0.04	0.34	0.54	0.76	1.23	809	1.00
beta_pi [4]	D2CR	0.52	0.01	0.32	-0.17	0.32	0.54	0.73	1.12	1113	1.00
beta_pi [5]	D2CR	1.53	0.01	0.38	0.77	1.28	1.51	1.77	2.31	888	1.00
beta_pi [6]	D2CR	1.37	0.02	0.44	0.47	1.10	1.37	1.66	2.22	846	1.00
beta_pi [7]	D4CR	-1.69	0.01	0.35	-2.34	-1.92	-1.70	-1.46	-1.02	1895	1.00
beta_pi [8]	D4CR	-1.70	0.01	0.29	-2.31	-1.88	-1.68	-1.49	-1.15	997	1.00
beta_pi [9]	D4CR	-1.06	0.01	0.21	-1.50	-1.20	-1.05	-0.92	-0.68	873	1.00
beta_pi [10]	D4CR	0.49	0.00	0.17	0.17	0.37	0.49	0.60	0.81	1138	1.00
beta_pi [11]	D4CR	2.10	0.01	0.28	1.56	1.90	2.10	2.28	2.67	919	1.00
beta_pi [12]	D4CR	2.79	0.01	0.47	1.95	2.45	2.78	3.08	3.79	1177	1.00
beta_pi [13]	I7CR	4.66	0.03	0.95	3.00	3.99	4.59	5.33	6.59	1043	1.00
beta_pi [14]	I7CR	-4.40	0.03	0.90	-6.20	-5.02	-4.38	-3.74	-2.77	1276	1.00
beta_pi [15]	I7CR	0.32	0.02	0.06	-0.81	-0.08	0.31	0.70	1.51	1381	1.00
beta_pi [16]	I7CR	-1.56	0.02	0.59	-2.72	-1.93	-1.53	-1.17	-0.47	1326	1.00
beta_pi [17]	I7CR	1.02	0.01	0.51	0.07	0.69	1.02	1.33	2.03	1274	1.00
beta_pi [18]	I7CR	1.75	0.02	0.64	0.51	1.35	1.72	2.16	3.07	1243	1.00
beta_pi [19]	I7CR	-2.25	0.02	0.64	-3.58	-2.69	-2.21	-1.78	-1.07	869	1.00
beta_pi [20]	I7CR	2.46	0.02	0.57	1.40	2.06	2.45	2.85	3.63	1339	1.00
beta_pi [21]	I8CR	2.06	0.02	0.51	1.21	1.71	2.01	2.39	3.18	856	1.00
beta_pi [22]	I8CR	-0.14	0.01	0.46	-1.13	-0.44	-0.13	0.16	0.72	1470	1.00
beta_pi [23]	I8CR	0.88	0.01	0.47	0.02	0.58	0.86	1.18	1.86	1311	1.00
beta_pi [24]	I8CR	0.52	0.01	0.49	-0.51	0.22	0.51	0.83	1.49	1310	1.00
beta_pi [25]	I8CR	1.86	0.02	0.63	0.65	1.44	1.84	2.26	3.15	1093	1.00
beta_pi [26]	I8CR	2.11	0.02	0.76	0.66	1.60	2.09	2.60	3.72	1328	1.00
beta_pi [27]	I8CR	4.84	0.04	1.17	2.67	4.05	4.79	5.59	7.34	1045	1.00
beta_pi [28]	I8CR	2.21	0.03	1.46	-0.50	1.20	2.21	3.20	5.08	1771	1.00
beta_pi [29]	I12CR	4.37	0.03	0.91	2.77	3.73	4.31	4.88	6.36	879	1.00
beta_pi [30]	I12CR	-2.26	0.02	0.85	-4.00	-2.81	-2.22	-1.69	-0.66	1655	1.00
beta_pi [31]	I12CR	1.64	0.03	0.88	-0.02	1.00	1.61	2.25	3.41	1144	1.00
beta_pi [32]	I12CR	-5.06	0.03	0.94	-7.06	-5.70	-4.97	-4.38	-3.44	904	1.00
beta_pi [33]	I12CR	-1.17	0.01	0.43	-2.05	-1.47	-1.17	-0.85	-0.36	1307	1.00
beta_pi [34]	I12CR	2.93	0.02	0.54	1.98	2.55	2.87	3.28	4.13	839	1.00
beta_pi [35]	I12CR	0.59	0.01	0.57	-0.51	0.21	0.60	0.96	1.74	1634	1.00
beta_pi [36]	I12CR	5.03	0.03	1.00	3.30	4.33	4.96	5.63	7.26	1123	1.00
beta_pi [37]	I14CR	2.12	0.02	0.69	0.91	1.64	2.08	2.54	3.59	881	1.00
beta_pi [38]	I14CR	-1.58	0.02	0.67	-2.89	-2.02	-1.57	-1.10	-0.31	926	1.00
beta_pi [39]	I14CR	0.31	0.02	0.62	-0.81	-0.10	0.30	0.71	1.57	1043	1.00
beta_pi [40]	I14CR	-0.69	0.02	0.65	-1.95	01.13	-0.65	-0.25	0.52	1420	1.00
beta_pi [41]	I14CR	-1.40	0.02	0.58	-2.62	-1.78	-1.40	-1.00	-0.34	1103	1.00
beta_pi [42]	I14CR	7.70	0.05	1.18	5.68	6.86	7.62	8.46	10.15	632	1.00
beta_pi [43]	I14CR	1.90	0.03	1.21	-0.42	1.14	1.84	2.69	4.37	1446	1.00
beta_pi [44]	I14CR	-2.10	0.03	1.16	-4.42	-2.90	-2.03	-1.30	0.07	1114	1.00
beta_pi [45]	I15CR	3.27	0.03	0.75	1.96	2.76	3.22	3.72	4.94	639	1.00
beta_pi [46]	I15CR	-5.29	0.03	0.77	-6.94	-5.75	-5.23	-4.76	-3.95	549	1.00
beta_pi [47]	I15CR	0.43	0.01	0.25	-0.05	0.26	0.43	0.60	0.92	1546	1.00
beta_pi [48]	I15CR	-0.18	0.01	-0.25	-0.68	-0.34	-0.16	-0.01	0.27	962	1.00
beta_pi [49]	I15CR	7.81	0.04	1.13	5.90	6.99	7.75	8.54	10.21	855	1.00
beta_pi [50]	I15CR	0.65	0.04	1.28	-1.96	-0.18	0.65	1.51	3.17	1138	1.00
beta_pi [51]	I15CR	0.19	0.03	1.17	-2.26	-0.60	0.22	1.01	2.43	1386	1.00
beta_pi [52]	I15CR	1.97	0.02	1.07	-0.05	1.25	1.94	2.66	4.15	1839	1.00

Table 4 (Continued). Parameter estimates with Bayesian approach

Item parameter	Item	Mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta_pi [53]	I19CR	4.97	0.03	0.86	3.42	4.39	4.86	5.51	6.76	627	1.00
beta_pi [54]	I19CR	-0.86	0.02	0.63	-2.15	-1.25	-0.84	-0.42	0.28	1138	1.00
beta_pi [55]	I19CR	1.98	0.02	0.66	0.74	1.54	1.96	2.40	3.31	1087	1.00
beta_pi [56]	I19CR	-0.36	0.02	0.70	-1.76	-0.82	-0.33	0.12	0.99	1229	1.00
beta_pi [57]	I19CR	3.36	0.02	0.86	1.76	2.73	3.30	3.96	5.09	1193	1.00
beta_pi [58]	I19CR	0.56	0.02	0.97	-1.48	-0.06	0.58	1.26	2.39	1616	1.00
beta_pi [59]	I19CR	1.59	0.03	0.96	-0.43	0.96	1.59	2.21	3.51	1276	1.00
beta_pi [60]	I19CR	0.98	0.02	1.03	-1.04	0.32	0.97	1.68	2.93	1950	1.00

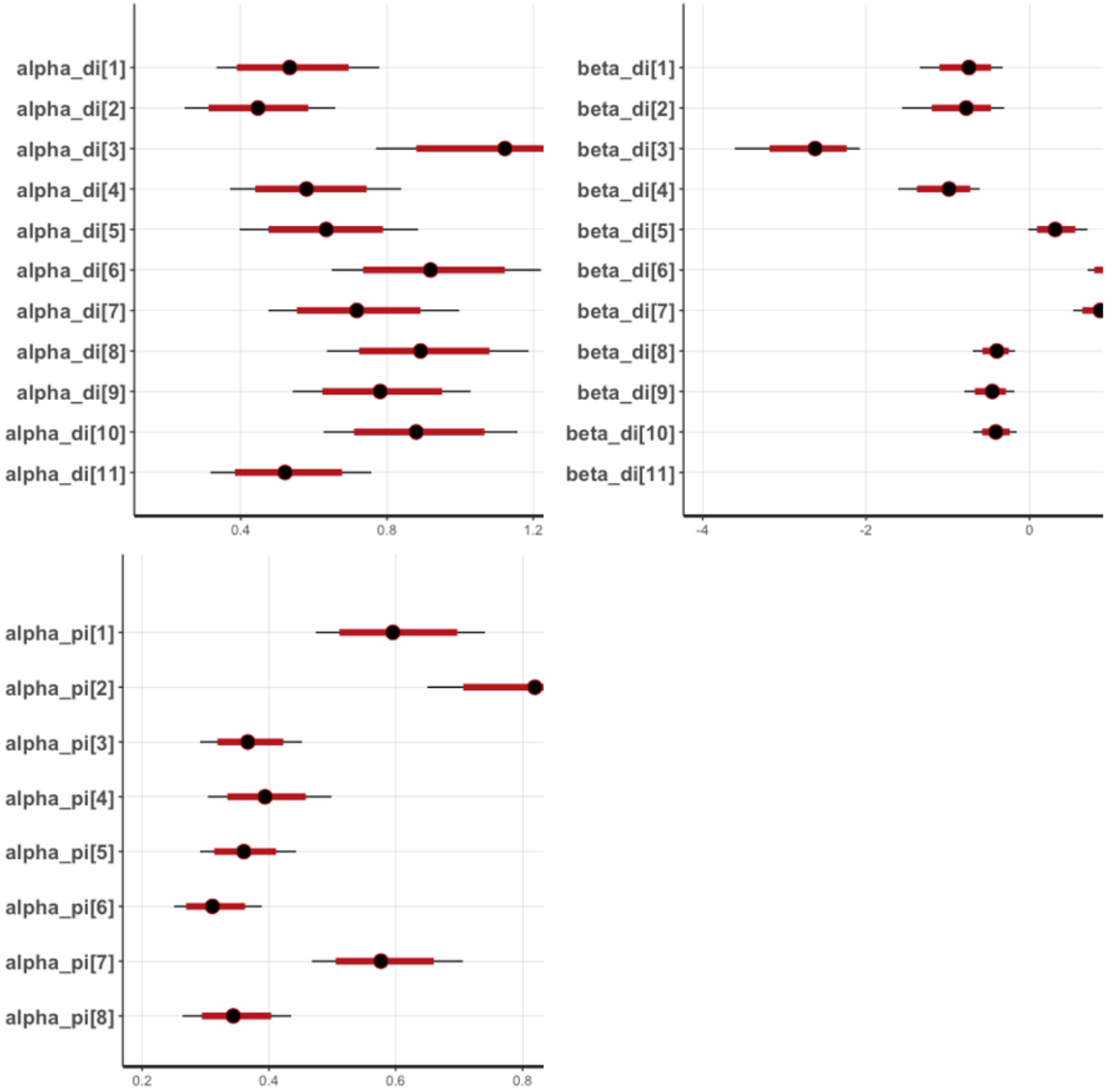

Figure 2. Posterior intervals and point estimates of item difficulty and discrimination

Figure 2 and **Figure 3** show posterior intervals and point estimates of item difficulty and discrimination and kernel density graphs for the parameters, respectively. The distribution of item difficulty and discrimination in the kernel density graphs are normal.

DISCUSSION

The overarching goal of this study is to determine if the ISACC items are helpful in assessing students' interdisciplinary understanding using the Bayesian IRT

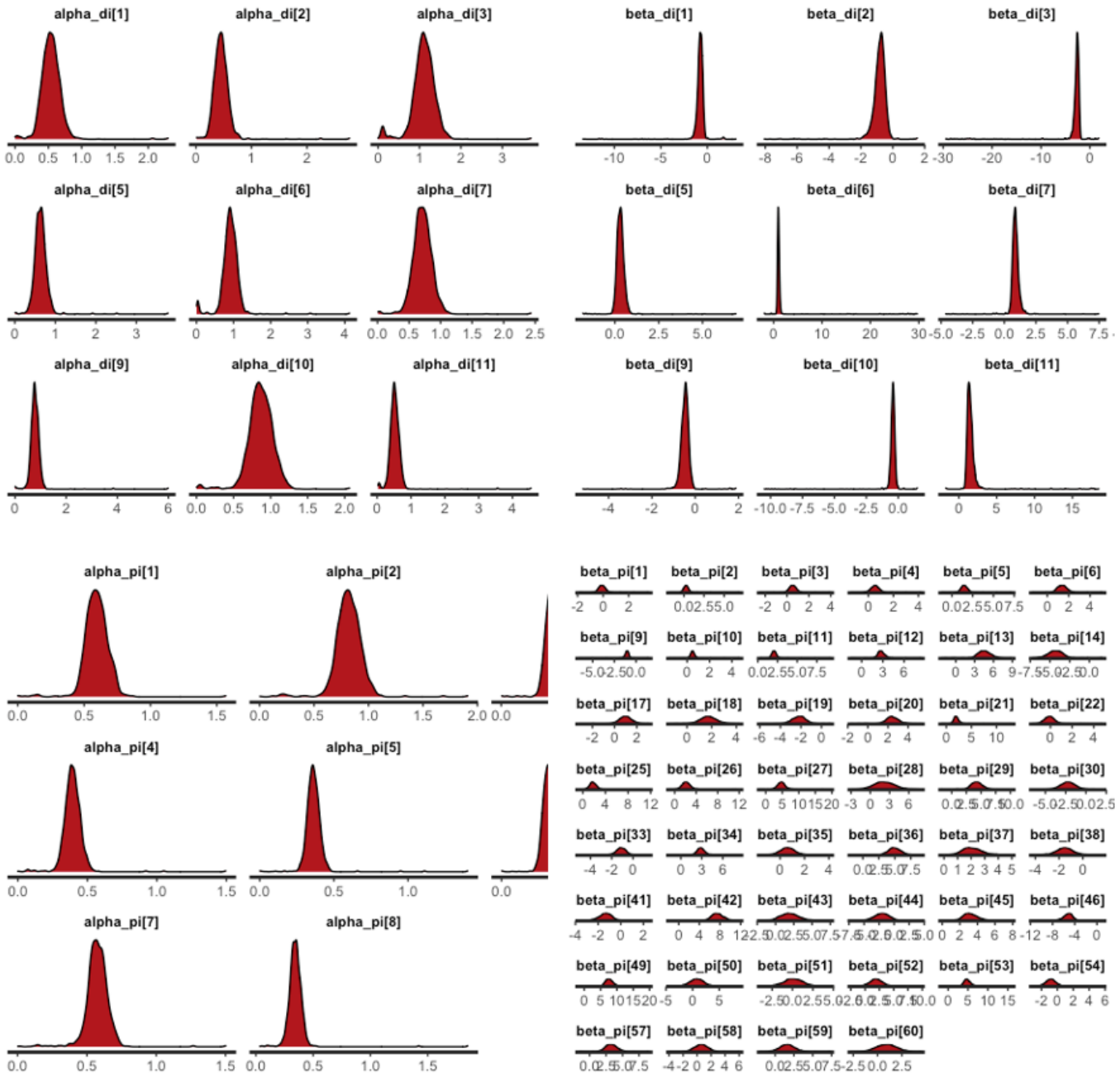

Figure 3. Kernel density graphs for items

models by determining if the results of the item parameters generated by the traditional IRT models are similar or not. Thus, the approach using Bayesian IRT analysis provides an opportunity for cross-validation to generalize the use of the ISACC. The item parameters from the traditional and Bayesian methods are similar and consistent. The Bayesian models tend more toward the center of the distribution than the conventional models. This result indicates that the Bayesian models are likely to estimate the difficulty of the items to be easier than the traditional models. However, the values are not significantly different from one another. The 95% credible interval for each item showed zero in the interval for only one item, I9MC. Note that, unlike a confidence interval, a credible interval is defined as a probability that the true parameter value is in the interval. The application of the MCMC in estimating item parameters is robust. The trace plots of each item represented the convergence through which the MCMC was implemented successfully.

However, the chosen priors yielded similar results to maximum likelihood-based inferences, which often happens in large samples. Subjective priors have been the most controversial aspect of Bayesian statistics. Some researchers believe that subjective priors can compromise the integrity of the study results and can even lead to conclusions driven not by the data but by a prior. Thus, choosing informative priors that quantify prior beliefs or empirical evidence about the possible values for the previous data is recommended. Even though the results in the Bayesian approach are similar to the results of the traditional IRT, there are some advantages in Bayesian modeling; it makes inferences about basic parameters, intermediate parameters, and hyperparameters

simultaneously, where the frequentist approach does not. Ironically, in frequentist IRT modeling, prediction of theta (i.e., actual measurement of persons) must appear in a second, post-estimation process (Furr, 2017). In addition, estimates through Bayesian methods are asymptotically distribution free, which lowers the dependency on the data distribution. The most crucial advantage of Bayesian methods is when the data have small samples and/or inaccurate estimation of parameters with extreme response patterns (Lord, 1986). Luo et al. (2013) reported that Bayesian IRT modeling using an MCMC approach produces more accurate results when normality assumptions on errors are violated and these violations are taken into account. As data in this study included a large sample size of 454 with a not-skewed normal distribution, the results may not benefit from the strength of Bayesian modeling.

The systematic development and analysis processes used can be expected to yield assessment tools that have strong psychometric properties and will be valuable for teachers in the classroom. If developing another version of the ISACC, a larger sample would be needed to have more robust statistical results. A larger sample size makes it possible to conduct different IRT models with more parameters (e.g., 3PLM).

Additionally, the results of the ISACC should be inferenced with generalizability. Different demographic information or contextual factors (e.g., a low socioeconomic status sample versus a high socioeconomic status sample, different raters, different geographical education settings) lead to different assessments' inferences (Kane, 2006; Nitko & Brookhart, 2010). In order to investigate the impact of the differences, the items can be administered to a sample of students from different backgrounds, for example, students in different counties. Such differences influence the cognitive process of ID understanding and demonstrate different patterns in the ISACC. Statistical models or qualitative models can be used as methods or tools to interpret the patterns of the data collected through assessment tasks. For example, differential item functioning (DIF) analysis is concerned with identifying significant differences across subgroups (e.g., commonly gender or ethnicity). The DIF provides helpful evidence to determine the measurement bias across the groups, identifying assessment items that are differentially difficult for examinees with the same ability regarding a construct. The results of the study imply additional research that requires the development and validation of the ISACC. A new version, ISACC II has been finalized based on the current psychometrical feedback and the statistical results are under analysis. The results will be reported elsewhere in the near future.

Funding: No external funding is received for this article.

Declaration of interest: The author declares that there are no competing interests.

Ethics approval and consent to participate: Not applicable.

Availability of data and materials: All data generated or analyzed during this study are available for sharing when appropriate request is directed to author.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Addison-Wesley.

Boix Mansilla, V., & Duraisingh, E. D. (2007). Targeted assessment of students' interdisciplinary work: An empirically grounded framework proposed. *The Journal of Higher Education, 78*(2), 215-237. https://doi.org/10.1080/00221546.2007.11780874

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Furr, D. C. (2017). *Bayesian and frequentist cross-validation methods for explanatory item response models*. University of California, Berkeley.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351-380. https://doi.org/10.1207/s15324818ame1804_2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472. https://doi.org/10.1214/ss/1177011136

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25*, 165-173. https://doi.org/10.2307/271064

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530-543. https://doi.org/10.3102/1076998615606113

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593-1623.

Hsieh, M. I., Proctor, T. P., Hou, J. I., & Teo, K. S. (2010). A comparison of Bayesian MCMC and marginal maximum likelihood methods in estimating the item parameters of the 2PL IRT model. *International Journal of Innovative Management, Information & Production*, *1*(1), 81-89.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). American Council on Education/Macmillan.

Klein, J. T. (1990). *Interdisciplinarity: History, theory, and practice*. Wayne State University Press.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Elsevier Science. https://doi.org/10.1016/B978-0-12-405888-0.00008-8

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2) 157-162. https://doi.org/10.1111/j.1745-3984.1986.tb00241.x

Luo, S., Ma, J., & Kieburtz, K. D. (2013). Robust Bayesian inference for multivariate longitudinal data by using normal/independent distributions. *Statistics in Medicine, 32*(22), 3812-3828. https://doi.org/10.1002/sim.5778

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314. https://doi.org/10.1007/s10648-014-9287-x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. https://doi.org/10.1177/014662169201600206

Nishio, M., Akasaka, T., Sakamoto, R., & Togashi, K. (2020). Bayesian statistical model of item response theory in observer studies of radiologists. *Academic Radiology*, *27*(3), e45-e54. https://doi.org/10.1016/j.acra.2019.04.014

Nitko, A. J., & Brookhart, S. M. (2010). *Educational assessment of students*. Pearson Education.

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Reise, S. P., & Waller, N. G. (2002). Item response theory for dichotomous assessment data. In F. Drasgow, & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 88-122). Jossey-Bass.

Reiska, P., Soika, K., & Cañas, A. J. (2018). Using concept mapping to measure changes in interdisciplinary learning during high school. *Knowledge Management & E-Learning: An International Journal, 10*(1), 1-24. https://doi.org/10.34105/j.kmel.2018.10.001

Schaal, S., Bogner, F. X., & Girwidz, R. (2010). Concept mapping assessment of media assisted learning in interdisciplinary science education. *Research in Science Education, 40*(3), 339-352. https://doi.org/10.1007/s11165-009-9123-3

Shen, J., Liu, O. L., & Sung, S. (2014). Designing interdisciplinary assessments in sciences for college students: An example on osmosis. *International Journal of Science Education, 36*(11), 1773-1793. https://doi.org/10.1080/09500693.2013.879224

Spelt, E. J., Biemans, H. J., Tobi, H., Luning, P. A., & Mulder, M. (2009). Teaching and learning in interdisciplinary higher education: A systematic review. *Educational Psychology Review, 21*(4), 365-378. https://doi.org/10.1007/s10648-009-9113-z

Tripp, B., Voronoff, S. A., & Shortlidge, E. E. (2020). Crossing boundaries: Steps toward measuring undergraduates' interdisciplinary science understanding. *CBE—Life Sciences Education, 19*(1), ar8. https://doi.org/10.1187/cbe.19-09-0168

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Yang, Y., He, P., & Liu, X. (2017). Validation of an instrument for measuring students' understanding of interdisciplinary science in grades 4-8 over multiple semesters: A Rasch measurement study. *International Journal of Science and Mathematics Education, 16*(4), 639-654. https://doi.org/10.1007/s10763-017-9805-7

You, H. S., Marshall, J. A., & Delgado, C. (2018). Assessing students' disciplinary and interdisciplinary understanding of global carbon cycling. *Journal of Research in Science Teaching*, *55*(3), 377-398. https://doi.org/10.1002/tea.21423

You, H. S., Marshall, J. A., & Delgado, C. (2021). Toward interdisciplinary learning: Development and validation of an assessment for interdisciplinary understanding of global carbon cycling. *Research in Science Education, 51*, 1197-1221. https://doi.org/10.1007/s11165-019-9836-x

You, H. S., Park, S., Marshall, J. A., & Delgado, C. (2022). Interdisciplinary science assessment of carbon cycling: Construct validity evidence based on internal structure. *Research in Science Education*, *52*(5), 473-492. https://doi.org/10.1007/s11165-020-09943-9

APPENDIX A

Code for 2PLM and GPCM

```
###load Rstan package
library(rstan)
rstan_options(auto_write=TRUE)
options(mc.cores=parallel::detectCores())
install.packages("mirt")
library(mirt)
library("acnr")
library("bayesplot")
################################################################################
irt_code<-"
data{
int<lower=2, upper=10> k_pi[8]; //number of categories for each polytomous item
int <lower=0> nk; //total number of categories in all polytomous items
int<lower=0> k_index[9]; //categories index used for ragged structure
int <lower=0> n_student; //number of individuals
int <lower=0> n_item; //number of items
int <lower=0> n_di; //number of dichotomous items
int <lower=0> n_pi; //number of polytomous items
int<lower=0> Y[n_student,n_item]; //array of responses
}
parameters {
real<lower=0> alpha_di [n_di]; //item discrimination for dichotomous items
real beta_di [n_di]; //item difficulty for dichotomous items
real<lower=0> alpha_pi [n_pi]; //item discrimination for polytomous items
vector[nk] beta_pi; //item difficulty parameter for polytomous items
real mu_beta_di; //mean difficulty of dichotomous items
real<lower=0> sigma_beta_di; //difficulty sd of dichotomous items
real mu_beta_pi; //mean difficulty of polytomous items
real<lower=0> sigma_beta_pi; //difficulty sd of polytomous items
vector[n_student] theta; //latent trait
}
model{
theta ~ normal(0,1);
beta_di ~ normal(mu_beta_di,sigma_beta_di);
mu_beta_di ~ normal(0,5);
sigma_beta_di ~ cauchy(0,5);
beta_pi ~ normal(mu_beta_pi,sigma_beta_pi);
mu_beta_pi ~ normal(0,5);
sigma_beta_pi ~ cauchy(0,5);
alpha_di ~lognormal(0, 1);
alpha_pi ~lognormal(0, 1);
for (i in 1:n_student){
for (j in 1:n_di){Y[i,j] ~ bernoulli_logit(alpha_di[j]*(theta[i]-beta_di[j]));
}
for (j in (n_di+1):n_item){vector[k_pi[j-n_di]+1] p;
vector[k_pi[j-n_di]] beta=beta_pi[(k_index[j-n_di]+1):k_index[j-n_di+1]];

p=softmax(cumulative_sum(append_row(rep_vector(0.0, 1), alpha_pi[j-n_di]*(theta[i] - beta))));
Y[i,j] ~ categorical(p);
}}
}
```

```
generated_quantities <-"
generated quantities {
real theta_rep[n_student]
int y_rep[n_student];
for (i in 1:n_student)
theta_rep[n_student] <- normal_rng(0.1))
y_rep[i] <-bernoulli_rng(inv_logit((alpha_di[j]*(theta[i]-beta_di[j])))
}
"
################################################################################
resp<-read.table("resp.csv",header=T,sep=",")
apply(resp,2,table)
#number of polytomous and dichotomous items
n_pi<-length(which(apply(resp,2,max)>1))
n_di<-length(which(apply(resp,2,max)==1))
#number of categories for each polytous item
k_pi<-apply(resp,2,max)[1:n_pi]
#total number of categories
nk<-sum(k_pi)
resp<-resp[,c(9:19,1:8)]
resp[,12:19]<-resp[,12:19]+1
###index used in ragged data structure in stan
k_index<-c(seq(0,12,by=6),seq(20,60,by=8))
I<-dim(resp)[1]
J<-dim(resp)[2]
data_irt<-list(n_student=I,n_item=J,n_pi=n_pi,n_di=n_di,Y=resp,
k_index=k_index,k_pi=k_pi,nk=nk)
irt_mixed <- stan(model_code=irt_code, data=data_irt, iter =500,chains=4)
irt_mixed1 <- stan(fit=irt_mixed, data=data_irt, iter =600,chains=4)
print(irt_mixed1,par=c("alpha_di","alpha_pi","beta_di","beta_pi"))
y_rep <- extract(irt_mixed1, pars="y_rep", permuted=true)$y_rep
str(y_rep)
stan_trace(irt_mixed1, inc_warmup=TRUE, pars='alpha_di[1]')
stan_trace(irt_mixed1, inc_warmup=TRUE, pars='alpha_pi[1]')
stan_trace(irt_mixed1, inc_warmup=TRUE, pars='beta_di[1]')
stan_trace(irt_mixed1, inc_warmup=TRUE, pars='beta_pi[1]')
stan_plot(irt_mixed1, inc_warmup=TRUE, pars='alpha_di')
stan_plot(irt_mixed1, inc_warmup=TRUE, pars='alpha_pi')
stan_plot(irt_mixed1, inc_warmup=TRUE, pars='beta_di')
stan_plot(irt_mixed1, inc_warmup=TRUE, pars='beta_pi[1]')
stan_dens(irt_mixed1, inc_warmup=TRUE, pars='alpha_di')
stan_dens(irt_mixed1, inc_warmup=TRUE, pars='alpha_pi')
stan_dens(irt_mixed1, inc_warmup=TRUE, pars='beta_di')
stan_dens(irt_mixed1, inc_warmup=TRUE, pars='beta_pi')

sink("stan_output.txt")
print(irt_mixed1,digit=3)
sink()
################################################################################
#estimate the 2pl/gpcm models with MMLE
library(mirt)
#path<-"C:\\Users\\jacky\\Dropbox\\Collaboration Project"
#setwd(path)
resp<-read.table("resp.csv",header=T,sep=",")
resp<-resp[,c(9:19,1:8)]
```

```
#locate max item point for each item
mp<-apply(resp,2,max)
#item type
it_ty<-rep(NA,19)
it_ty[which(mp==1)]="2PL"
it_ty[which(mp!=1)]="gpcm"
mod<-mirt(resp,1,itemtype=it_ty)
par<-coef(mod,IRTpars=TRUE,simplify=TRUE)$item
print(par)
```