Psychometric Models for a New State Science Assessment Aligned to the Next Generation Science Standards

Jing Chen, Jonghwan Lee, Paul Nichols, and M. Christina Schneider

1 Introduction

Unlike traditional unidimensional science standards, the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) emphasize three distinct dimensions: Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs). These dimensions are combined to form performance expectations that reflect the inherent complexity in scientific understanding and reasoning. The complexity of the standards and the new task types they require poses significant challenges for psychometric modeling (Gorin & Mislevy, 2013).

The explicit dimensionality in the construct as defined by the NGSS impacts the choice of measurement models for an NGSS assessment. Meanwhile, to measure the NGSS, performance tasks are designed to elicit responses that are more aligned with the targeted reasoning and higher cognitive skills. These tasks often include contextualized and multidimensional items to measure real-world problem-solving skills, which may violate the assumptions of traditional psychometric models (Martineau, 2017). The psychometric challenges introduced by the NGSS require appropriate models to assess the dimensionality and to estimate item and person parameters.

The goal of this study is to identify an appropriate measurement model for an NGSS-aligned state summative science assessment. The assessment was recently created to align to the state's college and career ready standards for science designed around NGSS' three-dimensional science learning. Because of the multidimensional nature of the assessment, the most appropriate measurement model that could be supported by learning theories, capture the patterns within the data, and be

J. Chen (✉) · J. Lee · P. Nichols · M. C. Schneider
NWEA, Portland, OR, USA
e-mail: jing.chen@nwea.org; jay.lee@nwea.org; paul.nichols@nwea.org; christina.schneider@nwea.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 353, https://doi.org/10.1007/978-3-030-74772-5_36

feasible to use in an operational setting was investigated. The following sections provide more details about the science assessment and its pilot administration, the dimensionality analyses and results, and a discussion of the findings.

2 Science Pilot Overview

This study was conducted based on data from a pilot test of a new state science assessment administered in Grade 5 and Grade 8 in Spring 2019. The assessment is based on performance tasks, which are phenomena-based scenarios with multiple items to elicit responses that show students' understanding of the DCIs, SEPs, and CCCs. The items are minimally two dimensional. A variety of technology-enhanced item types are used that allow students to show their thinking more fully. For example, the drag-and-drop technology-enhanced item type requires students to drag and drop items into groups. Within each group, students can rank items by dragging and dropping them into place.

Each grade-level pilot test had two test forms (Form A and Form B) that each consisted of two tasks and several items. The two forms at Grade 5 had 11 and 14 items, respectively, and the two forms at Grade 8 had 17 and 18 items, respectively. All items were scored dichotomously. The pilot test was intentionally short to reduce the time students spent away from the classroom.

The student sample for this study was a convenience sample based on schools' availability and willingness to participate. Table 1 presents the total number of students who took the test by grade and form. The student sample's demographic information (including sex and ethnicity) presented in Table 2 suggests that the sample had demographic characteristics similar to the state's general student population at these two grade levels. The differences in percentages between the sample and the general population are all smaller than 5%. In addition, because the two forms at each grade were randomly administered to students within the same school, students were comparable across the forms in terms of their demographics.

Table 1 Pilot sample

	Number of Students		
Grade	Form A	Form B	Total number of students
5	2739	2495	5234
8	3081	2770	5851
Total			11,085

Table 2 Demographic information: Pilot sample vs. general population of the state

		Pilot sample				General population			
		Grade 5		Grade 8		Grade 5		Grade 8	
Demographic variable		N	%	N	%	N	%	N	%
Sex	Female	2351	48.5	2531	48.9	11,789	48.8	11,579	48.9
	Male	2501	51.5	2641	51.1	12,375	51.2	12,117	51.1
Ethn-icity	AIAN[a]	68	1.4	82	1.6	307	1.3	320	1.4
	Asian	144	3.0	111	2.1	664	2.7	638	2.7
	Black	181	3.7	193	3.7	1603	6.6	1654	7.0
	Hispanic	899	18.5	941	18.2	4886	20.2	4660	19.7
	White	3380	69.7	3674	71.0	15,666	64.8	15,513	65.5
	Two or more races	169	3.5	160	3.1	1038	4.3	911	3.8
Total		4841	100.0	5161	100.0	24,164	100.0	23,696	100.0

Note: Around 10% of the students did not have demographic information available and were excluded from Table 2. However, their responses were included in all other analyses
[a]AIAN: American Indian or Alaskan Native

Table 3 Study datasets

	N	Number of tasks	Number of items	Total score points
Grade 5 Form A	2739	2	11	11
Grade 5 Form B	2495	2	14	14
Grade 8 Form A	3081	2	18	18
Grade 8 Form B	2770	2	17	17

3 Dimensionality Analysis

3.1 Description of Four Datasets and Three IRT Models

Four datasets were used in the analyses, one for each form and grade. Table 3 provides the number of students who took the form, the number of tasks and items, and the total score points for each form.

Three IRT models based on content specifications were fit to the data to compare the model fitness and investigate the dimensionality of the assessment: 1) a unidimensional IRT model, 2) a three-dimensional IRT model, and 3) a testlet model. Figure 1 shows a graphic illustration of each model. All the analyses were conducted using the R mirt package (Chalmers, 2012).

3.2 Unidimensional IRT Model (Model 1)

First, unidimensional models were applied to fit the data. Three unidimensional models were examined to determine the best fit: Rasch one-parameter logistic (1PL;

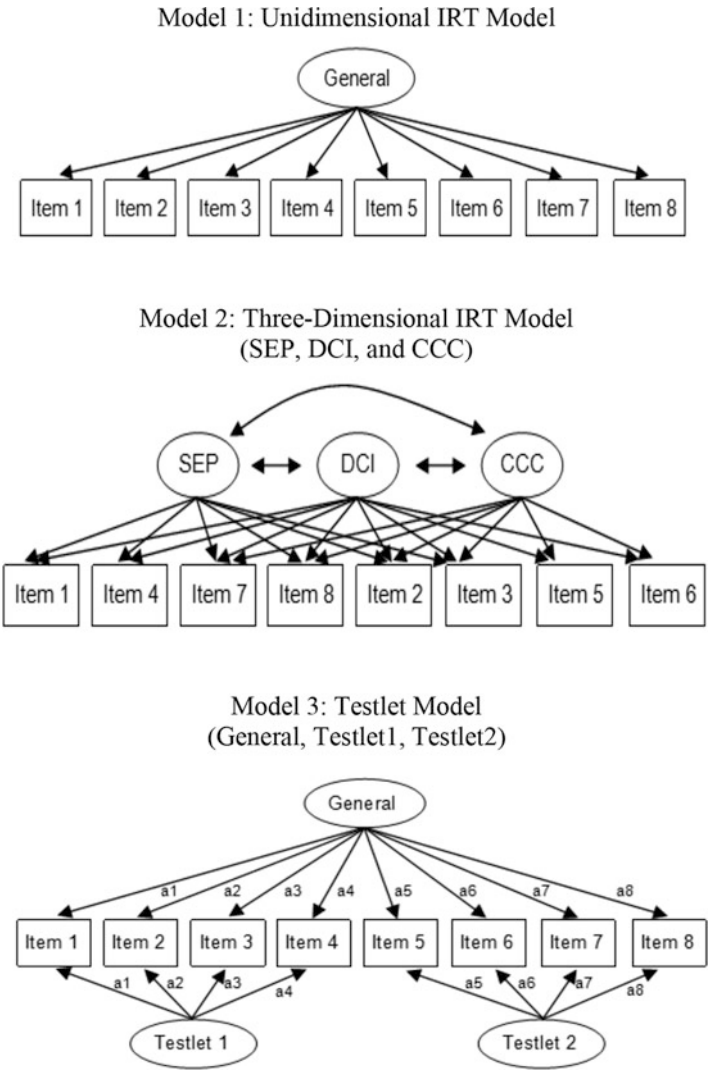

Fig. 1 Graphic illustrations of IRT Models 1, 2, and 3

Rasch, 1960), two-parameter logistic (2PL; Birnbaum, 1968), and three-parameter logistic (3PL; Lord, 1980). The equations for each model are presented below.

$$P\left(U_{ij}=1|\theta_j,b_i\right)=\frac{e^{\theta_j-b_i}}{1+e^{\theta_j-b_i}} \qquad \text{(1PL)}$$

$$P\left(U_{ij}=1|\theta_j,b_i\right)=\frac{e^{a_i\left(\theta_j-b_i\right)}}{1+e^{a_i\left(\theta_j-b_i\right)}} \quad \text{(2PL)}$$

$$P\left(U_{ij}=1|\theta_j,b_i\right)=c_i+\left(1-c_i\right)\frac{e^{a_i\left(\theta_j-b_i\right)}}{1+e^{a_i\left(\theta_j-b_i\right)}} \quad \text{(3PL)}$$

where θ_j, b_i, a_i and c_i are the person, item difficulty, discrimination, and guessing parameters, respectively.

To evaluate model fit, Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were consulted. The better-fitting model is the one with a lower AIC or BIC value. BIC penalizes model complexity more heavily than AIC, which may result in an inconsistent model preference. Table 4 presents the fitting results from the Rasch, 2PL, and 3PL models for each test form. The lowest AIC and BIC values for each dataset are bolded. Though the 3PL model fits the data best for two of the four forms as indicated by the lowest AIC and BIC values, the model has a convergence problem for Grade 8 Form B, and the BIC value indicates that the 2PL model fit better than the 3PL model for the dataset from Grade 5 Form A. Lack of convergence is an indication that the data do not fit the model well because there are too many poorly fitting observations. The 2PL model generally fits much better than the 1PL model. Though it fits the data slightly worse than the 3PL model in some cases, it does not have the same convergence problem as the 3PL model. Thus, a 2PL model was preferred and was selected as Model 1 for the study analyses.

3.3 *Three-Dimensional IRT Model (Model 2)*

Second, a three-dimensional IRT model (Model 2) was applied to fit the data. This model assumes the underlying domains as DCIs, SEPs, and CCCs. This three-

Table 4 Model-fit comparison between unidimensional 1PL, 2PL, and 3PL models

		Form A		Form B	
Grade	Model	AIC	BIC	AIC	BIC
5	Rasch 1PL	23265.85	23337.02	38991.28	39078.74
	2PL	23152.84	**23283.33**	38504.10	38667.36
	3PL	**23136.74**	23332.48	**38327.11**	**38572.01**
8	Rasch 1PL	55042.77	55157.40	51828.75	51935.45
	2PL	53889.39	54106.58	50425.67	50627.22
	3PL	**53341.93**	**53667.72**	NA[a]	NA[a]

Note: The highlighted data indicate the best-fit model
[a]NA indicates that the model did not converge

dimensional model is the multidimensional extension of the 2PL model (Reckase, 2009). The form of the model is given by

$$P\left(U_{ij}=1|\boldsymbol{\theta}_j,\boldsymbol{a_i},d_i\right)=\frac{e^{\boldsymbol{a_i}\boldsymbol{\theta}'_j+d_i}}{1+e^{\boldsymbol{a_i}\boldsymbol{\theta}'_j+d_i}}$$

where $\boldsymbol{a}$ is a $1 \times m$ vector of item discrimination parameters and $\boldsymbol{\theta}$ is a $1 \times m$ vector of person coordinates with m indicating the number of dimensions in the coordinate space (i.e., m is 3 in this case). The intercept term, d, is a scalar. The exponent of e in this model can be expanded to show how the elements of the $\boldsymbol{a}$ and $\boldsymbol{\theta}$ vectors interact.

$$\boldsymbol{a_i}\boldsymbol{\theta}'_j+d_i=a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+\cdots+a_{im}\theta_{jm}+d_i$$

The latent traits of this three-dimensional model were set to be correlated because students' abilities in these dimensions are expected to be related to some extent. The empirical results also suggest that the model fits the data better when the latent traits are set to be correlated.

3.4 Testlet Model (Model 3)

A 2PL testlet model (Bradlow et al., 1999) was also applied to fit the data. Because the pilot test was composed of testlet-based items, which may violate the local independence assumption of IRT models, a testlet model was applied to the data to examine the testlet effect. The testlet model assumes a single primary dimension (i.e., general knowledge and abilities in science) and several uncorrelated specific dimensions according to testlets (i.e., tasks) after accounting for the primary dimension. For a testlet model, an item's slope for the specific dimension is constrained to equal the item's slope for the general dimension (Cai, 2010). The 2PL testlet model is given as

$$P_j\left(\theta_i\right)=\frac{1}{1+e^{-a_j\left(\theta_i-b_j-\gamma_{id(j)}\right)'}}$$

where $p_j(\theta_i)$ is the probability of a correct response to item j for examinee i, θ_i is examinee i's latent ability, a_j and b_j are the item discrimination and difficulty parameters, and $\gamma_{\mathrm{id}(j)}$ is a person-specific testlet effect that is assumed to follow a distribution $N(0, \sigma^2\gamma\mathrm{id(j)})$.

Table 5 Model-fit comparison between Models 1, 2, and 3

			Form A		Form B	
Grade	Model description	Model #	AIC	BIC	AIC	BIC
5	Unidimensional	1	23152.8	23283.3	38504.1	38667.4
	3D (SEP, CCC, DCI)	2	**23074.4**	23317.6	**38206.9**	**38481.0**
	Testlet model	3	23127.6	**23269.9**	38464.3	38639.3
8	Unidimensional	1	53889.4	54106.6	50425.7	50627.2
	3D (SEP, CCC, DCI)	2	**53147.3**	**53521.4**	**49821.7**	**50159.6**
	Testlet model	3	53479.6	53708.8	50399.7	50613.1

3.5 IRT Model-Fit Comparisons

Model fit among Models 1, 2, and 3 was compared. Each model was applied to the four datasets. Table 5 presents the model-fit comparison results for all four datasets. The lowest AIC and BIC statistics are bolded. All the AIC and BIC statistics suggest that Model 2 fits the data best with the exception of the BIC statistics for Grade 5 Form A. Overall, Model 2 (three-dimensional IRT model) provides the best fit across all four datasets.

3.6 Item Fit Statistics

Overall, the three-dimensional IRT model (Model 2) fit the data better than the other two models. To further examine the fitness of the three-dimensional model, the chi-squared-based item-level fit index ($S\text{-}X^2$; Orlando & Thissen, 2000, 2003) was evaluated to see if the model fits the data well at the individual item level. Item fit statistics from the 2PL unidimensional model were used as a baseline for the comparison. The results from the chi-square-based item-level goodness-of-fit tests suggest that more items have bad fit (i.e., p-value <0.05) from the three-dimensional model than from the unidimensional model. For example, four items on Grade 8 Form B showed poor fit to the unidimensional model. However, for the three-dimensional model, these four items and five additional items showed poor model fit. Similar patterns were discovered for the other forms.

All four items that did not fit well to the unidimensional model were technology-enhanced items that required students to enter a short response that is scored as either correct or incorrect. It is possible that students rely on different abilities to respond to these items compared to the abilities measured by the multiple-choice items. A close look of the items by content experts is needed to identify the potential causes of item misfit.

3.7 Local Dependency Among Items Within a Task

Although the testlet model fits slightly better than the unidimensional model, the extent to which the local independence assumption is violated was examined using a popular local independence statistic, Yen's Q3 index. Index values greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997). Among the 435 item pairs across forms, only two pairs of items had a residual correlation greater than 0.20, suggesting that local item independence generally holds for all forms.

4 Discussion

In general, based on the pilot test data, the model fit statistics suggest that the three-dimensional IRT model that aligns with the DCI, SEP, and CCC dimensions (Model 2) provides slightly better overall fit than the unidimensional model (Model 1) and the testlet model (Model 3). However, the fit of the three-dimensional model at the item level is poor. Another issue to consider for this model is that the NGSS dimensions may not be conceptualized in the same manner that test score dimensionality has been conceptualized, which may create some confusion (Martineau, 2017). The use of the term "dimensionality" in NGSS may be better described as "complex" performance (Dunbar et al., 1991), which involves knowledge and skills across a number of domains or subjects.

Local independence is a fundamental assumption of unidimensional models. Fitting a unidimensional model in the presence of local dependencies may result in biased item parameters and standard errors of measurement (Yen, 1993). The American Institute of Research (AIR) applied a Rasch testlet model (Wang & Wilson, 2005) to calibrate NGSS-aligned science assessments for multiple states (Rijmen, 2018). However, for the new science assessment used in this study, the local independency assumption still generally holds and the testlet model only provides slightly better fit than the unidimensional model.

It is important to note that the data used in this study were collected from a pilot test, so the quality of some items may be low. These items may impact the model fitness results. Students' low motivation for the pilot test may also have affected the quality of the data. The relatively short test length compared to a regular state assessment limited the number of items to be administered for each dimension. All these factors may cause the structure of the pilot data to not strongly resemble the structure of data from operational assessments. It will be worth conducting the dimensionality analysis again using data from the operational test to identify the most appropriate measurement model for the assessment.

Unidimensional IRT models are widely used in testing programs. In contrast, MIRT models are rarely implemented in any state testing program due to its complexity. They require a large sample size to obtain accurate parameter estimates

and take a much longer estimation time, which pose challenges in an operational setting. The sample size of an operational test will be much larger than the sample size of this study that used pilot data. Applying a multidimensional model will significantly increase computation time. Implementing MIRT models in operation will likely be a new practice for most vendors working with states. The need for more complex measurement models needs to be further evaluated. Data from the operational test will be collected to further evaluate the need of using MIRT models and examine the robustness of the unidimensional model under various test conditions in future studies.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

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.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581–612.

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chen, W. H., & Thissen, D. (1997). Local dependence indices from item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment.* Commissioned paper presented at the K–12 Center at ITS Invitational Research Symposium on Science Assessment, Washington DC.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Martineau, J. (2017). *The intersection of measurement model, equating, and the Next Generation Science Standards*. Center for Assessment. https://www.nciea.org/sites/default/files/inline-files/Martineau_RILS%20-%20Brief%20on%20NGSS%20Measurement%20Models%20and%20Equating%20-%20Final.pdf

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academic Press. https://www.nextgenscience.org/search-standards

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research (Expanded edition, 1980. University of Chicago Press).

Reckase, M. (2009). *Multidimensional item response theory*. Springer.

Rijmen, F. (2018). *Scoring and reporting for assessments developed for the new science standards*. Paper presented at the National Conference on Student Assessment.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126–149.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.