Measuring Proficiency using Interactive Simulation Data: Empirical Comparison of Evidence Aggregation Methods

Jinnie Choi, Kristen DiCerbo, Matthew Ventura, Emily Lai

Pearson

Jim Wood, Judi Iverson

Minnesota Department of Education

Paper to be presented at the
2019 National Council on Measurement in Education Annual Meeting
at Toronto, Ontario, Canada, April, 2019.

Abstract

From a learner's interaction with a simulation, we can derive evidence of learning. In the parlance of ECD, we create a scoring model by identifying particular actions a learner takes that are related to our construct of interest. We then require a means by which to aggregate this evidence to make inferences about the knowledge, skills, and attributes of interest. This is the measurement model. In this study we empirically compared four methods that can be used to aggregate the individual pieces of evidence: item response model, cognitive diagnostic model, Bayesian networks, and percent correct scores. Results showed that all four results were positively associated with near-transfer and far-transfer scores, implying external validity of evidence identification. They were also all highly correlated with each other. We suggest that selection of evidence aggregation method should be made in part based on assessment purpose and context.

Introduction

Recently, computer-based interactive simulations have been introduced as a tool for science instruction in K-12 classrooms. Multiple studies have supported the value of interactive science simulations as effective instructional tools (e.g, Hensberry, Moore, Perkins, 2015; Moore, Chamberlain, Parson, & Perkins, 2014), and also as assessment tools. However, most simulation-based assessment relies on learner responses to questions *after* interacting with simulations. A critical gap in research exists in understanding how to make inferences about what learners know and can do through the use of rich process data produced *during* the interaction with simulation activities, potentially for additional insight and feedback.

Measuring proficiency using interactive simulation data requires careful consideration of what counts as evidence and how we combine all of the many different events that could serve as evidence in order to make inferences about the knowledge, skills, and abilities of the learner. This paper focuses specifically on evidence aggregation, or the statistical models we can use to aggregate the individual pieces of evidence that can be derived from a learner's interaction with a simulation. Among many approaches to aggregate evidence of learning (Mislevy et al., 2014), we consider four different methods: item response model, cognitive diagnostic model, Bayesian networks, and percent correct scores.

Research Questions

In this paper, we describe and compare different ways to aggregate evidence of learning during learners' use of interactive science simulations, by answering the following research questions. (a) How many latent variables explain evidence data the best? (b) How are the four methods different or the same empirically? (c) What inferences can one draw from each of the four evidence aggregation methods about learner proficiency? In what ways are findings from these approaches congruent, conflicted, or enhanced by each other?

Methods

Participants

During the 2017-2018 school year, 253 (89 5th and 164 8th grade) learners from nine schools from the state of Minnesota completed the study activities. Each learner was given login information for a computer-based test form that included five task variants about designing and optimizing a submarine, and five near-transfer traditional multiple choice items about the

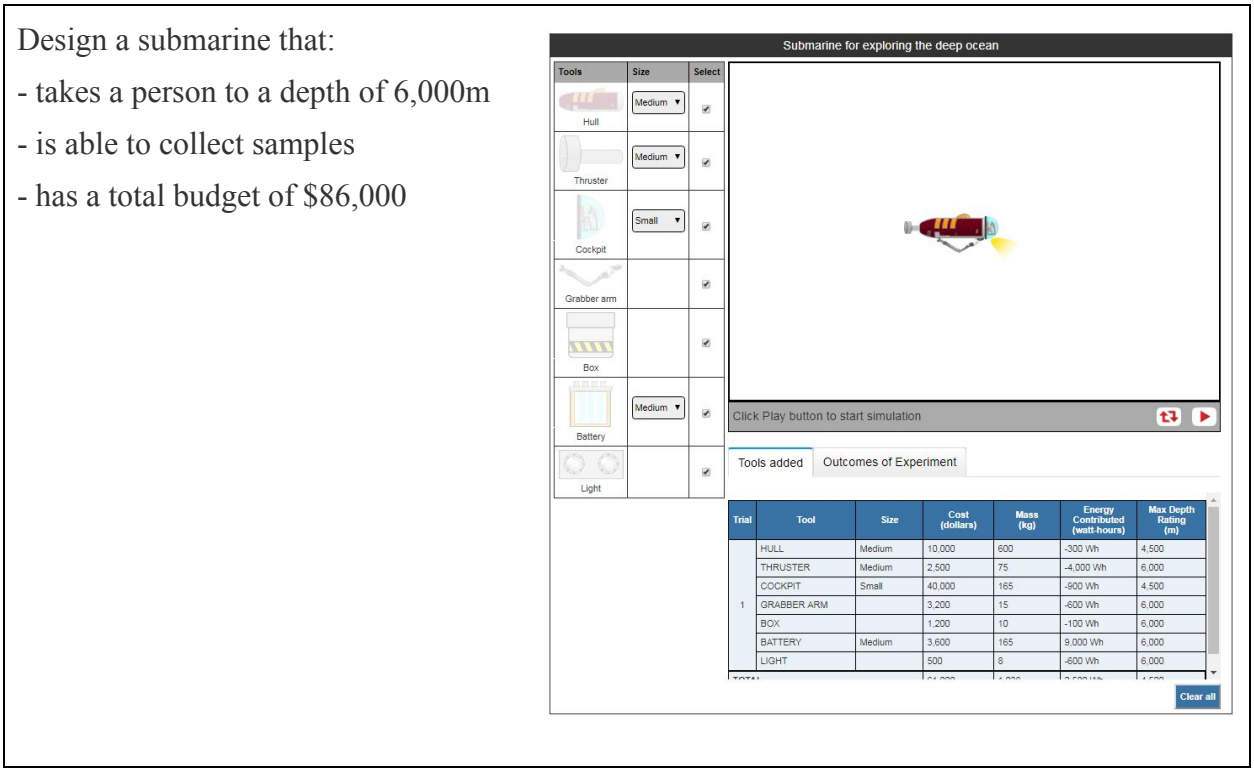
targeted standard. In addition, each learner's state assessment score in science (a composite scaled score) was collected at the end of the school year.

Materials

Interactive simulation for engineering. Students completed five task variants that involved creating and refining the design of a submarine. The tasks were designed to align with Minnesota science and engineering standards, for example is the engineering standard 6.1.2.2.1 "The learner will understand that engineering design is the process of devising products, processes and systems that address a need, capitalize on an opportunity, or solve a specific problem" (Minnesota Department of Education, 2009).

The construct that we targeted to measure was 'applying an engineering design process', with two subskills 'developing solution' and 'optimizing solution'. 'Developing solution' is what is referred to in the standard as "apply an engineering design process that includes identifying criteria and constraints, making representations" and was associated with actions such as building a successful solution without failure. 'Optimizing solution' was what is referred to in the standard as "refining the design as needed to construct a problem or system to solve a problem" and was associated with actions such as defining a design problem of a failure, testing tools in a systematic way, and not making the same mistakes in multiple attempts.

Figure 1. A screenshot of the simulation interface with a task.



Based on the targeted construct and subskills, we designed and developed the simulation interface template that could be used to create multiple task variants, the task variants themselves, and evidence pieces that would elicit evidence of proficiency/mastery of the skills. Specifically, learners were required to design and optimize a submarine for exploring the deep ocean. Figure 1 presents a task with the simulation interface. Learners used tools to build a submarine to meet desired constraints. Learners kept track of their progress on things like budget, maximum depth, and power required in a data table. When satisfied, they tested the simulation, and received notification of whether the launch was successful or not. The learners could then revise based on the results of their trial.

Evidence Identification

Each learner's interactions with the simulation activities were recorded in a .json log file. We developed a set of scoring algorithms to observe whether learners engaged in engineering practices, for example, testing different tools in a systematic way. A total of 32 pieces of evidence from a learner's interaction with a simulation were collected in the form of dichotomous indicators. Each evidence piece was coded as 1 if it was present in the interaction, 0 if it was not present in the interaction, and missing if the learner did not interact with the simulation to produce any evidence related to the task. Each of these variables were related to one of two subskills of the target construct (15 items for developing solution, 17 items for optimizing solution). Following is the list of evidence pieces for the submarine simulation.

Develop solution - 15 evidence pieces
Selects tools that are relevant before failure for each of the 5 tasks
Selects correct sizes for relevant tools for each of the 5 tasks
Selects a minimum number of tools and trials to complete the mission for each of the 5 tasks

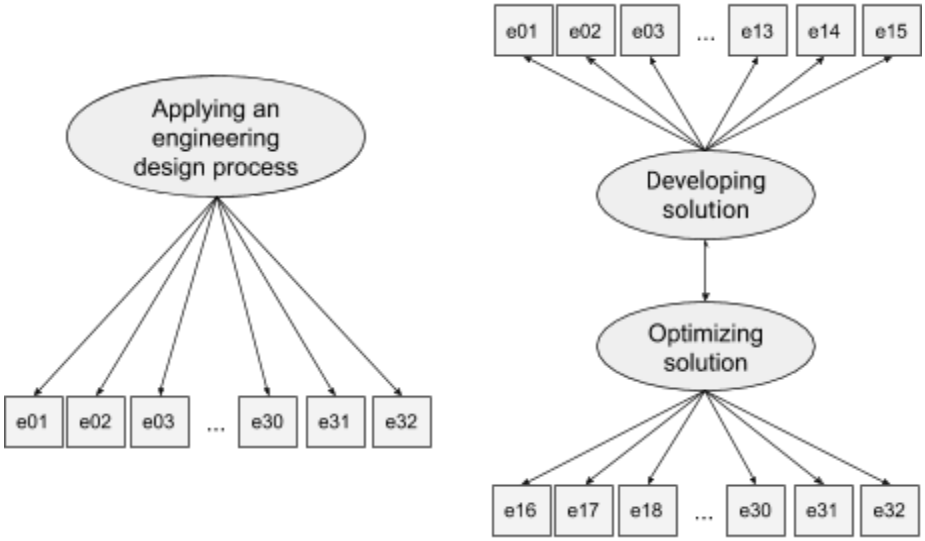
Optimize solution - 17 evidence pieces
Tests individual tools to identify the tool functions for each of the 5 tasks
Does not repeat the same error after failure for each of the 5 tasks
Selects tools that are relevant after failure for each of the 5 tasks
Explains the relevance of parts and sizes selected to solve mission (forced-choice item in a task)
Explains what parts and sizes caused a failure (forced-choice item in a task)

Competing Model Assumptions

With the constructs and evidence identified, there existed two different assumptions about the latent variables. The first assumption ('1LV') was that all 32 pieces of evidence are associated with one latent variable, 'apply an engineering design process'. The second

assumption (‘2LV’) was that 15 items are associated with one latent variable ‘developing solution’ and the remaining 17 items are related with another latent variable ‘optimizing solution’. It was assumed that the two latent variables are correlated with each other.

Figure 2. Two assumptions: one latent variable vs. two latent variables.



Statistical Methods of Evidence Aggregation

Among many approaches to aggregate evidence of learning (Mislevy et al., 2014), we explored four evidence aggregation options for dichotomous evidence data: item response model (R package ‘TAM’), cognitive diagnostic model (R package ‘CDM’), Bayesian networks (R package ‘bnlearn’), and the percent correct method. The percent correct was calculated for each learner as the sum of the scores on the evidence pieces divided by the total number of evidence pieces submitted. Below we describe typical specifications of each of the three methods that are based on latent variable modeling.

Item Response Model.

Adams, Wilson, & Wang (1997) proposed a multidimensional random coefficient multinomial logit model. In this model, the conditional probability of a response in a category *k* of the item *i* is defined as

$$P(y_{ij} = 1 : A, B, \xi|\theta) = \frac{exp(b_{ij}\theta + a'_{ij}\xi)}{\sum_{k=1}^{K_i} exp(b_{ij}\theta + a'_{ij}\xi)}$$

The relationship between items and dimensions is specified by a scoring function, such that a response in category *j* in *j* = (1, ... , *J*) to item i is represented by a V-dimensional column

vector $b_{ik} = (b_{ik1}, ..., b_{ikV})'$ where *j* is a sequential ordering of existing response category k_i for each *i*. These column vectors are then collected into a scoring submatrix B_i for item *i*, and these submatrices collected into an overall scoring matrix **B**. In this study, we used $J = 2$ for all items and tested V = 1 and V = 2 as competing hypotheses of the latent variable structure.

There are *p* item parameters, which are collected into a parameter vector $\xi = (\xi_1, ..., \xi_P)$. The relationship between item responses and parameters is specified by a design matrix **A**. This is composed of design vectors a_{ij}, each of length *p*, which, when multiplied by the parameter vector ξ, form a linear combination of the parameters operative in response *j* to item *i*. The focus of this study was the proficiency of person N, which is described as a V-dimensional vector $\theta = (\theta_1, ..., \theta_V)$ where V = 1 or 2. The marginal maximum likelihood estimation (MMLE) was used for estimating the item parameters, while empirical Bayes estimation was used to estimate the latent proficiencies.

Cognitive Diagnostic Model.

de la Torre (2011) proposed the G-DINA (generalized deterministic inputs, noisy " and" gate) model for dichotomous attributes. This model assumes that two skills a_1 and a_2 are required for mastering item *j*. The model specifies

$$g[P(y_{nj} = 1|a_n)] = \delta_{j0} + \delta_{j1}a_{n1} + \delta_{j2}a_{n2} + \delta_{j12}a_{n1}a_{n2}$$

with a link function *g*. δ_j is the item parameters of the jth item. δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to a_k and $\delta_{j1...K}$ is the interaction effect due to a_1 and a_2. The model requires a $J \times K$ Q-matrix of which the element in row *j* and column *k*, q_{jk}, is equal to 1 if the k^{th} attribute is required to answer item *j* correctly; otherwise it is equal to zero. In this study, we used $K = 1$ or $K = 2$ as competing hypotheses of the latent variable structure. The focus of this study was the mastery group classification, which is predicted to be $\widehat{M}_n = 1$ if $\widehat{a_{n1}} > \widehat{a_{n2}}$ and 0 otherwise. Weighted maximum likelihood estimation (WLE) was used to estimate the item parameters and attribute mastery probabilities.

Bayesian Networks.

A Bayesian network is a directed acyclic graph $\mathcal{G} = (V, A)$ that describes a joint probability distribution X over the sets of random variables X_i. In this study, we used $I = 1$ or $I = 2$ as competing hypotheses of the latent variable structure. Each variable corresponds to a node $v_i \in V$. The arcs $a_{ij} \in A$ represent direct correlations between the variables. The network assumes conditional independence: each node is independent of its nondescendants in the graph, given the state of its parents. Based on this assumption, a joint distribution is formed over the local conditional distributions of each of the variables given its parent.

$$P(X) = \prod_{i=1}^{p} P(X_i|\Pi_{X_i}) \text{ where } \Pi_{X_i} = \{parents\ of\ X_i\}$$

When all variables are discrete, the conditional probabilities associated with each variable are the parameters of interest. In calculating the Bayesian Dirichlet scores, the standard priors for the conditional probability table, provided in Table 1, were used for the estimation of the parameters associated with each $X_i|\Pi_{X_i}$.

Table 1. Priors for the conditional probability table.

	Mastery	Non-mastery
Correct	.75	.25
Incorrect	.25	.75

The focus of this study was the mastery classification. Bayesian network classifiers take the assumptions of Bayesian networks to solve classification problems. We used naive Bayes (Cortes & Vapnik, 1995) classifier for discrete variables which uses the posterior probability of the target variable for classification. For both CDM and BN, probability cutoff for mastery prediction was .5.

Analysis

Model selection between one (1LV) and two latent variable model (2LV).

First, before we compared the four methods, we tested two different assumptions about the latent variables. The first assumption ('1LV') was that all 32 pieces of evidence are associated with one latent variable, 'applying an engineering design process'. The second assumption ('2LV') was that 15 items are associated with one latent variable 'developing solution' and the remaining 17 items are related with another latent variable 'optimizing solution'. It was assumed that the two latent variables are correlated with each other.

In order to examine whether two latent variables are significantly different, we conducted an exploratory factor analysis using the polychoric correlation matrix to determine the number of factors. Also, we examined the latent correlation between the estimates for the two latent variables. Based on all the results above, we made a decision on either the 1LV or 2LV model.

Comparison of the four methods.

We compared the four methods using several validation approaches that include examining general model fit, and correlation of the scores with near-transfer and far-transfer scores.

General model fit. First, we examined log likelihood, deviance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) to compare the model fit of the model-based methods. AIC and BIC are calculated based on likelihood and are measures of efficiency. The model with the lowest value of AIC and BIC is preferred over other models. Percent correct was non-parametric method so it was excluded in this comparison.

Correlation between the estimates from the four methods with each other. Second, we examined the correlation between the results from the four methods. In order to compare the linear associations between the pairs of results among all four methods, we used the posterior probability estimates for the mastery, instead of discrete mastery predictions, from the CDM and BN results. To account for the non-normal distributions of the CDM and BN results, we also examined rank-based correlations.

Correlation between the estimates from the four methods and near- and far- transfer scores. Third, we evaluated how closely related the results from the four methods are with other measures of engineering performance. To do this, for the continuous results (i.e., IRT proficiency estimates and percent correct scores), we correlated the proficiency estimates with scores from the near-transfer items as well as those from the far-transfer state summative test. For the discrete results (i.e., mastery classification results from CDM and BN), we compared the near- and far-transfer score distributions between mastery and non-mastery groups.

In addition, for discrete mastery classification results from CDM and BN, we used sensitivity and specificity measures (Altman & Bland, 1994; Powers, 2011) to evaluate whether the predicted mastery was accurate when compared to the performance on the corresponding near-transfer items and far-transfer state assessments. We used a dichotomized near transfer score (1 if the near transfer score on a scale of 0 to 5 was at least 3, 0 if less than 3) and a dichotomized far-transfer score (1 if the state score was at least 50 (set by the state), 0 otherwise) to compare with the dichotomous mastery estimates from CDM and BN, respectively (1 if predicted mastery, 0 otherwise). We calculated the sensitivity and specificity measures following the formulae below.

Sensitivity (true positive rate) = true positives / (true positives + false negative)
Specificity (true negative rate) = true negatives / (true negatives + false positives)

Combining all the findings from above, we evaluated what inferences one can draw from each of the four evidence aggregation methods about learner proficiency and in what ways the findings from these approaches are congruent, conflicted, or enhanced by each other.

Results

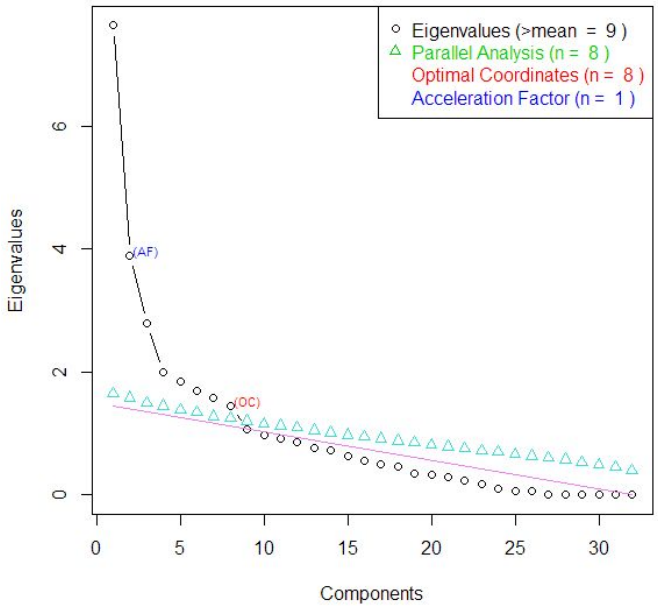
With the raw evidence data before any aggregation or statistical modeling, we first examined learners' response and success patterns. Specifically, first, we looked at the response rate: How many learners submitted at least one solution for the task and reached either success or failure (instead of skipping the task or only showing irrelevant actions)? The results showed that Tasks 1, 2, and 3 had the most responses and the number of responses decreased for Task 4 and Task 5. Second, we looked at the correctness: Within those who submitted a solution to the tasks, what percentage reached success or failure? Overall, in terms of developing a successful solution, we observed more success than failure across all five task variants. Task 5 was the most difficult task.

Figure 3. Number of responses (top) and correctness (bottom).

How Many Latent Variables Explain Data the Best?

To address the first research question, we first conducted an exploratory factor analysis to examine how many latent factors explain the data the best. The scree plot in Figure 4 supported a single factor solution, showing that from the second factor, the line is almost flat, meaning that each successive factor after the first did not account for a change in explanation of the total variance. The acceleration factor, a non-graphical solution of the scree test (Raiche, Walls, Magis, Riopel, & Blais, 2013), also indicated that the first point is where the slope of the curve changes with maximum acceleration (most abruptly).

Figure 4. Scree plot from the exploratory factor analysis showing one factor solution.



Next, using each of the four evidence aggregation methods, we calculated and estimated learners' performance on two latent variables: develop solution and optimize solution. With the three model-based methods (IRT, CDM, and BN), we imposed a structure with two latent variables and for the percent correct method we separated the evidence pieces into two groups and calculated the percent correct for each group. Table 2 shows that the resulting correlation coefficients from the latent variable models were close to one across three of the four methods, supporting the claim that the two latent variables are almost identical. The percent correct method produced a lower correlation, but this was likely due to missing data. While the three latent variable models accounted for missingness through likelihood estimation, the percent correct method had no such procedure available.

Table 2. Correlation between the estimates for two latent variables: develop solution and optimize solution.

Method	Type	Value
IRT	Latent correlation	0.901
CDM	Polychoric correlation	0.999
BN	Polychoric correlation	0.937
Percent Correct	Pearson correlation	0.452

How do assumptions in evidence model match with the one factor solution?

To examine how the single latent variable model fits the data, we conducted a principal component analysis with one factor. The model fit was good, with a root mean square error (RMSE) of .16 and $\chi^2 = 6260.04$ with $p<.001$. The indicators showed zero to positive factor loadings, with standardized loadings ranging from .00 to .71. Interestingly, the evidence pieces related to getting the correct answer without failure (i.e. correct at first try) were among the least associated with the latent factor. This result met with our expectation that designing a submarine and finding a solution is an engineering proficiency that requires more of the skills such as testing, evaluating, and refining the design as needed to construct a product or system to solve a problem; and less of getting a correct solution at first try.

Based on the results in the current section, we chose and proceeded with the one latent variable structure in the following analyses. We only present the results with one latent variable (i.e., applying an engineering design process) in the following sections.

How Are the Four Evidence Aggregation Methods Different or the Same Empirically?

General model fit.

First, we compared the general model fit of the three model-based methods by calculating log likelihood, deviance, AIC and BIC. As shown in Table 3, the model fit estimates are better for the IRT method than for the other two methods, both before and after considering the number of parameters.

Table 3. Comparison of model fit statistics.

Model	IRT	CDM	BN
logLL	-3259.07	-3376	-3951
Deviance	6518	6752	7903
Npars	33	65	32
AIC	6584	6882	7966
BIC	6701	7111	8079

Correlation between the estimates from the four methods with each other.

Second, we examined how closely the estimates from the four methods are correlated with each other. In order to compare the linear associations between the pairs of results among all four methods, we used the posterior probability estimates for the mastery, instead of discrete mastery predictions, from the CDM and BN results. Overall, as shown in Figure 5, the Pearson product-moment correlations between the four measures were positive and strong, with coefficients ranging from 0.62 to 0.99. All coefficients were significantly different from zero at

alpha = 0.001. One noticeable pattern was that the correlations between CDM and other methods were generally lower at around .6 and .7 than other correlations, and the distribution of the results looked different from others, suggesting less similarity of the CDM method to the other methods.

Figure 5. Pearson correlation coefficients between the estimates from the four methods (est_irt, est_cdm, est_bn, est_pc).

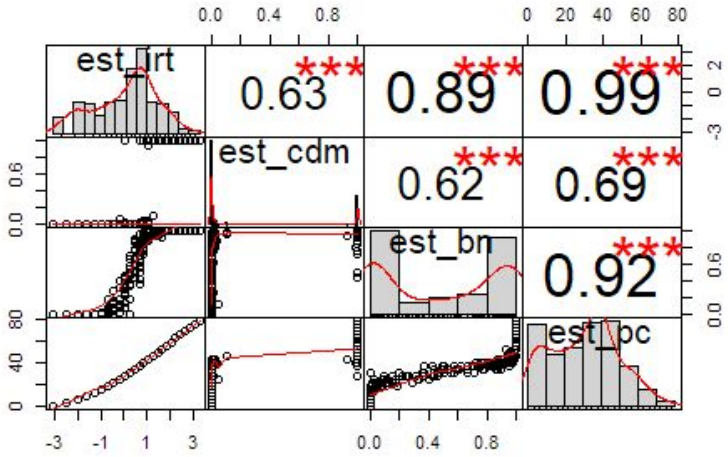


Table 4. Spearman's rho (upper diagonal) and Kendall's tau (lower diagonal) coefficients between the estimates from the four methods.

Estimates	IRT	CDM	BN	PercentCorrect
IRT		.958	.977	.998
CDM	.865		.947	.959
BN	.892	.829		.978
PercentCorrect	.983	.880	.907	

Given that the distributions of the CDM and BN probability estimates were not normal, we also calculated rank-based correlations such as Spearman's rho and Kendall's tau coefficients. The results in Table 4 suggest that the measures are strongly associated with each other and are in strong monotonic relationships. The pattern of lower correlations found for CDM from Figure 5 is less evident in rank-based correlations. This means that even though the linear association between the CDM results and other measures are not as strong, there is a strong association in terms of the increasing pattern: as CDM measure increases, other measures also increase. Overall, the correlation between the four methods are strong and positive.

Correlation between the four methods and near- and far- transfer scores.

Evidence of validity based on relations to other variables: correlations. Third, we evaluated how closely related the results from the four methods are with other measures of science and engineering performance, as evidence of validity based on relations to other variables. We first separated the two groups of learners based on their grade level (5th or 8th grade), to account for grade-level differences in the state assessments. Then we estimated correlation between the proficiency estimates and the scores from the near-transfer multiple choice items as well as those from the far-transfer state assessments.

Table 5. Pearson correlation, Spearman's rho and Kendall's tau coefficients between the proficiency estimates and near- and far-transfer scores for 5th graders (N=89) and for 8th graders (N=164).

Grade	Estimates	Pearson correlation		Spearman's rho		Kendall's tau	
		Near-transfer	Far-transfer	Near-transfer	Far-transfer	Near-transfer	Far-transfer
5	IRT	.32**	**.42*****	.29**	**.47*****	.22**	**.33*****
	CDM	.18.	**.23***	.23.	**.48*****	.17.	**.34*****
	BN	.25*	**.48*****	.33**	**.45*****	.24**	**.32*****
	PercentCorrect	.32**	**.42*****	.28**	**.47*****	.22**	**.34*****
	Near- and Far-transfer	.27*		.19.		.15.	
8	IRT	**.41*****	.35*****	**.40*****	.36*****	**.29*****	.25*****
	CDM	.22**	**.28*****	**.39*****	.37*****	**.30*****	.26*****
	BN	**.42*****	.41*****	**.42*****	.39*****	**.31*****	.26*****
	PercentCorrect	**.40*****	.35*****	**.40*****	.37*****	**.30*****	.25*****
	Near- and Far transfer	.35*****		.37*****		.28*****	

Note. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

As shown in Table 5 with Pearson correlation coefficients, all four sets of proficiency estimates captured from simulations were significantly and positively correlated with both near- and far-transfer scores for both the 5th and 8th graders. When we considered the rank-based correlation coefficients (Spearman's rho and Kendall's tau), we found similar results as shown in the right four columns in Table 5. The bolded numbers show stronger correlations.

Interestingly, these results showed a general pattern across the four methods: there were differences between grade levels in terms of how the proficiency estimates correlate with near-

or far-transfer measures. For the 5th graders, the correlations were stronger for the far-transfer scores than for the near-transfer scores, implying that the proficiencies measured by the simulation activities are more closely related with a measure of proficiency for the general science domain. For the 8th graders, the correlations were stronger for the near-transfer scores than for the far-transfer scores, implying that what is measured by the simulation activities are more closely related with a near-transfer measure of proficiency for the specific engineering practices we targeted. It is important to note that the engineering standard is actually a 6th grade standard and the 5th graders have not been taught this standard. Thus, it is not surprising to see that the results are more related to general science ability.

Because mastery classification results from CDM and BN are discrete, in addition to the traditional correlations and scatterplots, we also compared the near- and far-transfer score distributions between mastery and non-mastery groups predicted by CDM and BN. Figure 6 shows red mastery group (N=58) and blue non-mastery group (N=195) groups predicted by CDM and Figure 7 shows red mastery group (N=128) and blue non-mastery group (N=125) groups predicted by BN. As expected, for both near- and far-transfer scores, and for both 5th and 8th grades, the mastery groups predicted from CDM and BN had generally higher near-transfer and far-transfer scores than the non-mastery groups.

Evidence of validity based on relations to other variables: simple regressions. To confirm the observation from Figures 6 and 7, we examined whether the near- and far-transfer scores were significantly different for grade levels and the mastery groups predicted by CDM and BN by fitting regression models. In particular, we examined whether the predicted mastery groups have significant differences in near- and far-transfer means after controlling for grade level. Table 5 summarises the results. The results supported the finding from Figure 7: across CDM and BN results, predicted mastery groups had higher near- and far-transfer mean scores than non-mastery groups after controlling for grade level. Between CDM and BN, the estimated mean differences were greater for BN than for CDM, and the model fit was also better for BN, implying that the BN model explained the variance in dependent variables better than the CDM model did.

Figure 6. Near-transfer (on the left) and far-transfer scores (on the right), by predicted CDM mastery classifications (Grade 5 on the top, Grade 8 at the bottom).

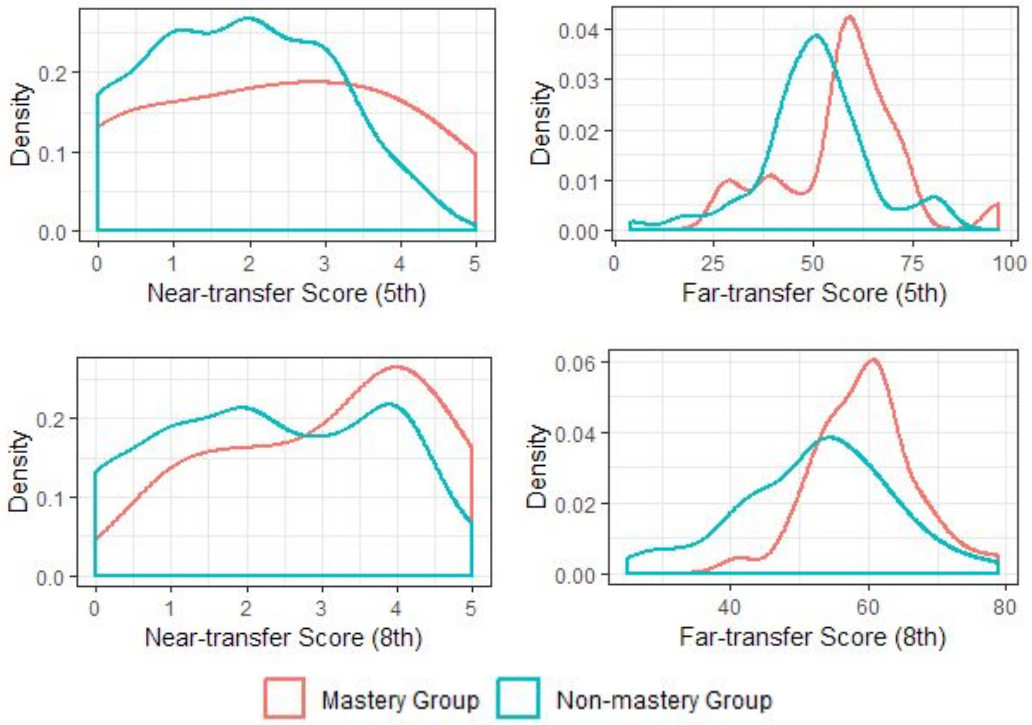


Figure 7. Near-transfer (on the left) and far-transfer scores (on the right), by predicted Bayesian Network mastery classifications (Grade 5 on the top, Grade 8 at the bottom).

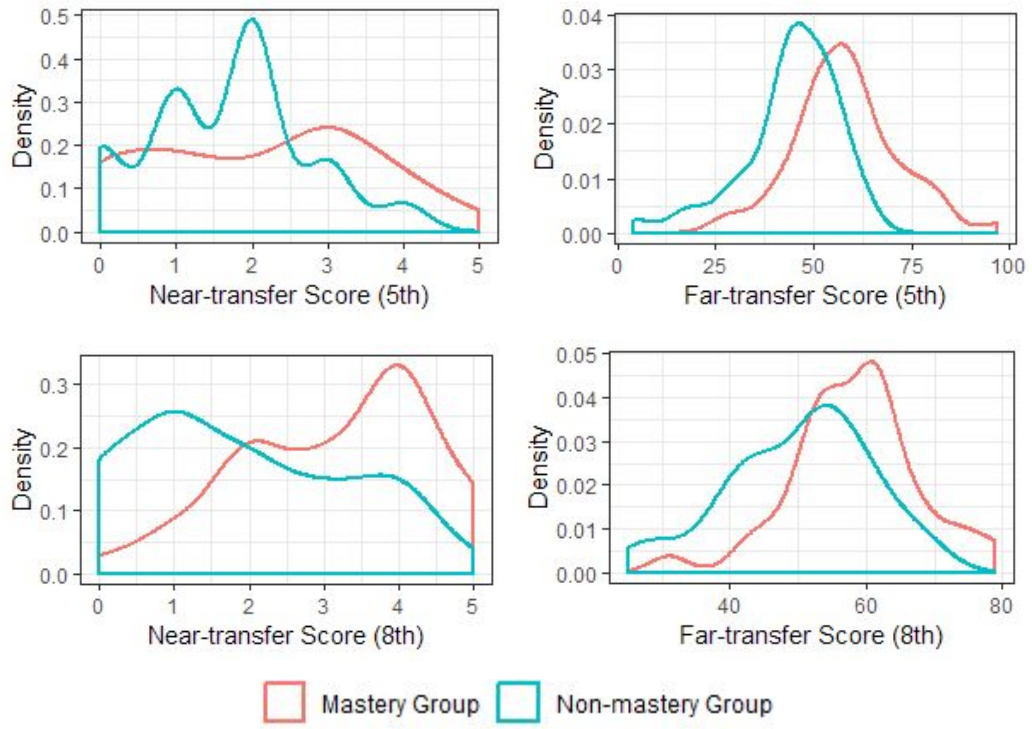


Table 6. Summary of regression results: do predicted mastery groups have significant difference in near- and far-transfer means?

Method	Dependent Variable	Independent Variables	Est (SE)	AIC
CDM	Near-transfer (0 to 5)	Grade 8 (vs. grade 5)	0.585 (0.188) **	902.83
		Mastery (vs. non-mastery)	0.690 (0.213) **	
	Far-transfer (0 to 100)	Grade 8 (vs. grade 5)	1.920 (1.574)	1978.50
		Mastery (vs. non-mastery)	7.281 (1.788) ***	
BN	Near-transfer (0 to 5)	Grade 8 (vs. grade 5)	0.680 (0.181) ***	882.04
		Mastery (vs. non-mastery)	0.990 (0.173) ***	
	Far-transfer (0 to 100)	Grade 8 (vs. grade 5)	2.896 (1.488) .	1947.70
		Mastery (vs. non-mastery)	10.162 (1.422) ***	

Note. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Evidence of reliability based on precision. Precision is related to the question of how close our estimation is to the truth. Typical measures of precision in most standardized testing are standard errors of measurement or test information function, which are based on Fisher information function that is defined for continuous outcome variables. For IRT results, we calculated expected-a-posteriori (EAP) reliability (Adams, 2005) for IRT results following the formula below.

EAP Reliability = Average of (1 - (posterior variance) / (population variance))

where posterior variance represents the uncertainty of skill estimates driven by the data and population variance represents the uncertainty of skill estimates posed by the model. The EAP reliability was .885, which indicated that data explained about 89% of the true variance of skill estimates posed by the model. We did not depend on either near- or far-transfer measure to calculate this reliability and thus we did not necessarily need to separate out the grade level groups.

Because the main outcomes from CDM and BN are discrete, other approaches were necessary. For discrete mastery classification results from CDM and BN, we used sensitivity and specificity measures (Altman & Bland, 1994; Powers, 2011) to evaluate how accurate the predicted mastery was when compared to the performance on the corresponding near-transfer items and far-transfer state assessments. We used a dichotomized near transfer score (1 if the

near transfer score on a scale of 0 to 5 was at or above average (2.3), 0 if below average) and a dichotomized far-transfer score (1 if the state score was at least 50 (set by the state), 0 otherwise) to compare with the dichotomous mastery estimates from CDM and BN, respectively (1 if predicted mastery, 0 otherwise). We calculated the sensitivity and specificity measures following the formulae shown in the methods section.

The results in Table 7 showed a couple of interesting patterns. Stronger results were bolded. First, overall, BN results were better than CDM in terms of both sensitivity and specificity. Second, both CDM and BN were better in predicting failure in far-transfer tests, than success. At the same time, both CDM and BN were better in predicting success in near-transfer tests, than failure, especially for the younger learners. CDM showed low sensitivity for predicting success in far-transfer scores from simulation results.

Table 7. Accuracy of mastery prediction compared to near- and far-transfer results

Model	Dependent Variable	Grade	Sensitivity (how well the model predicted success)	Specificity (how well the model predicted failure)
CDM	Near-transfer (success means above average)	Overall	**.600**	.586
		Grade 5	**.691**	.476
		Grade 8	.551	**.649**
	Far-transfer (success means at least 50 out of 100)	Overall	.354	**.879**
		Grade 5	.397	**.762**
		Grade 8	.331	**.946**
BN	Near-transfer (success means above average)	Overall	**.712**	.594
		Grade 5	**.815**	.471
		Grade 8	.667	**.675**
	Far-transfer (success means at least 50 out of 100)	Overall	.456	**.852**
		Grade 5	.579	**.804**
		Grade 8	.402	**.883**

A score closer to 1 indicates higher accuracy of mastery prediction compared to the other results. The results in Table 7 show that the mastery prediction based on simulation activities are mostly accurate in predicting success in the near future and failure in later assessment. For example, we can safely conclude that 81.5% of 5th grade learners who score above average in near-transfer quiz will have prediction of ‘mastery’ from BN based on the simulation activities. Similarly, 88.3% of 8th grade learners who scores below 50 points in state assessment will have prediction of ‘non-mastery’ from BN based on the simulation activities.

Concluding Remarks

In the present study, we were interested in using evidence we can observe from the activity within the simulation to make inferences about learners' knowledge and skills. We described and compared different ways to aggregate evidence of learning during students' use of interactive science simulations. The result from this study will contribute to increased understanding of different evidence aggregation methods for interactive simulation data.

We first examined the evidence related to the internal structure of the data, examining how many latent variables explain evidence data the best. Ideally based on the design of the simulation, the levels of correlation between the two hypothesized subskills were expected to be moderate (typically within .4 to .7 range), because if correlations were too strong and close to one, there is no reason to identify two separate skills; also because if correlations were too weak, close to zero, or negative, the two subskills cannot be measuring one broader construct of the engineering standard. The results from all four methods, along with ones from exploratory factor analysis, indicated that a two-dimensional assumption ('develop solution' and 'optimize solution') was not fully supported by data than an assumption for one underlying construct ('applying an engineering design process'). An implication of this result could be that in the future when we design a simulation-based assessment around this specific engineering design standard, having one broad construct rather than the two subskills may work as well as having two skills. Or, in order to separate apart the two subskills, we need to design and elicit more pieces of evidence in each subskill that might be more distinguishable from each other.

Then we examined which one among the four evidence aggregation methods works the best, and how the four methods are different or the same empirically. Table 8 summarises the comparison of the four methods in terms of ease of interpreting and communicating the results to inform teaching of engineering practice, general model fit, correlation among the four methods, predictive accuracy against near- and far-transfer performances and group differences between different grade levels.

Throughout different analyses, we observed positive and strong correlations among the results from the four methods. Among the four methods, BN showed better results in terms of model fit and relations with other variables. It was not straightforward to compare the evidence of reliability between the methods because the measure of reliability had to be defined in different ways for different types of variables. Both CDM and BN offered specific evidence of precision based on classification accuracy, which can be useful in providing diagnostic feedback for specific groups of learners.

One promising pattern was that results from all four methods were positively associated with near-transfer and far-transfer scores, implying external validity of the simulation activity for predicting these scores. For the IRT proficiency estimates and percent-correct scores, correlations with near-transfer and far-transfer scores were positive and significant. The mastery

groups predicted from CDM and BN had generally higher near-transfer and far-transfer scores than the non-mastery groups.

The finding that grade 5 results are correlated higher with far-transfer but the accuracy of mastery prediction is higher for near-transfer success can be puzzling. However, when the focus is on how well the methods predicted failure (i.e., specificity), the particular trend makes sense.

Table 8. Summary of model comparisons.

<table>
<tr><th colspan="2"></th><th>IRT (MRCML[a])</th><th>CDM (GDINA[b])</th><th>Bayesian Network[c]</th><th>Percent Correct[d]</th></tr>
<tr><td colspan="2">Outcome</td><td>Proficiency estimate (-3 to 3 logit scale)</td><td>Mastery group classification (mastery or non-mastery)</td><td>Mastery group classification (mastery or non-mastery)</td><td>Percent correct score (0 to 1)</td></tr>
<tr><td colspan="2">Easy to interpret?</td><td>Fair</td><td>Good[e]</td><td>Good[e]</td><td>Best</td></tr>
<tr><td colspan="2">Model fit (BIC)</td><td>Good</td><td>Good</td><td>Fair</td><td></td></tr>
<tr><td colspan="2">Correlation among four methods</td><td>Better</td><td>Good</td><td>Best</td><td>Better</td></tr>
<tr><td rowspan="3">Correlation with external measures (near- and far-transfer)</td><td>Overall</td><td>Positively correlated Similar to PCorrect</td><td>Mastery group means > Non-mastery group means Not as good as BN</td><td>Mastery group means > Non-mastery group means Better than CDM</td><td>Positively correlated Similar to IRT</td></tr>
<tr><td>G5</td><td colspan="4">Correlations with Near-transfer < Correlations with Far-transfer</td></tr>
<tr><td>G8</td><td colspan="4">Correlations with Near-transfer > Correlations with Far-transfer</td></tr>
<tr><td rowspan="3">Accuracy was high for predicting..</td><td>Overall</td><td>‘True’ proficiency</td><td colspan="2">Near-transfer success > NT failure
Far-transfer failure > FT success</td><td></td></tr>
<tr><td>G5</td><td></td><td colspan="2">Near-transfer failure > NT success
Far-transfer failure > FT success</td><td></td></tr>
<tr><td>G8</td><td></td><td colspan="2"></td><td></td></tr>
</table>

Note. a) Adams, Wilson, & Wang (1997); b) de la Torre (2011); c) Scutari (2010); d) non-parametric method; e) Interpretation of mastery/nonmastery from CDM and BN requires understanding of a probability cutoff.

Across the types of evidence we have looked at, empirical results did not support any clear “winner” as overall best model. The model fit indices are good for IRT; BN results correlated well with the results from other methods; CDM prediction was accurate for the 8th

grade far-transfer failures; the estimates from the four methods showed consistent results when comparing how simulation scores correlated with near- and far-transfer measures.

In designing the simulation we expected that correlations between near-transfer score and simulation estimates would be stronger than those between far-transfer score and simulation estimates. In the results, we observed that this pattern only shows for the 8th graders, implying that what is measured by the simulation activities are more closely related with a near-transfer measure of proficiency for the specific engineering practices we targeted. Given that the engineering design standard was targeted at a 6th grade level and the 5th graders have not been taught this standard, it is not surprising to observe that the 5th graders did not show the expected pattern. Another possible explanation is that the reliability of the near-transfer items (alpha = .49) can be a factor: when it is low, it is less reasonable to assume that the near-transfer items are measuring the same construct as what is measured by the simulation activities. Indeed, when we calculate disattenuated estimate of the correlation by dividing the correlation by the square root of the product of Cronbach's alpha, the results (in Table 9) suggest that simulation activities are more closely related with near-transfer results than with far-transfer results in both grade levels.

Table 9. Corrected for attenuation (to compare with Table 5): Pearson correlation, Spearman's rho and Kendall's tau coefficients between the proficiency estimates and near-transfer scores for 5th graders (N=89) and for 8th graders (N=164).

Grade	Estimates	Pearson correlation		Spearman's rho		Kendall's tau	
		Near-transfer	Far-transfer	Near-transfer	Far-transfer	Near-transfer	Far-transfer
5	IRT	**.65**	.42***	**.59**	.47***	**.45**	.33***
	CDM	**.37**	.23*	.47	**.48*****	**.35**	.34***
	BN	**.51**	.48***	**.67**	.45***	**.49**	.32***
	PercentCorrect	**.65**	.42***	**.57**	.47***	**.45**	.34***
	Near- and Far-transfer	.55		.39		.31	
8	IRT	**.84**	.35***	**.82**	.36***	**.59**	.25***
	CDM	**.45**	.28***	**.80**	.37***	**.61**	.26***
	BN	**.86**	.41***	**.86**	.39***	**.63**	.26***
	PercentCorrect	**.82**	.35***	**.82**	.37***	**.61**	.25***
	Near- and Far transfer	.71		.76		.57	

Given that the methods produced highly related results, we would also encourage consideration of the purposes of assessment when selecting the method. If the assessment is to be used as a formative measure to inform immediate instructional choices, percent correct may be sufficient. It is highly interpretable by teachers, and yields similar results to much more complex measures. Alternately, in a research context where the relationship between the evidence pieces and the latent variables is unknown, the BN can learn this relationship as new data are collected. If the goal is to predict how students might do on the summative end of year test, IRT might make sense because the IRT model had the best fit and good relationships with other measures.

The results from this study contribute to increased understanding of different evidence aggregation methods for interactive simulation data, as well as the value of interactive science simulations as a tool that supports both instruction and formative assessment. Throughout this study, we also demonstrated that we can provide some evidence of validity and reliability of the simulation-based assessment data. 1) Our design work provided evidence based on test content. We reviewed research literature in science education focusing on the targeted MN science standards. We documented our principled design process of describing the knowledge, skills, and attributes, potential work products, task features, mapping of evidence to targeted knowledge, skills, and practices to guide specification of design and analysis. 2) We examined evidence based on internal structure. The work done to answer the first research question provided evidence based on internal structure. We examined the internal structure of the data against our hypothesized models by exploratory and confirmatory factor analysis as well as examining latent correlations. 3) We examined evidence based on relations to other variables. We examined the relationship between simulation evidence and performance on external science measures. Also, we investigated the relationship between mastery predictions and near- and far-transfer results using logistic regression. 4) We examined evidence based on classification accuracy. CDM and BN showed moderate classification accuracy when compared to near- and far-transfer results. Both CDM and BN were good in predicting success in near-transfer than success in far-transfer; predicting failure in far-transfer than failure in near-transfer. 5) We examined evidence based on internal structure by looking at traditional reliability estimate for the near-transfer items and EAP reliability from IRT.

Further study will examine whether we can validate the results from this study with other science simulations, whether we can demonstrate the use of activity data that are only related to learners' skills in following instructions for training the interface use (by estimating the proficiency of 'interface use' score and using that score in interpreting the overall results), and whether we can demonstrate how results could be used to guide instructional decisions.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R Package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24.

Hensberry, K., Moore, E., & Perkins, K. (2015). Effective student learning of fractions with an interactive simulation. *Journal of Computers in Mathematics and Science Teaching*, *34*(3), 273-298.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., and & John, M. (2014). Psychometric considerations in game-based assessment. *GlassLab Report*.

Moore, E. B., Chamberlain, J. M., Parson, R., & Perkins, K. K. (2014). PhET interactive simulations: Transformative tools for teaching chemistry. *Journal of Chemical Education*, *91*(8), 1191-1197.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States*. Retrieved from https://www.nextgenscience.org/.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2018). CDM: Cognitive diagnosis modeling. R package version 6.3-45. https://CRAN.R-project.org/package=CDM

Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. R package version 2.12-18. https://CRAN.R-project.org/package=TAM

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software, 35*(3), 1-22. URL http://www.jstatsoft.org/v35/i03/.