PSYCHOMETRIKA—VOL. 89, NO. 3, 747–773
SEPTEMBER 2024
https://doi.org/10.1007/s11336-024-09971-8

A DIAGNOSTIC FACET STATUS MODEL (DFSM) FOR EXTRACTING INSTRUCTIONALLY USEFUL INFORMATION FROM DIAGNOSTIC ASSESSMENT

CHUN WANG

UNIVERSITY OF WASHINGTON

Modern assessment demands, resulting from educational reform efforts, call for strengthening diagnostic testing capabilities to identify not only the understanding of expected learning goals but also related intermediate understandings that are steppingstones on pathways to learning goals. An accurate and nuanced way of interpreting assessment results will allow subsequent instructional actions to be targeted. An appropriate psychometric model is indispensable in this regard. In this study, we developed a new psychometric model, namely, the diagnostic facet status model (DFSM), which belongs to the general class of cognitive diagnostic models (CDM), but with two notable features: (1) it simultaneously models students' target understanding (i.e., goal facet) and intermediate understanding (i.e., intermediate facet); and (2) it models every response option, rather than merely right or wrong responses, so that each incorrect response uniquely contributes to discovering students' facet status. Given that some combination of goal and intermediate facets may be impossible due to facet hierarchical relationships, a regularized expectation–maximization algorithm (REM) was developed for model estimation. A log-penalty was imposed on the mixing proportions to encourage sparsity. As a result, those impermissible latent classes had estimated mixing proportions equal to 0. A heuristic algorithm was proposed to infer a facet map from the estimated permissible classes. A simulation study was conducted to evaluate the performance of REM to recover facet model parameters and to identify permissible latent classes. A real data analysis was provided to show the feasibility of the model.

Key words: Regularized expectation–maximization algorithm, Cognitive diagnostic model, Facet map.

1. Introduction

The "Every Student Succeeds Act" emphasizes that "*High-quality assessments are essential to effectively educating students, measuring progress, and promoting equity*." For such assessments to scaffold learning, they should provide actionable diagnostic information (Kluger & DeNisi, 1996; Pellegrino et al., 2001) such as learners' mastery of learning goals and related intermediate understandings. According to theories of conceptual change and constructivism, accurate and reliable information about *both* aspects of student understanding (hereafter called "facets," which represents psychological constructs such as knowledge states, conceptual understandings, etc.) is key to targeted instruction. Unfortunately, current assessment approaches have largely remained rooted in a summative paradigm, with scoring rubrics that predominantly gradate performance (i.e., right/wrong, or partial credit) (Dadey, 2017; Lanting, 2000; Underwood et al., 2018).

A plethora of online learning platforms have emerged aiming for better-differentiated and more responsive instruction both in the classroom and at a distance. But the quality of a responsive, adaptive system hinges on an embedded assessment architecture providing feedback at a sufficiently fine-grain size. Assessments constructed under a formative paradigm and grounded in research on learner cognition would much better serve the needs of various stakeholders (e.g.,

The study is funded by IES R305D200015 and NSF EDU-CORE #2300382.

Correspondence should be made to Chun Wang, 312E Miller Hall Measurement and Statistics, College of Education, University of Washington, 2012 Skagit Ln, Seattle, WA 98105, USA. Email: wang4066@uw.edu

teachers, students, parents). However, questions remain as to how to strengthen the diagnostic capabilities of such assessments for generating a more nuanced, yet robust profile of student learning that produces actionable information. An appropriate psychometric model is indispensable in this regard. In the past two decades, cognitive diagnostic modeling (CDM) has emerged as a flexible tool to extract diagnostic information from assessments and its application has been on the rise in various domains (Bradshaw & Templin, 2014; Morphew et al., 2018; Tjoe & de la Torre, 2014) . CDMs can provide a profile on every learner, pinpointing their mastery status on multiple attributes. However, existing CDMs have limited capability for modeling both goal and intermediate understandings, handling ultra-large number of attributes, and identifying intricate attribute relationships. This paper aims to develop (1) a psychometric model to identify learners' mastery of goal attributes (i.e., accurate understanding, practice, etc.) *and* their intermediate thinking (i.e., intuitive, alternative, partial understanding, etc.) and (2) a statistical learning method to identify relations among attributes. Together these will enhance the diagnostic power of assessment and support targeted instructions (DiBello et al., 2015) .

1.1. A Facets Approach to Instruction and Learning

In the realm of research on learner cognition and learning sciences, the knowledge in pieces (KiP) theory presumes that when faced with a new task, learners draw on relevant pieces of knowledge and reasoning to construct their understanding (diSessa, 2017) . Taking this "fragmented" stance (diSessa, 1993) , the theory assumes that students' ideas consist of many quasi-independent elements. Conceptual change (i.e., learning), therefore, involves learners picking and choosing the most productive ideas and refining them to create normative concepts (diSessa, 2014a) rather than abandoning entire ideas as a coherent whole. In this view, students' naïve ideas are resources for learning to develop scientific understanding instead of roadblocks to conceptual change (Minstrell, 1989, 1991) . The KiP perspective holds strong implications for instruction. First, KiP has a small enough grainsize of analysis to allow the tracking of individual learning so instructional design can benefit from formative feedback (diSessa, 2014b; Kapon & diSessa, 2012) . Second, KiP acknowledges complex contextuality, advocating that students need to be exposed to multiple contexts so that their knowledge can be developed in each. The knowledge a learner draws on may be productive in one context but problematic in another context. One example is the notion of "property" which physics learners commonly apply to both force and energy (e.g., "One cart colliding with a second stopped cart gives [suggesting transfer of property] some force of motion to the second cart.") But whereas regarding force as a property of matter is problematic to fully understand forces and motion (FM), it is productive in the learning about energy.

Facets initially emerged from classroom research of learner's conceptual understanding in physics. What started as a list of interesting alternative ideas and reasoning related to core physics concepts, was eventually organized, roughly ranked (in terms of degrees of instructional friction) and formalized into facet clusters. Most facets are paraphrases or slight abstractions of student expressions of their conceptual understanding as recorded in the classroom and through the research of others focused on conceptual learning in science. In brief, a "facet" is an individual piece or a construction of several pieces of knowledge or reasoning strategies that the learner uses to explain a phenomenon or solve a problem. Some pieces reflect accurate understanding (i.e., goal facets) while others (i.e., intermediate facets) may be alternative intuitive ideas, partial conceptions, or even useful ideas misapplied in a context. In some literature, intermediate facets are sometimes referred to as "misconceptions," which originate from students' prior learning and lived experience with the world or in the classroom (Smolleck & Hershberger, 2011; Thompson & Logue, 2006) . For instance, in Newtonian mechanics, students' intermediate understanding of force and motion is formed by their everyday experiences in the physical world (Clement, 1987) . In this paper, we use "intermediate facet" and "misconceptions" interchangeably as they are treated the same in psychometric models, although the former carries the asset-based connotations. A Facets approach to assessment and instruction is a good starting point for the development of robust diagnostic assessment (Minstrell, 1991, 1992, 2000; Minstrell et al., 2015) .

In contrast, the "attribute" in CDM is a generic term that represents psychological constructs including skills, knowledge states, cognitive processes, and rules. One subtle difference is that different condensation rules might be more appropriate for attributes vs. facets. For instance, the DINA model, the most frequently studied among CDMs, assumes a conjunctive rule, i.e., students need to master all the required attributes to answer an item correctly. This is most appropriate if considering attributes as skills and rules. For example, in order to respond correctly to the problem "4 5/7–1 4/7", students need to master two skills: basic fraction subtraction, as well as separating out the whole number subtraction from fraction subtraction. On the other hand, facets represent small identifiable units of knowledge/reasoning that are building blocks of core knowledge within a content domain. As a result, an "additive" rather than conjunctive rule may be more appropriate, as it represents how different pieces/blocks of knowledge combine to build complete understanding of a particular concept.

Although mathematically we do not differentiate facets from attributes, they have different philosophical orientations and hence differ in their formative value for instruction. Attributes, primarily rooted in a summative paradigm of assessment, are useful for describing mastery or non-mastery of skills that constitute the highest stage of learning. However, they reveal little to nothing about the pathways (of conceptual development) leading to that mastery. Therefore, while attributes can reveal students' *static* state of skill mastery to inform the general focus of subsequent instruction, they are less useful for pinpointing specific conceptual needs or for really promoting conceptual development. In contrast, facets stem from a formative paradigm, aimed at revealing the *dynamics* of conceptual learning and places of instructional opportunity. Moreover, recent studies have shown that both goal and intermediate understandings may coexist (Stavy et al., 2006) , even after a conceptual change has occurred or when a student responds correctly to an assessment item (Foisy et al., 2015; Potvin et al., 2015; Vosniadou & Verschaffel, 2004) . Being aware of the pieces of understanding that students are drawing together to explain a phenomenon or solve a problem is essential to providing instruction and feedback that are responsive to learners' needs.

1.2. Psychometric Models for Scoring

Currently, facet-based assessments are scored by computing subscores, that is, tallying the number of times a student chooses an option that measures each facet. However, this approach not only ignores the inherent probabilistic nature of students responding to items but is also unreliable when a facet is measured by a small number of items (Cizek et al., 2004; Haberman et al., 2009; Tate, 2004) . Since measurement errors are rarely reported for subscores, the decisions made therefrom may be invalid if errors are large. Alternatively, ordered multiple-choice (OMC) scoring (Briggs et al., 2006; Hadenfeldt et al., 2013) is a psychometrically better justified, flexible tool that overcomes these limitations. With OMC, each response option is linked to discrete, developmental levels of student understanding. Using the attribute hierarchy method (AHM), one can map out the facet mastery profile of each student with goal facets lined up along a *linear* hierarchy defined by the learning progression. However, the linear hierarchy assumption of OMC scoring is overly restrictive as it precludes the possibility that students may hold multiple intermediate facets simultaneously. In addition, certain intermediate facets may be at the same developmental levels such that strictly ordering them along a liner progression is too simplistic. In fact, research has shown that certain learning progressions, such as "forces and motion" in physics, are hard to map using the AHM framework (Alonzo & Steedle, 2009) . Instead, CDM

TABLE 1.
Three different modeling frameworks for modeling learning.

Framework	Foundation	Pros (no shading) and Cons (shaded in gray)
Knowledge space theory (Doignon & Falmagne, 1999)	Mathematical set theory	Adaptively assess student knowledge and provide advice for further study Deterministic theory that does not naturally account for stochastic human behavior Does not provide item characteristic information
Rule space model (Tatsuoka, 1983, 2009)	Pattern recognition	Non-parametric Difficult to make item level inferences due to the aggregate nature of the process Only allows for non-compensatory attribute relationships
Cognitive diagnostic models	Statistical measurement model	Probabilistic and flexible (i.e., accommodate dichotomous and polytomous responses, as well as dichotomous, polytomous, and continuous attributes) Allows for statistical check of model data fit; allows for data-driven estimation of attribute relationships Requires sophisticated estimation algorithm for complex models Requires high computation power for models with large number of attributes

is a more versatile psychometric tool than AHM, allowing facets to connect more freely to reflect complex, and potentially nonlinear, learning pathways.

Table 1 compares CDM with two other popular modeling frameworks for diagnostic assessment: knowledge space theory, and the rule space model wherein AHM is located. Compared with these, CDM is not only capable of handling the probabilistic nature of human behaviors such as allowing slipping and guessing but is also immensely flexible to accommodate various cognitive processes (e.g., compensatory vs. conjunctive) and outcome spaces (e.g., dichotomous, polytomous, and nominal).

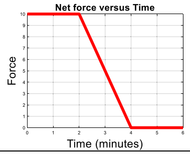
Within the CDM framework, we propose a new model, namely, the diagnostic facet status model (DFSM), which will handle nominal outcome space to maintain the information encapsulated in each response option. Meanwhile, it will also provide a fuller cognitive profile that covers both goal and intermediate facets. Existing CDMs (see Table 2) can only accommodate one of these two flexibilities. Specifically, the widely used generalized deterministic input noisy 'and' gate model (GDINA, de la Torre, 2011) and loglinear CDM model (Henson et al., 2009) only han-

TABLE 2.
Unique features of DFSM compared to other CDMs.

CDM approach	A	B	C	D	E
Generalized Deterministic Input Noisy ‘And’ Gate model (GDINA); Loglinear CDM	Y		Y		Y
Scaling Individuals and Classifying Misconceptions (SICM)		Y		Y	
Multiple choice-DINA	Y		Y	Y	
Generalized diagnostic classification model for MC option-based scoring (GDCM-MC)	Y	Y		Y	
Simultaneously identifying skills and misconceptions model (SISM)	Y	Y	Y		
DFSM	Y	Y	Y	Y	Y

A. model goal facets; B. model intermediate facets; C. scale well to large number of facets using efficient algorithm; D. model option level response (i.e., nominal response); and E. explore facet relationships.

The graph at right shows the net force (sum of all the forces) on the block during different time intervals as Jahmelia moves the block across the table. When was the block moving at constant speed?



A. From 0 to 2 minutes, when the net force is 10.
B. From 2 to 4 minutes, when the net force is decreasing.
C. From 4 to 6 minutes, when the net force is 0.
D. Both from 0 to 2 minutes and 4 to 6 minutes when the net force is constant

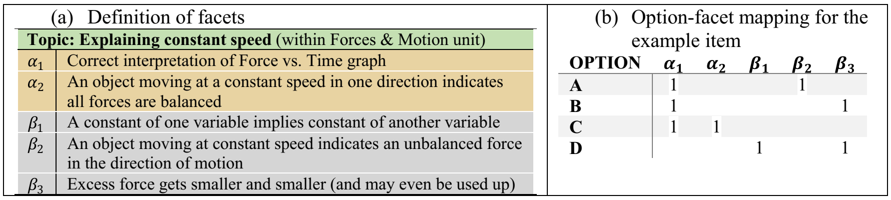
FIGURE 1.
An example item from “Forces and Motion” Unit.

dles goal facets and dichotomously (i.e., right or wrong) or polytomously scored responses (i.e., partial credits); the scaling individuals and classifying misconception (SICM, Bradshaw & Templin, 2014) model models option level responses and only intermediate facets; the multiple-choice DINA also models option level responses but only goal facets; the simultaneously identifying skills and misconceptions model (SISM, Kuo et al., 2018) models both goal and intermediate facets but at item level instead of option level hence it is only suitable for dichotomously scored items. One notable exception is the generalized diagnostic classification model for multiple choice option-based scoring (GDCM-MC, DiBello et al., 2015), which satisfies the needs of modeling goal and intermediate facets at item optional level. However, GDCM-MC requires a three-value coding scheme per response option, placing higher requirements on experts and hence a greater chance of human-coding errors. Furthermore, GDCM-MC is estimated by Bayesian Markov chain Monte Carlo (MCMC) algorithm, which does not scale well to large numbers of facets. DFSM requires less complicated coding of a Q-matrix and has more compact parameterization which affords use of the faster expectation–maximization (EM) algorithm. As detailed in the next section, the EM algorithm with a proper penalty term scales well to high-dimensional space.

2. Method

Before introducing DFSM, we first present an example item that motivates the proposal of DFSM. Figure 1a provides an example item from a topic of “Explaining constant speed” in the “Forces and Motion (FM)” unit in Diagnoser[1] (Thissen-Roe et al., 2004) , which is a set of online assessment and instruction tools designed to elicit and develop students thinking to arrive at deeper conceptual understanding. Figure 1 provides an example item from a topic of “Explaining constant speed” in the “Forces and Motion (FM)” unit. Figure 2a presents the definitions of facets from

[1] http://www.diagnoser.com/



(a) Definition of facets

Topic: Explaining constant speed (within Forces & Motion unit)	
α_1	Correct interpretation of Force vs. Time graph
α_2	An object moving at a constant speed in one direction indicates all forces are balanced
β_1	A constant of one variable implies constant of another variable
β_2	An object moving at constant speed indicates an unbalanced force in the direction of motion
β_3	Excess force gets smaller and smaller (and may even be used up)

(b) Option-facet mapping for the example item

OPTION	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
A	1			1	
B	1				1
C	1	1			
D			1		1

FIGURE 2.
Illustration of facet definition and option-to-facet mapping.

this FM topic, and Fig. 2b shows the mapping of item response options from the example item to the defined goal and intermediate facets. As shown, option C is the correct answer for this item, and this option measures both the goal facets but none of the intermediate facets. In contrast, all distractors measure one or multiple intermediate facets and some (or none) goal facets.

2.1. Diagnostic Facet Status Model (DFSM)

DFSM models the log-odds of student i with a facet profile, $\left(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\right)$, choosing a distractor of item j against a correct answer, as follows:

$$\log\left(\frac{P\left(x_{ij}=k\right)}{P\left(x_{ij}=K\right)}\right)=\lambda_{j,0}+\boldsymbol{\lambda}_{j,1}\left[\mathbf{h}\left(\boldsymbol{\alpha}_i,\mathbf{q}_{j,k}^{g}\right)-\mathbf{h}\left(\boldsymbol{\alpha}_i,\mathbf{q}_{j,K}^{g}\right)\right]^T+\boldsymbol{\lambda}_{j,2}\mathbf{h}\left(\boldsymbol{\beta}_i,\mathbf{q}_{j,k}^{p}\right)^T, \quad (1)$$

where $k = 1, .., K$ denotes different response options and K is the correct response without loss of generality. $\lambda_{j,0}$ is the intercept representing the logit of selecting a distractor over a correct response for someone with none of the facets. If $\lambda_{j,0} = 0$, then the probability of selecting the key is the same as selecting any other distractors, i.e., any response option has an equal chance to be selected. If $\lambda_{j,0} < 0$, then the chance of selecting the key for students with none of the facets is higher than that of selecting any other distractors, and smaller $\lambda_{j,0}$ leads to a higher probability of selecting the key, hence it implies that the item is easy. In contrast, if $\lambda_{j,0} > 0$, then the probability of selecting the key is lower than that of selecting any of the distractors. $\boldsymbol{\lambda}_{j,1}$ and $\boldsymbol{\lambda}_{j,2}$ are row vectors of slopes quantifying the effect of having additional goal and intermediate facets on the logit. They are constrained to be positive. Larger slopes indicate better discrimination, i.e., the probability of endorsement pattern will differ more for students with and without the relevant facets.$\mathbf{q}_j^g$ and $\mathbf{q}_j^p$ are the binary vectors denoting the mapping of response k on goal and intermediate facets, respectively. Both $\boldsymbol{\lambda}_{j,1}^T\mathbf{h}\left(\boldsymbol{\alpha}_i,\mathbf{q}_{j,k}^g\right)$ and $\boldsymbol{\lambda}_{j,2}^T\mathbf{h}\left(\boldsymbol{\beta}_i,\mathbf{q}_{j,k}^p\right)$ are linear combinations of facet main effects and interactions.

In Eq. 1, if only the main effects are considered, the term $\boldsymbol{\lambda}_{j,1}^T\mathbf{h}\left(\boldsymbol{\alpha}_i,\mathbf{q}_{j,k}^g\right)$ can be spelt out as

$$\sum\nolimits_{m=1}^{G}\lambda_{j,1,m}q_{j,k,m}^{g}\alpha_{i,m} \quad (2)$$

where G is the total number of goal facets, $\lambda_{j,1,m}$ is the main effect due to α_m. Similar forms are used for$\boldsymbol{\lambda}_{j,2}^T\mathbf{h}\left(\boldsymbol{\beta}_i,\mathbf{q}_{j,k}^p\right)$ as well. We constrain all coefficients λ to be non-negative, which means a student with more goal facets not matched with the response option will be less likely to choose a distractor over the correct option, whereas a student with more matched problematic facets will be more likely to choose a distractor. Comparing Eq. 2 to the LCDM model or the GDINA model, note that we do not include an intercept term. This is because the intercept is absorbed in $\lambda_{j,0}$ in

Eq. 1 automatically. Like in GDINA and LCDM, both $\boldsymbol{\lambda}_{j,1}^{T}\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right)$ and $\boldsymbol{\lambda}_{j,2}^{T}\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)$ can also include interaction terms among facets as needed.

Given the logit form defined in Eq. 1, the conditional probability that a student i with a facet profile, $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$, chooses distractor option k for item j is

$$\begin{aligned} &P(X_{ij} = k|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\lambda}_j) \\ &= \frac{\exp\left(\lambda_{j,0}+\boldsymbol{\lambda}_{j,1}\left[\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right) - \mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,K}^{g}\right)\right]^T +\boldsymbol{\lambda}_{j,2}\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)^T\right)}{1+\sum_{k=1}^{K-1}\exp\left(\lambda_{j,0}+\boldsymbol{\lambda}_{j,1}\left[\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right) - \mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,K}^{g}\right)\right]^T +\boldsymbol{\lambda}_{j,2}\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)^T\right)}, \end{aligned} \quad (3)$$

whereas the conditional probability that a student i chooses a correct response option for item j is

$$\begin{aligned} &P(X_{ij} = K|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\lambda}_j) \\ &= \frac{1}{1+\sum_{k=1}^{K-1}\exp\left(\lambda_{j,0}+\boldsymbol{\lambda}_{j,1}\left[\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right) - \mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,K}^{g}\right)\right]^T +\boldsymbol{\lambda}_{j,2}\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)^T\right)}. \end{aligned} \quad (4)$$

Remark. To provide an intuition about DFSM, let us consider a simple item example with 2 goal facets and 2 intermediate facets. For an item with 4 response options, the Q-matrix takes the form of [1, 1, 0, 0; 0, 0, 1, 0; 0, 0, 0, 1; 0, 0, 1, 1]. Each row refers to one response option and the first option is the correct response. The four columns refer to α_1, α_2,β_1, and β_2, respectively. Set $\boldsymbol{\lambda}_{j,1} = (1.5, 2)$ and $\boldsymbol{\lambda}_{j,2} = (0.75, 1.25)$, then the probability of responding to each option given a facet profile is presented in Table 14. Several conclusions can be drawn from Table 14. First, when the facet profile is (1, 1, 0, 0), implying that a student masters all required goal facets and none of the intermediate facets, the chance of selecting the key is the highest (i.e.,.909), whereas the chance of selecting any distractor is the same (i.e.,.030). The chances of selecting distractors are not always the same, they are the same in this case because none of the distractors measure any goal facets. Second, when a student masters all required goal facets and one (or several) intermediate facet, the chance of selecting the key drops understandably. For a student with a facet profile of $(\boldsymbol{\alpha}, \beta)$, and for an item with a q-vector (i.e., row vector) denoted by $\mathbf{q}_k$ for option k. If $(\boldsymbol{\alpha}, \beta) - \mathbf{q}_{k_1} = (\boldsymbol{\alpha}, \beta) - \mathbf{q}_{k_2}$, then the chance of selecting option k_1 and k_2 will be the same. Third, if the slope on a facet is high, then possessing that facet will lead to a higher chance of selecting an option that measures the same facet. For instance, in Table 14, because the slope on α_2 is higher than the slope on α_1, the third facet profile yields a higher chance of selecting the key (i.e., .69) than the second facet profile. The same conclusion applies to intermediate facets as well.

Note that in Eq. 1 $\sim$ 3, the λ's are defined at *item* level instead of response *option* level, hence there is no subscript k. This choice is made because of two reasons: (1) the model is more parsimonious in general; (2) if for all items, each facet is measured by only one response option (that is, if one computes the column sums of a q-matrix for an item, the column sum vector only has 0's and 1's in it), then whether or not adding the subscript k will lead to the same model. We tried to fit the two versions of the DFSM model, one with item-level intercepts and slopes and the other with item-option level intercepts and slopes, on the real data set and noted that the latter model did not converge. However, we provide estimation code for both models on https://github.com/wang4066/HARLI1 and readers may consider this more flexible version if they have enough data to support the model estimation.

2.2. *Regularized EM Algorithm*

For model estimation, the main idea of the regularized EM will proceed as follows assuming $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ reside in the same space. Let $\boldsymbol{\lambda}$ denote the set of item parameters, the form of $\mathbf{h}$ is specified in advance (i.e., additive form for instance), and let π denote the vector of $\pi_{\boldsymbol{\alpha},\boldsymbol{\beta}}$'s representing the latent class mixing proportions. Then the marginal log-likelihood of $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ given a response matrix $\boldsymbol{Y}$ is

$$logL_N(\boldsymbol{\lambda}, \boldsymbol{\pi}|\mathbf{Y}) = \sum\nolimits_{i=1}^{N} \log \left[\sum\nolimits_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^K} \pi_{\boldsymbol{\alpha},\boldsymbol{\beta}} \prod\nolimits_{j=1}^{J} \prod\nolimits_{k=1}^{K} \theta_{j,k}^{y_{ij}=k} \right], \tag{5}$$

where N is the sample size, and $\theta_{j,k}$ is the conditional probability defined in Eq. 3 (or 4) which is a function of $\boldsymbol{\lambda}$. A regularized EM (REM) algorithm with log-penalty on π_l is proposed so that the mixing proportion associated with impermissible latent classes will shrink strictly to 0. In the REM algorithm, the objective function that will be maximized takes the following form

$$\text{arg}max_{\boldsymbol{\theta},\boldsymbol{\pi}} \left[logL_N(\boldsymbol{\lambda}, \boldsymbol{\pi}|\mathbf{X}) + \gamma \sum\nolimits_{l}^{2^M} \log_{\rho_N}(\pi_l) \right], \text{subject to the constraint of} \sum\nolimits_{l}^{2^M} \pi_l{=}1. \tag{6}$$

In Eq. 6, M is the total number of goal and problematic facets. $\log_{\rho_N}(\pi_l) = \log(\pi_l) \times I(\pi_l > \rho_N)$ $+ \log(\rho_N) \times I(\pi_l \le \rho_N)$ where $\rho_N \approx \frac{1}{N}$ is a small threshold parameter to suppress singularity issue of the log function (Gu & Xu, 2019; Wang, 2021a) . $\gamma\epsilon(-\infty, 0)$ is a tuning parameter and smaller γ yields more sparsity in $\boldsymbol{\pi}$. At the convergence of the REM algorithm, $\hat{\pi}_l$ will be compared to ρ_N and only classes with $\hat{\pi}_l > \rho_N$ are retained as permissible classes. Details of the REM algorithm for the proposed DFSM model are given below.

2.3. *E-step*

In the E-step, one computes the conditional expectation of complete data log-likelihood, $l_N(\boldsymbol{\lambda}, \boldsymbol{\pi}|\mathbf{X}, \alpha, \beta)$, with respect to the posterior distributions of $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ $(i = 1, ..., N)$ .First, write the complete data log-likelihood as follows,

$$logL_N(\boldsymbol{\lambda}, \boldsymbol{\pi}|\mathbf{X}, \alpha, \beta) = \sum\nolimits_{i=1}^{N} \left\{ \sum\nolimits_{j=1}^{J} \sum\nolimits_{k=1}^{K_j} I\left(x_{ij} = k\right) \log \theta_{jk,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i} + \log f(\boldsymbol{\alpha}_i \boldsymbol{\beta}_i|\boldsymbol{\pi}) \right\}, \tag{7}$$

where $\theta_{jk,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}$'s are conditional probabilities defined in Eq. 3 or 4. Then its expectation is,

$$\begin{aligned} & E_{\boldsymbol{\alpha},\boldsymbol{\beta}}[logL_N(\boldsymbol{\lambda}, \boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ & = \sum\nolimits_{i=1}^{N} \left\{ \sum\nolimits_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^M} P_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}^{r} \left[\sum\nolimits_{j=1}^{J} \sum\nolimits_{k=1}^{K_j} I\left(x_{ij} = k\right) \log \theta_{jk,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i} + \log f(\boldsymbol{\alpha}_i \boldsymbol{\beta}_i|\boldsymbol{\pi}) \right] \right\} \\ & = \sum\nolimits_{j=1}^{J} \left[\sum\nolimits_{i=1}^{N} \sum\nolimits_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^M} P_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}^{r} \sum\nolimits_{k=1}^{K_j} I\left(x_{ij} = k\right) \log \theta_{jk,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i} \right] \\ & + \sum\nolimits_{i=1}^{N} \left[\sum\nolimits_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^M} P_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}^{r} \log \pi_l \right] \end{aligned} \tag{8}$$

where $P_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}^{r} \equiv P_{i(\boldsymbol{\alpha},\boldsymbol{\beta})\in l}^{r} = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l, \boldsymbol{\beta}_i = \boldsymbol{\beta}_l|\boldsymbol{\lambda}^r, \boldsymbol{\pi}^r)$ is the posterior distribution of $(\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l)$ for person i given the parameter estimates from the rth iteration.

Let $H_j^r \equiv \sum_{i=1}^{N} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^M} P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})} \sum_{k=1}^{K_j} I\left(x_{ij}=k\right)\log\theta_{jk,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}$ for notational simplicity hereafter, and hence Eq. 8 is simplified as follows

$$E_{\boldsymbol{\alpha},\boldsymbol{\beta}}[logL_N(\boldsymbol{\lambda},\boldsymbol{\pi}|\mathbf{X},\alpha,\beta)] = \sum_{j=1}^{J} H_j^r + \sum_{i=1}^{N}\left[\sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\{0,1\}^M} P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}\log\pi_l\right]. \tag{9}$$

Given that the last term in Eq. 9 is unrelated to $\boldsymbol{\lambda}$, it will be considered separately.

2.4. M-step

In the M-step, we will maximize $E_{\boldsymbol{\alpha},\boldsymbol{\beta}}[l_N(\boldsymbol{\lambda},\boldsymbol{\pi}|\mathbf{X},\alpha,\beta)] + \gamma\sum_l^{2^M}\log_{\rho_N}(\pi_l)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, respectively. Specifically, let $P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})} \equiv P^r_{il}$ for notational simplicity, the objective function of π_l that needs to be maximized becomes,

$$logL_N(\boldsymbol{\pi}) = \sum_{i=1}^{N}\left[\sum_l^{2^M} P^r_{il}\log\pi_l\right] + \gamma\sum_l^{2^M}\log_{\rho_N}(\pi_l) + s(1-\sum_l^{2^M}\pi_l) \tag{10}$$

where s is a Lagrange multiplier, and the last term in Eq. 10 reflects the constraint that $\sum_l^{2^M}\pi_l = 1$. Then

$$\frac{\partial logL_N(\boldsymbol{\pi})}{\partial\pi_l} = \frac{\gamma+\sum_{i=1}^{N}P^r_{il}}{\pi_l} - s = 0,$$

such that $\pi_l = \frac{\gamma+\sum_{i=1}^{N}P^r_{il}}{s}$. The key is to derive the term s. Since $\sum_l^{2^M}\pi_l = 1$, we have $\sum_l^{2^M}\left(\frac{\gamma+\sum_{i=1}^{N}P^r_{il}}{s}\right) = 1$, and hence $s = \sum_{l=1}^{2}{}^{M}\left(\gamma+\sum_{i=1}^{N}P^r_{il}\right)$. Based on the derivation, $\pi_l$will be updated using the following two steps.

1. Let $\nu_l^{(r)} = \max\left\{c, \gamma+\sum_{i=1}^{N}P^r_{il}\right\}$, where $c > 0$ is a pre-specified value and $P^r_{il} \equiv P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})}$ where $(\boldsymbol{\alpha},\boldsymbol{\beta})$ belongs to the lth group.
2. $\pi_l^{(r)} = \frac{\nu_l^{(r)}}{\sum_{l=1}^{L}\nu_l^{(r)}}$, where L denotes the total number of latent classes.

Here, "c" was chosen to be a very small value, in our case, $c = .00001$. Like Gu and Xu (2019) as well as Wang (2021b), "c" was prespecified and included to avoid negative counts. Note that $\gamma < 0$, depending on the value of γ, it is likely that $\gamma+\sum_{i=1}^{N}P^r_{il} < 0$. This is impermissible as $\Delta_l^{(r)}$ denotes the expected number of people in latent class l which cannot be negative. So, adding "c" is a numeric trick to make the algorithm stable, it will not affect the results as it is set to be very small. Each item's parameters will be updated by maximizing H_j^r separately. The full REM algorithm is presented in Table 3.

2.5. Model Identifiability

The DFSM is identified when the model parameters can be uniquely estimated given the observed responses. The identifiability of DFSM can be studied using the conclusions established by Liu and Culpepper (2023). They derived the strict and generic identifiability conditions for restricted latent class models (RLCM) for nominal response data. Because DFSM is a special CDM for nominal responses, their conclusions can be generalized to our context. Specifically, we first need to construct a Δ-matrix for an item from its Q-matrix. For example, for an item j with 4

TABLE 3.
The REM algorithm for DFSM.

1. Initialize all model parameters. Set $r = 1$.

2. Compute $P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})\in l} = \frac{L(\mathbf{x}_i|\boldsymbol{\lambda}^r,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l,\boldsymbol{\beta}_i=\boldsymbol{\beta}_l)\pi^r_l}{\sum_{l=1}^{L} L(\mathbf{x}_i|\boldsymbol{\lambda}^r,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l,\boldsymbol{\beta}_i=\boldsymbol{\beta}_l)\pi^r_l}$ for $i = 1, \ldots N$.

3. Obtain item parameters by maximizing H^r_j for each item separately, for $j = 1, \ldots J$ using the 'optim' function in R [One may choose "L-BFGS-B" algorithm, this step can be parallelized as maximizing each H^r_j is independent of the others.]

$$H^r_j \equiv \sum_{i=1}^{N} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in l} P^r_{i(\boldsymbol{\alpha},\boldsymbol{\beta})\in l} \sum_{k=1}^{K_j} I\left(x_{ij} = k\right) \log P(x_{ij} = k|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l, \boldsymbol{\beta}_i = \boldsymbol{\beta}_l, \boldsymbol{\lambda}^r_j)$$

4. Compute $\pi^{(r)}_l = \frac{\nu^{(r)}_l}{\sum_{l=1}^{L} \nu^{(r)}_l}$, where $\nu^{(r)}_l = \max\left\{c, \gamma + \sum_{i=1}^{N} P^r_{il}\right\}$ and set c=0.0001.

5. Repeat steps 2-4 until convergence. The convergence criterion could be the maximum difference between all parameter estimates (including both item parameters and class proportions) in the last two consecutive EM steps is smaller than.0001.

6. After convergence, compare $\pi^{(r+1)}_l$ to ρ_N and if $\pi^{(r+1)}_l < \rho_N$, set $\pi^{(r+1)}_l = 0$, g_l =0. Otherwise, g_l =1. And g_l is an indicator indicating whether latent class l is permissible. Re-standardize all $\pi^{(r+1)}_l$ such that they still sum up to 1.

7. Re-conduct EM algorithm without penalty to obtain final parameter estimates and use the final estimates to compute BIC. In this EM algorithm, one goes through steps 1-5, except that in step 4, compute $\pi^{(r)}_l = \frac{\nu^{(r)}_l}{\sum_{l=1}^{L} \nu^{(r)}_l}$ only for those latent classes with g_l =1, where $\nu^{(r)}_l = \sum_{i=1}^{N} P^r_{il}$. Still make sure all $\pi^{(r)}_l$ add up to 1. [In fact, we will use an initial value of $\pi^{(0)}_l$ such that those impermissible latent classes will have $\pi^{(0)}_l = 0$. They will remain 0 during the EM algorithm].

total response options, i.e., $K_j = 3$, and assume the total number of facets M=3 (e.g., 1 goal facet and 2 intermediate facet) for simplicity, and presume the Q-matrix takes the form (note that the first option is the key), $\mathbf{Q}_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Different from Liu and Culpepper (2023), we have a term $\left[\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}^g_{j,k}\right) - \mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}^g_{j,K}\right)\right]$ in DFSM, hence, we need to derive a $\mathbf{Q}^*_j$ by subtracting row 1 from row 2 and row 3, respectively, before obtaining the Δ-matrix. That is,

$$\mathbf{Q}^*_j = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \rightarrow \boldsymbol{\Delta}_j = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\rightarrow \boldsymbol{\lambda}_j = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_{j,0} & -\lambda_{j,1,1} & \lambda_{j,2,1} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{j,0} & -\lambda_{j,1,1} & 0 & \lambda_{j,2,2} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here, $\boldsymbol{\Delta}_j$ is a K_j-by-2^{M} matrix, and as shown, the first row refers to the first option which is the "reference" category in the nominal response DFSM model. The last two rows refer to the two distractors. Each column refers to one possible facet profile, and the first column refers to the intercept. Consistent with previous notation, the three subscripts in $\lambda_{j,1,1}$ refers to item j, goal facet, and α_1, respectively. Similarly, the subscripts in $\lambda_{j,2,2}$ refer to item j, intermediate facet, and β_2, respectively. Since we only consider "main" effects of facets on item responses, column 4 to 8 are all 0's in both $\boldsymbol{\Delta}_j$ and $\boldsymbol{\lambda}_j$.

According to Liu and Culpepper (2023), the nominal RLCM is generically identifiable if the sparse three-dimensional array $(\mathbf{\Delta})_{J\times\sum K_j\times 2^M} = \begin{bmatrix}\mathbf{\Delta}^1\\ \mathbf{\Delta}^2\\ \mathbf{\Delta}^{'}\end{bmatrix}$ satisfies the following two conditions:

(a) For $j = 1, ..., M$, Δ_j^1 and Δ_j^2 meet the structure of

$$\mathbf{\Delta}_j = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0\\ * & \cdots & * & \delta_{j,j,1} & * & \cdots & *\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ * & \cdots & * & \delta_{j,j,K_j-1} & * & \cdots & * \end{bmatrix}_{K_j\times 2^M}$$

where the first row refers to the reference option, which, in the case of DFSM, refers to the answer key, $*$ can be either 0 or 1, and $\sum_{k=1}^{K_j-1}\delta_{j,j,k} > 1$. The latter constraint implies that for item j, at least one of the response options must measure facet j. Or in other words, this condition implies that, in the test, for each facet m, there are at least 2 items and each item has at least one response option measuring m. Hence, the test length is at least $2 \times M$.

(b) For every $m = 1, ..., M$, $\mathbf{\Delta}^{'}$ satisfies that there exists an item $j > 2M$ that $\sum_{k=1}^{K_j-1}\delta_{j,m,k} \geq 1$. This implies that there is at least one item in the last $J - 2M$ items that has a response option that measures facet m.

Note that the generic identification condition described above is directly adopted from Theorem 2 of Liu and Culpepper (2023). Although in the current DFSM model set up, the loadings and slopes are at item (instead of item option) level (i.e., see the $\boldsymbol{\lambda}_j$ matrix), because every response option of an item will measure different facets, the rank of the $\boldsymbol{\lambda}_j$ matrix will be the same regardless of whether the parameters are defined at item level or item option level. Furthermore, the proof of Theorem 2 only requires $\mathbf{p}_{j0} \neq \mathbf{p}_{j1}$ for $j = 1, ..., M$ (see Equation A15 in their paper), where $\mathbf{p}_{j0}$ is the two-dimensional vector including the probability of selecting the key for a given facet profile and one minus that probability (i.e., sum of two elements in $\mathbf{p}_{j0}$ =1). $\mathbf{p}_{j0}$ is the two-dimensional vector including the probability of selecting a distractor for a given facet profile and one minus that probability. The inequality $\mathbf{p}_{j0} \neq \mathbf{p}_{j1}$ easily holds when the above condition (a) is satisfied. In fact, using the example item presented in the earlier remark, as shown in Table 14, the probability of selecting the key is always different from the probability of selecting the distractor for all possible facet profiles.

2.6. *Facet Map*

Another instructionally formative piece of information generated by DFSM is a comprehensive facet map, which includes both directional and non-directional links between facets. A directional link implies a hierarchical structure (Dahlgren et al., 2006; Simon & Tzur, 2004) , reflecting theories about students' learning, where a piece of knowledge learned earlier serves as a pre-requisite for the more advanced knowledge, forming a so-called learning trajectory (LT, Corcoran et al., 2009; Daro et al., 2011; Templin & Bradshaw, 2014). A LT will allow for optimizing instructional design by offering theory-driven sequences of cognitive progression. In contrast, facets joined by a non-directional link would result in a "conjoined facet" which represents cross-cutting thematic ideas. Identifying conjoined facets could provide teachers insight into persistent ideas and reasoning that a student is applying across multiple learning units (with variable success), and which may require specific attention. For example, a dominant physics learner conception is that "motion implies an on-going force of motion" (see Option A in the item in Fig. 1), which means students are treating force as a *property* of the object. Perhaps if net force on an object were

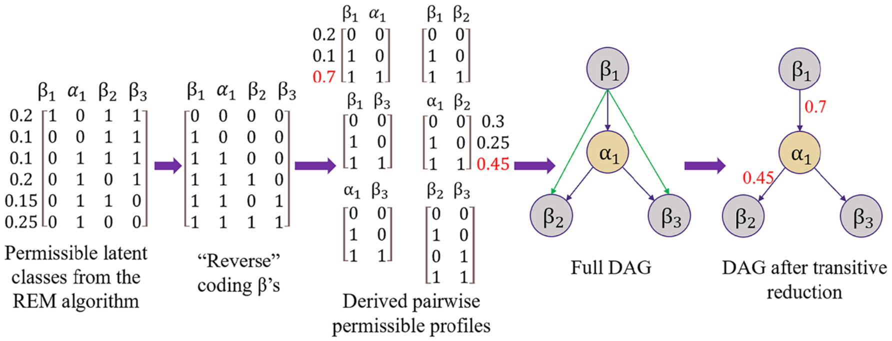

β₁ α₁ β₂ β₃
0.2
0.1
0.1
0.2
0.15
0.25
Permissible latent classes from the REM algorithm
“Reverse” coding β’s
0.2
0.1
0.7
0.3
0.25
0.45
Derived pairwise permissible profiles
β₁
α₁
β₂
β₃
Full DAG
0.7
0.45
DAG after transitive reduction

FIGURE 3.
Illustration of the deductive algorithm to construct DAG and added strength of the relationship.

taught as a *mechanism* conjoined with a resulting *change in the property* of momentum or kinetic energy of the object, student learning would be more naturally and productively facilitated.

Both facet hierarchy and conjoined facet can be inferred from the permissible facet profiles. Specifically, a hierarchical relationship exists between a pair of facets when mastery of a facet m is required prior to mastery of another facet $m^{'}$. That is, for person i, $\alpha_{im} = 1$ is a necessary but not sufficient condition for $\alpha_{im'} = 1$. As a result, for the combination of $(\alpha_{ik}, \alpha_{ik'})$, instead of having 4 possible combinations, only 3 would be possible, i.e., (0, 0), (1,0), and (1, 1). For conjoined facets (i.e., they are connected topics and should be treated as a unity), only 2 would be possible, i.e., (0, 0) and (1, 1). That is, a student will either master both facets or master neither of the facets, resulting in a non-directional link between them. The non-directional link implies that both facets in the conjoined set are at the same level, none is pre-requisite of the other.

We propose a deductive algorithm to infer the links among facets and eventually to construct a facet map. First, based on the significant latent classes, we will check pairs of facets separately to deduce the feasible profiles based on them and put a directional arrow (if any) between them, resulting in a directed acyclic graph (DAG). The DAG is denoted as $G(V, E)$ where V is a set of nodes (i.e., facets) and E is a set of directed edges that imply the hierarchical relationships among facets. Then we will perform transitive reductions to construct a simple structure. A transitive reduction is a subgraph with fewest edges that maintain the same reachability (i.e., directed connection) as G (Chen & Wang, 2023) .

When we put both goal and intermediate facets in the same map, we first need to “reverse” the profile of intermediate facets before using the same algorithm to determine G, that is, coding 0 as 1 and 1 as 0, respectively. This is because the possession of an intermediate facet implies lack of relevant knowledge. Using α_1, $\beta_1 \sim \beta_3$ in Fig. 2 as an example, Fig. 3 illustrates the proposed deductive algorithm for constructing DAG. It follows three steps:

(1) Step 1: For the final N-by-K (total number of goal and intermediate facets) estimated facet profile matrix, reverse code the columns related $\boldsymbol{\beta}$’s.
(2) Step 2: For each pair of facets, check the number of permissible patterns. If such a number is smaller than 4, that means there is a potential link between the pair. When the number is 2, it implies the existence of a conjoined facet, i.e., a non-directional arrow. When the number is 3, it implies a hierarchy. The direction of the arrows depends on the specific permissible pattern. In the example in Fig. 3, the permissible patterns between (β_1, α_1) are [0, 0], [1, 0], and [1, 1], resulting in an arrow from β_1 to α_1, denoting that β_1 is pre-requisite of α_1.

(3) Step 3: Transitive reduction. We only want to include direct links. If a facet has two or more arrows pointing toward it, it implies that there may exist redundancy. For instance, in Fig. 3, β_1 connects to β_2 through two routes: $\beta_1 \rightarrow \beta_2$ and $\beta_1 \rightarrow \alpha_1 \rightarrow \beta_2$. As a result, the direct connection between β_1 and β_2 can be eliminated to make DAG more parsimonious without changing its meaning. With a small number of facets, this step can be accomplished by inspecting the full DAG. When the number of total facets is large, we can construct an adjacency matrix of the DAG, A, based on the pair-wise relationships among facets. The A-matrix for the example in Fig. 3 will the following form with the order of facet $(\beta_1, \alpha_1, \beta_2, \beta_3)$:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \text{(after transitive reduction)} \tilde{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

That is, for any pair of facets denoted by (m_1, m_2), if $(A^2)_{m_1,m_2} > 0$ then we set $A_{m_1,m_2} = 0$, resulting in the second matrix above. The full, parsimonious DAG can be constructed based on $\tilde{A}$.

Once the DAG is constructed, we propose to further infer the strength of the arrow based on the estimated mixing proportions, as illustrated in the probabilities marked in red in Fig. 3. The strength value will be a probability measure between 0 and 1, with larger values indicating stronger relationships. As shown, such values are simply the proportion of students who possess both facets (again with intermediate facets reverse coded). This is intuitive because if a high proportion of students simultaneously possess both facets rather than merely the prerequisite, there is a strong pathway between them, i.e., say from β_1 to α_1, for a student who has matured from the intermediate facet β_1, they are highly likely to master α_1 as well.[2] A conjoined facet will be formed by adding a *bidirectional* arrow between two facets, which happens when 2 out of 4 latent profiles are permissible, implying the co-occurrence of two facets.

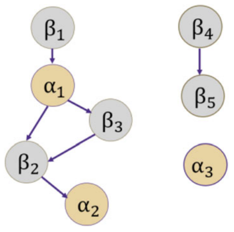
3. Simulation Study

A simulation study was conducted to evaluate the REM algorithm for DFSM parameter recovery. The sample size was fixed at 4,000 and we intentionally selected the large sample size due to model complexity. Test length was fixed at 27, and each item had 4 response options. It was assumed that the test measures 3 goal facets and 5 intermediate facets just to be consistent with our real data example. The Q-matrix was constructed as follows:

- Item 1–9: key requires 1 goal facet; none of the distractors require any goal facets. Each distractor only measures 1 intermediate fact.
- Item 10–18: key requires 2 goal facets; 1 distractor requires 1 goal facet and 1intermediate facets. The rest of the distractors only measure 1 intermediate facet and none of the goal facet.
- Item 19–27: key requires 2 goal facets; 1 distractor requires 1 goal facet and 1 intermediate facts; 1 distract requires 2 intermediate facets and 0 goal facet, and 1 distractor only requires 1 intermediate facet.

[2] One may argue that the strength of the link can be quantified as the probability of simultaneously possess plus simultaneously not possess both facets. We acknowledge that our selected statistic is somewhat confounded by the difficulty of the facets. That is, when the two facets are easy and basic, the probability of the class (1, 1) would be high, leading to a seemingly strong link between the two facets. As we mention in the discussion, other better ways to measure strength of links between facets are needed in the future.

(a) facet hierarchies



(b) correlation among facets

$$\begin{bmatrix} 1 & & & & & & & \\ .31 & 1 & & & \textit{symmetric} & & & \\ -.00 & .00 & 1 & & & & & \\ -.62 & -.19 & -.00 & 1 & & & & \\ -.50 & -.63 & .01 & .31 & 1 & & & \\ -.71 & -.44 & .00 & .44 & .71 & 1 & & \\ .00 & .01 & .02 & .01 & -.02 & -.01 & 1 & \\ -.00 & .02 & .00 & -.00 & -.00 & -.00 & .51 & 1 \end{bmatrix}$$

FIGURE 4.
Hierarchical facet structure in the simulation study.

As a result, the number of items measuring each attribute were 21, 21, 21, 18, 18, 18, 18, and 18, respectively, for the goal facets (the first 3) and intermediate facets (the last five). $\lambda_{j,0}$ was generated from Uniform (-1, 1), both $\lambda_{j,1}$ and $\lambda_{j,2}$ were generated from Uniform (1.75, 2.25), resulting in good informative items (de la Torre, 2011). We only considered main effects in both $\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right)$ and $\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)$. Twenty-five replications were conducted per condition. Different sets of item parameters were simulated per replication.

We manipulated facet relationship, as follows:

- Moderate correlation among facets: Assuming the attributes were generated from a higher-order DINA model (de la Torre & Douglas, 2004), i.e., $P(\alpha_k|\theta) = \frac{1}{1+e^{-1.7\gamma_{1k}(\theta-\gamma_{0k})}}$, where $\theta \sim N(0,1)$, $\gamma_{0k} \sim U(-1, 0.5)$, γ_{1k} =[1.35, 1.5, 1.65, −1.35, −1.425, −1.5, −1.575, −1.65]. Note that γ_{1k}'s are positive on the goal facets, but they are negative on the intermediate facet. This reflects the intuition that higher overall ability θ is associated high chance of mastery of goal facets and lower chance of possession of intermediate facets. The parameters were chosen to induce 30% sparsity, which means out of $2^8 = 256$ all possible patterns, there were about 180 permissible patterns. The absolute correlation among attributes were in the range of 0.4~0.5.
- Independent facets: We simulated all profiles from a uniform distribution, i.e., assume every facet profile among all 256 possible profiles are equally likely to occur. Then we imposed a constraint to ensure that those who have more goals facets will be less likely to have many intermediate facets, i.e., if $\sum_{k=1}^{3} \alpha_k \geq 2$ then $\sum_{k=1}^{5} \beta_k \leq 2$. This manipulation results in 25% sparsity and almost zero correlation among facets.
- Hierarchical facets: We simulated profiles based on a hypothetical theory-driven hierarchical structure, as shown in Fig. 4a. This structure was chosen because it contains divergent (i.e., α_1 is the pre-requisite of both β_2 and β_3), convergent (i.e., β_2 requires bothα_1 and β_3), and independent (i.e., α_3 is independent of the rest) structures (Liu et al., 2017; Wang, 2021b; Wang & Lu, 2021) in one map. Therefore, it is, to some extent, comprehensive to cover popular attribute relationships in practice. As a result, there were only 36 permissible patterns (listed in Table 15 in the Appendix), and the correlation among attributes are shown in Fig. 4b.

Evaluation criteria include the mean bias, mean relative bias, and mean absolute bias for all item parameters. They were averaged across 25 replications. For person's facet profiles, we computed goal facet recovery rate, i.e., $\sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\mathrm{I}_{|\alpha_{ik}=\hat{\alpha}_{ik}|}}{N \times K}$. Similarly, we computed the pattern recovery for intermediate facets as well as the combined profile. For population facet relationship, we compared the estimated sparsity structure with the true sparsity structure and computed (1) true positive rate (TPR) = Correct recovery of true permissible patterns, (2) 1-FDR (False discovery

TABLE 4.
Recovery of item parameters (values in the parathesis are standard deviations across items and across replications).

	Mean bias	Mean relative bias	Mean absolute bias
Moderately correlated facets			
$\lambda_{j,0}$	.002 (.031)	24.93% (89.20%)	.113 (.023)
$\lambda_{j,1}$	−.009 (.025)	−.47% (1.24%)	.102 (.012)
$\lambda_{j,2}$	−.013 (.016)	−.67% (.77%)	.085 (.010)
Independent facets			
$\lambda_{j,0}$	−.005 (.019)	−17.98% (47.88%)	.067 (.011)
$\lambda_{j,1}$	−.005 (.013)	−.24% (.69%)	.078 (.008)
$\lambda_{j,2}$	−.003 (.014)	−.17 (.72%)	.066 (.007)
Hierarchical facets			
$\lambda_{j,0}$	−.003 (.001)	−10.38% (27.15%)	.078 (0.041)
$\lambda_{j,1}$	−.004 (.021)	−.17% (1.09%)	.097 (.027)
$\lambda_{j,2}$	−.021 (.034)	−.36% (2.03%)	.080 (.021)

rate) = proportion of detected permissible patterns that are true patterns, and (3) true negative rate (TNR)=correct recovery of true impermissible patterns. We aimed for these three rates to be as close to 1 as possible.

Table 4 presents the recovery of item parameters. Values in the parathesis are standard deviations across items and across replications. As shown, all item parameters appear to recover well, and there is no appreciable difference among the three facet relational conditions, or among different types of item parameters. Of note, the relative bias of $\lambda_{j,0}$ is high (so is its standard deviation across replications) which is merely because some of the true $\lambda_{j,0}$'s is close to 0, hence when they appear in the denominator to compute relative bias, the relative bias shoots up.

Table 5 presents the recovery of facets and facet profiles. It is unsurprising that recovery at facet level is much higher than recovery at profile level. Moreover, recovery at overall profile level is the hardest, followed by recovery of the five-dimensional intermediate facet profile, whereas recovery of the three-dimensional goal facet profile is the most accurate. In addition, under the hierarchical facet condition, the facet and facet profile recovery are the best, whereas their recoveries are the worst under the independent facet condition. This is much expected because when the facets display a hierarchical structure, only 36 out of 256 facet profiles are permissible, which largely reduces the space for searching the optimal facet profile for each student. When the facets are independent, there is not much shared information among different facets, making the recovery of the entire profile hardest.

Table 6 presents the recovery of facet sparsity structure. First to notice is under both hierarchical and independent conditions, all TPR, TNR and 1-FDR are close to 1, implying that the sparsity structure is mostly accurately recovered under these two conditions. The hierarchical condition slightly outperforms the independent condition simply because it has less density and therefore easier to locate true permissible profiles. Second, where TNR and 1-FDR are close to 1 under the moderately correlated facets condition, whereas the average TPR is only.526. This means that the algorithm tends to generate a sparser solution, thereby missing about 48% of the true permissible pattern. This is not too surprising as it is well known in Lasso regression that when the predictors are correlated, correct identification of significant predictors is harder, especially when true density is high (i.e., less sparsity).

Table 5.
Recovery of facets and facet profiles.

	Profile	Facet
Moderately correlated facets		
Overall	.693 (.020)	.950 (.004)
α	.874 (.008)	.956 (.003)
β	.780 (.020)	.947 (.005)
Independent facets		
Overall	.625 (.014)	.938 (.003)
α	.827 (.009)	.936 (.004)
β	.742 (.015)	.940 (.004)
Hierarchical facets		
Overall	.773 (.019)	.963 (.002)
α	.891 (.021)	.962 (.001)
β	.852 (.018)	.964 (.001)

Table 6.
Recovery of facet sparsity structure.

	True pattern	TPR	1-FDR	TNR
Moderately correlated facets	186/256	.526 (.037)	.991 (.011)	.984 (.018)
Independent facets	192/256	.952 (.016)	.999 (.003)	.996 (.008)
Hierarchical facets	36/256	1 (0.008)	.964 (0.015)	.990 (0.002)

4. Real Data Example

The feasibility of DFSM was demonstrated using data from *Diagnoser* assessment. The assessment contains multiple-choice items with diagnostically rich response options, each mapped onto well-defined facets (Minstrell et al., 2015) . The *Diagnoser* items have gone through various psychometric analyses to validate interpretations of learning progression level diagnosis (Chattergoon, 2020; Steedle & Shavelson, 2009) , although our analysis would be the first to use DFSM to construct facet map. Nine items from the unit of "Identification of forces" were used, and the sample size was $N = 2,729$. Every item has 4 response options except item #3 which has 5 options. The Q-matrix for these 9 items is presented in the Appendix. Overall, the number of response options measuring each facet (i.e., the first 3 are goal facets and the remaining 5 are intermediate facets) is 19, 16, 6, 6, 5, 4, 4, and 8. As shown, the Q-matrix is sparse, and only the first two goal facets are measured by more than half of the response options. The definition of the facets is presented in Table 7. Table 8 presents the proportion of endorsement of each response option per item, and the key per item is bolded. As shown in Table 8, for 7 out of 9 items (except for items 4 and 5), the proportion of students selecting the key is the highest, indicating that the items are relatively easy for the student sample. We also use the reliability index proposed in Templin and Bradshaw (2014) and found that the reliability of the three goal facets are: 0.89, 0.87, 0.92, respectively, whereas the reliability of the five intermediate facts are 0.84, 0.82, 0.74, 0.43, and 0.68, respectively. It is worth noting that higher magnitude of slopes on the corresponding facets (i.e., $\boldsymbol{\lambda}_{j,1}$ and $\boldsymbol{\lambda}_{j,2}$) leads to higher reliability unsurprisingly, whereas the number of items options measuring each facet does not seem to affect facet reliability much.

TABLE 7.
Definition of goal and intermediate facets for the real data example using items from the unit of "Identification of Forces".

Goal Facets	1	Students can correctly identify the object or the mechanism that exerts the force. (α_1)
	2	Students can correctly identify the direction in which the force is acting. (α_2)
	3	Students can compare the relative sizes of the forces on an object. (α_3)
Intermediate Facets	1	Students express that the downward force is greater than the upward force on an object that is at rest or moving horizontally (to keep the object from lifting off). (β_1)
	2	Students express that force is a property of an object, and that the size of the force is dependent on the strength of the property (e.g., the faster, heavier, stronger object has more force). (β_2)
	3	Student cites an energy source as a force (e.g., the engine exerts a forward force on the car). (β_3)
	4	Students treat passive, non-active objects as unable to exert a force even if in contact with the object (e.g., the table under the book does not exert a force on the book). (β_4)
	5	Students determine the existence of a force by whether the object is in motion (e.g., the rolling marble has a force of motion). (β_5)

Note. Learners frequently speak of "force" as though it were a property of an object rather than an action on an object (i.e., "the object *has* force"). While this idea is less productive to understanding forces, it is useful for thinking about momentum or kinetic energy. Therefore, an α attribute in one unit may be modeled as β attribute in another unit

TABLE 8.
Proportion of students selecting each response option per item.

Items	1	2	3	4	5	6	7	8	9
Option 1	0.045	0.254	0.114	0.498	**0.102**	0.259	0.066	0.027	**0.754**
Option 2	0.132	0.081	0.036	**0.343**	0.263	0.224	0.228	0.150	0.121
Option 3	**0.743**	**0.663**	0.321	0.127	0.347	0.134	**0.673**	0.034	0.035
Option 4	0.080	0.001	0.112	0.032	0.287	**0.383**	0.033	**0.789**	0.090
Option 5			**0.417**						

Note. The answer key per item is bolded

To cross-check the results, we randomly split the data into a training set (sample size = 2,000) and a test set (sample size = 729) 10 times. For each random split, we fit the DFSM using the REM algorithm on the training set and used the estimated model parameters (including both item parameters and population class mixing proportions) on the test set to estimate students' facet profiles. The estimated item parameters and their standard error are presented in Table 9. The point estimates were obtained by averaging over 10 analyses. The standard error was computed as standard deviation of the point estimates from the 10 random training sets. Several points can be spotted from Table 9. First, majority of the items have negative $\lambda_{j,0}$'s, which is consistent with the observations from Table 8 that most items are relatively easy. Second, some of the cells for $\boldsymbol{\lambda}_{j,1}$ and $\boldsymbol{\lambda}_{j,2}$ are blank simply because none of the response options of those items measure the corresponding facets. Among all the estimated slopes, there are six that turn out to be nearly 0 (see the values highlighted in italics), which implies that the items are not discriminative with respect to the related facet. Take item 9 as an example, the estimated slopes on both α_1 and $\beta_2$are 0. As a result, for students with true facet profiles of [1,1,1,0,0,0,0,0], [0,1,1,0,0,0,0,0], [1,1,1,0,1,0,0,0], [0,1,1,0,1,0,0,0], they all have exactly the same predicted response probabilities of [.904,.033,.033,.030] for the four response options, respectively. That is, the item can hardly differentiate students with or without any or both of α_1 and β_2. In fact, by inspecting Table 8, it is not surprising to note that the frequency of choosing option 1 (i.e., the key) is the highest, whereas the frequencies of choosing any of the distractors are all low. Third, a few of the standard errors seem to be large, which indicate the corresponding parameters were not stably estimated. The large standard errors mostly occur on slopes for α_2, and as we show below, α_2 turns out to be the most difficult facet.

Figure 5 presents an estimated facet map. It is worth noting that not every round of analysis produced the same facet map, instead, we put a solid link between two facets if such link appeared in 5 or more out of 10 rounds of analysis. We put a dashed link between two facets if such link appeared in less than 5 rounds of analysis. Table 10 presents the details of 11 links that were identified between two facets. All these links were directional as reflected by 3 instead of 4 permissible classes defined by two facets, indicating that no conjoined facets were detected. Table 10 also includes the frequency of occurrence of each possible link, as well as the estimated probabilities of each permissible latent class. These probabilities were averaged across 10 random splits. Figure 5 presents the results in Table 10 graphically for easier visualization. Several hierarchical facets showed up. For instance, α_3, β_1, and β_4 are all pre-requisites of α_2, making α_2 arguably the most difficult goal facet. There is also an interesting linear hierarchy between β_1, β_4, and β_3, as well as another linear hierarchy between β_1, β_4, and β_5. The strengths of the solid links were also displayed in Fig. 5.

Table 11 presents the probability of possession of each facet in both training and test sets, averaged across 10 rounds of analysis. As can be seen, the probabilities are quite consistent between training and testing sets. Further, the mastery probability of α_2 is the lowest, which is consistent with the conjectured facet map that α_2 tends to be most difficult. In addition, β_1 and β_4 tends to be at the lower end of the spotted linear hierarchy, and hence it is unsurprising to see low probabilities on these two facets, implying that there may not be a lot of misconceptions among students on these most basic understandings.

Lastly, we evaluated item fit by using a chi-squared statistic. Two versions of the statistics were considered. For version I, for an item j, among all the permissible facet profiles (i.e., denoted by G, and $G < 2^M$), we compute the following chi-squared statistic,

$$\chi_j^2 = \sum\nolimits_{g=1}^{G} \sum\nolimits_{k=1}^{K_j} n_g \frac{\left(O_{jgk} - E_{jgk}\right)^2}{E_{jgk}\left(n_g - E_{jgk}\right)}. \tag{11}$$

Here O_{jgk} and E_{jgk} are the observed and expected count of students who are in latent class $g$and select option k of item j. We also considered two versions of degree-of-freedom $G \times \left(K_j - 1\right) - d_j$ where d_j is the total number of item parameters for item j and $G - d_j$. The former case seems to fit better with Eq. 11, whereas the latter term is used in polytomous IRT model (e.g., Naumenko, 2014; Su et al., 2021).

For version II, for an item j, we will consider all permissible latent groups G_j (i.e., $G_j \leq G$, and $G_j \leq 2^{K_j}$). The latent groups are created based on the q-vector of item j, and each latent group may contain several latent classes with identical attribute vector with respect to the required attributes (e.g., if $q = (1, 1, 0, 0)$, the latent classes (1,1,0,0), (1,1,1,0), (1,1,0,1) and (1,1,1,1) are placed in the latent group. For DFSM, the element in the q-vector of item j is 1 if at least one

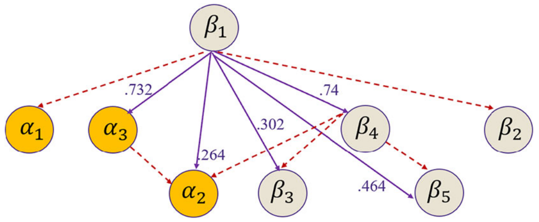


FIGURE 5.
Estimated facet map.

TABLE 9.
Estimated item parameters and their standard errors.

	$\lambda_{j,0}$	$\boldsymbol{\lambda}_{j,1}$			$\boldsymbol{\lambda}_{j,2}$				
Item		α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5
1	−0.69 (.12)	1.48 (.33)	0.88 (.56)	2.17 (.12)	*0.00 (.00)*			1.38 (.25)	0.64 (.26)
2	−1.90 (.13)	1.88 (.22)	2.43 (1.18)	0.86 (.14)	2.68 (.29)		3.92 (.17)		*0.00 (.00)*
3	−0.53 (.23)	*0.00 (.00)*	0.95 (.82)	1.24 (.18)	0.56 (.68)			0.96 (.46)	1.98 (.30)
4	1.06 (.38)	2.67 (.26)	3.12 (1.03)	0.51 (.43)	0.37 (.49)	0.57 (.59)			1.35 (.32)
5	−0.17 (.76)	1.17 (.79)				4.00 (.01)	3.61 (.28)		3.67 (.25)
6	−0.58 (.79)	2.43 (1.43)	1.39 (1.31)			3.65 (.23)	3.61 (.43)		3.81 (.25)
7	−0.81 (.29)	1.90 (.29)	2.18 (1.42)		0.07 (.14)		1.85 (.10)		0.44 (.17)
8	−0.27 (.22)	2.12 (.23)	2.22 (1.59)	3.89 (.17)	1.47 (.66)			*0.01 (.04)*	0.52 (.36)
9	−0.18 (.13)	*0.00 (.00)*	0.09 (.13)	3.32 (.07)		*0.00 (.01)*		0.28 (.17)	

Values in the parathesis are standard deviations computed across 10 random splits.

option of item j measures the corresponding attribute. Then the chi-squared statistic is

$$\chi_j^2 = \sum\nolimits_{g=1}^{G_j} \sum\nolimits_{k=1}^{K_j} n_g \frac{\left(O_{jgk} - E_{jgk}\right)^2}{E_{jgk}\left(n_g - E_{jgk}\right)}. \tag{12}$$

Here O_{jgk} and E_{jgk} are the observed and expected count of students who are in latent group $g$and select option k of item j. The two versions of degree of freedom are $G_j \times \left(K_j - 1\right) - d_j$ where d_j is the total number of item parameters for item j and $G_j - d_j$.

Before applying the item fit indices on real data, we tested them on the simulated data from two conditions: moderately correlated facets and hierarchical facets. Since we did not simulate any misfit intentionally, we only focused on Type I error of each index, as shown in Table 12. In this Table, we also considered whether correcting for family-wise error rate using Bonferroni correction. Most of the Type I error rates were below.05, and if Type I error is extremely small (such as 0), the test may not be powerful. We aim to look for indices that have Type I error closest to.05, and they are highlighted in red in Table 5. It is not surprising that implementing Bonferroni correction yields lower Type I error. Further, Type I error is considerably lower for moderated correlated facet condition because this condition yields many more permissible facet profiles (see Table 6). Because the real data mimics the "hierarchical" structure, we tried the two methods marked in Table 12 on the real data. Both methods flag the same items as having potential misfit: item # 1, 3, 7, and 9, which suggest that either the q-matrix of these items need to be revised, or the additive structure assumed in $\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}_{j,k}^{g}\right)$ and $\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}_{j,k}^{p}\right)$ need to be revisited.

TABLE 10.
Estimated links between pairs of facets.

	# of permissible latent classes	Frequency (out of 10 random splits)	Class 1	Probability	Class 2	Probability	Class 3	Probability	Implications
α_1 β_1	3	3	0 0	0.032	0 1	0.132	1 1	0.836	$\beta_1 \rightarrow \alpha_1$
α_2 α_3	3	4	0 0	0.246	0 1	0.541	1 1	0.212	$\alpha_3 \rightarrow \alpha_2$
α_2 β_1	3	10	0 0	0.067	0 1	0.668	1 1	0.264	$\beta_1 \rightarrow \alpha_2$
α_2 β_4	3	1	0 0	0.224	0 1	0.540	1 1	0.236	$\beta_4 \rightarrow \alpha_2$
α_3 β_1	3	5	0 0	0.044	0 1	0.224	1 1	0.732	$\beta_1 \rightarrow \alpha_3$
β_1 β_2	3	3	0 0	0.032	1 0	0.469	1 1	0.499	$\beta_1 \rightarrow \beta_2$
β_1 β_3	3	10	0 0	0.067	1 0	0.630	1 1	0.302	$\beta_1 \rightarrow \beta_3$
β_1 β_4	3	10	0 0	0.067	1 0	0.193	1 1	0.740	$\beta_1 \rightarrow \beta_4$
β_1 β_5	3	10	0 0	0.067	1 0	0.469	1 1	0.464	$\beta_1 \rightarrow \beta_5$
β_3 β_4	3	2	0 0	0.188	0 1	0.541	1 1	0.271	$\beta_4 \rightarrow \beta_3$
β_4 β_5	3	1	0 0	0.152	1 0	0.339	1 1	0.509	$\beta_4 \rightarrow \beta_5$

The displayed permissible classes for each pair of facets were after reversing the intermediate facets.

TABLE 11.
Average probability of possession of each facet in training and test sets.

	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5
Training	0.799	0.334	0.752	0.076	0.378	0.686	0.177	0.453
Test	0.809	0.334	0.751	0.069	0.353	0.672	0.168	0.457

TABLE 12.
Type I error of various item fit indices.

	Chi-squared statistics defined in Eq. 11				Chi-squared statistics defined in Eq. 12			
Degree of freedom	$G \times (K_j - 1) - d_j$		$G - d_j$		$G_j \times (K_j - 1) - d_j$		$G_j - d_j$	
Bonferroni correction	Y	N	Y	N	Y	N	Y	N
Hierarchical	0	.009	0	.092 [**Method 1**]	.002	.030 [**Method 2**]	.388	.716
Moderate correlation	0	0	0	0	0	.015	0	.053

5. Discussion

This study provides a valuable psychometric tool to greatly enhance diagnostic assessment. Current scoring rubrics in most assessments still primarily gradate performance (i.e., right/wrong, or partial credit). Such assessment results only provide an overview of student learning progress. Instead, the best feedback should identify both correct and intermediate states of understanding. This is exactly the type of information that DFSM will extract from assessment data. With targeted and individualized feedback, every student will get appropriate support on their greatest needs to be successful, a condition essential to achieving educational equity. Our simulation study and real data example show great promise for the DFSM and the regularized EM algorithm.

Note that another available model that could be used for facet-based items is the generalized diagnostic classification model for multiple-choice option-based scoring (GDCM-MC), which accounts for both goal and intermediate facets (DiBello et al., 2015) . However, the GDCM-MC has a complex formulation which makes Bayesian MCMC algorithm a preferable choice than the more efficient EM algorithm. Hence, the regularization technique that is proposed can hardly be applied for GDCM-MC in its current form. Second, as pointed out in a recent paper, the possible coexistence of skills and preconceptions is not adequately handled in GDCM-MC (Kuo et al., 2018). Third, GDCM-MC requires a three-valued coding scheme to specify which facets are strongly related to each option, and this puts a higher requirement on the expert-provided coding matrix. Instead, the DFSM model can naturally handle guessing and it requires less complicated coding of the option-facet mapping matrix. Compared to GDCM-MC, the DFSM model also has more compact parameterization such that it permits using the EM algorithm. If the option to facet mapping is sparse (i.e., not enough options measuring each facet), the parameterization of the DFSM model can be further reduced to make λ's as attribute-level parameters instead of item-by-attribute level parameters. Very recently, a model like DFSM was published (Levy, 2019) and it was used to analyze game-based assessment data where each step a student takes is cognitively coded. However, the authors used Bayesian estimation and did not consider exploring attribute relationships.

One innovation worth highlighting is that our REM algorithm coupled with deductive algorithm permits deriving a facet map. Despite the instructional benefits of a detailed facet map, identifying the multitude of links among facet pairs poses a methodological challenge. For example, just 9 facets could generate as many as 72 potential directional links. Templin and Bradshaw (2014) proposed a likelihood ratio (LR) test formatted within a hierarchical diagnostic classification model to statistically test one hypothesized link at a time. However, it is cumbersome to perform their LR testing at scale because the reference distribution needs to be simulated each time to derive the empirical p-value. Instead, we propose to leverage a fast ML method to directly identify massive potential facet hierarchies and create a facet map with data-driven directional and non-directional links. Such empirical evidence will not only complement existing theory driven LTs but may also reveal new insights to refine or expand theoretical trajectories.

This research can be expanded in several directions. First and foremost, as we mentioned in the real data analysis, the model may not be identified with the given Q-matrix because we noted that there does not seem to be an exact one-on-one relationship between the permissible latent classes and facet map. From our discussion about generic identifiable conditions, the Q-matrix in the real data example does not satisfy the conditions because one simple requirement of the condition is there are more than $2M$ items. In the real data example, since we have 8 total facets, there needs to be at least 16 items. However, the conditions we adopt from Liu and Culpepper (2023) are sufficient conditions and they may not be necessary. Plus, the conditions may change when there exist facet hierarchies or conjoined facets. Further studies need to be conducted to establish necessary conditions for DFSM to be generically identified, with and without facet relationships. Due to potential under identification, the revealed facet maps differed in certain aspects across 10 cross-validation analyses on one hand, and on the other hand, if we derived the facet map via the deductive algorithm from estimated permissible classes, and then used the derived facet map to map out the permissible classes, they did not exactly match the estimated permissible classes. This noted discrepancy implies the potential model identification issues, which did not happen in simulation study. Because the Q-matrix is much sparser and test length is much shorter in real data compared to simulated data, future study is needed to delineate the necessary (and sufficient) conditions of Q-matrix for DFSM to be identified. This information is essential for not only real data analysis, but also for designing future facet-based diagnostic assessments.

Along this line of inquiry, a second research direction would be to explore the types of items that provide high information. In classical item response theory, it is widely known that polytomously scored items tend to carry more information than dichotomously scored items, and items with higher discrimination are more informative. With DFSM, although it is intuitive to claim that higher $\lambda_{j,1}$ and $\lambda_{j,2}$ yield higher item information, it is still unknown what kind of Q-matrix structure makes an item more informative. For instance, should each distractor only measure one intermediate facet? Or should each distractor measure multiple intermediate facets to max out the number of options per facet? A future study should take a deeper dive to study the item information for DFSM, and to the extent possible, quantify the information gains by modeling nominal responses compared to binary scoring.

Further, our simulation study is limited to additive structure of $\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}^g_{j,k}\right)$ and $\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}^p_{j,k}\right)$, as well as one level of test length and sample size. Increasing sample size and/or increasing test length will certainly improve the estimation of item parameters and person parameters, respectively, determining the minimum sample size needed for adequate recovery of DFSM item parameters under different manipulated conditions is still needed because such information is essential to plan and interpret real data analysis. In addition, we provide a heuristic descriptive statistic to infer the strength of links between pairs of facets, which is essentially the estimated probability of having both the pre-requisite and current facets. Such a descriptive statistic is only a proxy to the actual strength of linkages, whereas future research may consider network models that can directly estimate and make inferences on the link strengths. Furthermore, our real data analysis revealed that 4 items may exhibit misfit. Not only will a separate study exploring the performance of various item fit indices in-depth be needed, but also developing more flexible versions of DFSM will be critical as well. For instance, allowing interaction terms in $\mathbf{h}\left(\boldsymbol{\alpha}_i, \mathbf{q}^g_{j,k}\right)$ and $\mathbf{h}\left(\boldsymbol{\beta}_i, \mathbf{q}^p_{j,k}\right)$ or allowing $\boldsymbol{\lambda}_j$ to be an item-option level parameter instead of an item-level parameter will make the model more flexible, albeit less parsimonious. Lastly, the derived facet map should be thoroughly validated through a well-planned validation study that will likely involve student cognitive interviews and teacher focus groups.

6. Appendix

See Tables 13, 14, 15.

TABLE 13.
Q-matrix of the sample *Diagnoser* items in the real data analysis.

Item	Response	β_1	β_2	β_3	β_4	β_5	α_1	α_2	α_3
1	1	0	0	0	0	1	0	0	0
1	2	0	0	0	1	0	1	0	0
1	3	0	0	0	0	0	1	1	1
1	4	1	0	0	0	0	1	1	0
2	1	0	0	1	0	0	0	0	0
2	2	1	0	0	0	0	1	1	0
2	3	0	0	0	0	0	1	1	1
2	4	0	0	0	0	1	0	0	0
3	1	0	0	0	0	0	1	1	0
3	2	1	0	0	0	0	1	1	0
3	3	0	0	0	0	1	0	0	0
3	4	0	0	0	1	0	1	0	0
3	5	0	0	0	0	0	1	1	1
4	1	0	1	0	0	0	1	0	0
4	2	0	0	0	0	0	1	1	1
4	3	0	0	0	0	1	0	0	0
4	4	1	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0
5	2	0	0	0	0	1	0	0	0
5	3	0	0	1	0	0	0	0	0
5	4	0	1	0	0	0	0	0	0
6	1	0	0	1	0	0	0	0	0
6	2	0	0	0	0	1	0	0	0
6	3	0	1	0	0	0	0	0	0
6	4	0	0	0	0	0	1	1	0
7	1	0	0	0	0	1	0	0	0
7	2	0	0	1	0	0	0	0	0
7	3	0	0	0	0	0	1	1	0
7	4	1	0	0	0	0	0	0	0
8	1	0	0	0	1	0	0	1	0
8	2	1	0	0	0	0	1	1	0
8	3	0	0	0	0	1	0	0	0
8	4	0	0	0	0	0	1	1	1
9	1	0	0	0	0	0	1	1	1
9	2	0	1	0	0	0	1	1	0
9	3	0	1	0	0	0	1	1	0
9	4	0	0	0	1	0	0	0	0

TABLE 14.
The probability of selecting each response option of an example item.

Latent class	α_1	α_2	β_1	β_2	A	B	C	D
1	0	0	0	0	0.232	0.256	0.256	0.256
2	1	0	0	0	0.575	0.142	0.142	0.142
3	0	1	0	0	0.690	0.103	0.103	0.103
4	0	0	1	0	0.147	0.345	0.163	0.345
5	0	0	0	1	0.102	0.113	0.393	0.393
6	1	1	0	0	0.909	0.030	0.030	0.030
7	1	0	1	0	0.437	0.228	0.108	0.228
8	1	0	0	1	0.337	0.083	0.290	0.290
9	0	1	1	0	0.561	0.178	0.084	0.178
10	0	1	0	1	0.456	0.068	0.238	0.238
11	0	0	1	1	0.065	0.152	0.251	0.532
12	1	1	1	0	0.851	0.060	0.028	0.060
13	1	1	0	1	0.790	0.026	0.092	0.092
14	1	0	1	1	0.238	0.124	0.205	0.433
15	0	1	1	1	0.340	0.108	0.177	0.375
16	1	1	1	1	0.697	0.049	0.081	0.172

TABLE 15.
The 36 permissible patterns under the hierarchical facet simulation condition.

Pattern	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5
1	0	0	1	1	1	1	1	1
2	0	0	1	0	1	1	1	1
3	1	0	1	0	1	1	1	1
4	1	0	1	0	1	0	1	1
5	1	0	1	0	0	0	1	1
6	1	1	1	0	0	0	1	1
7	0	0	0	1	1	1	1	1
8	0	0	0	0	1	1	1	1
9	1	0	0	0	1	1	1	1
10	1	0	0	0	1	0	1	1
11	1	0	0	0	0	0	1	1
12	1	1	0	0	0	0	1	1
13	0	0	1	1	1	1	0	1
14	0	0	1	0	1	1	0	1
15	1	0	1	0	1	1	0	1
16	1	0	1	0	1	0	0	1
17	1	0	1	0	0	0	0	1
18	1	1	1	0	0	0	0	1
19	0	0	0	1	1	1	0	1
20	0	0	0	0	1	1	0	1
21	1	0	0	0	1	1	0	1
22	1	0	0	0	1	0	0	1
23	1	0	0	0	0	0	0	1
24	1	1	0	0	0	0	0	1
25	0	0	1	1	1	1	0	0
26	0	0	1	0	1	1	0	0
27	1	0	1	0	1	1	0	0
28	1	0	1	0	1	0	0	0
29	1	0	1	0	0	0	0	0
30	1	1	1	0	0	0	0	0
31	0	0	0	1	1	1	0	0
32	0	0	0	0	1	1	0	0
33	1	0	0	0	1	1	0	0
34	1	0	0	0	1	0	0	0
35	1	0	0	0	0	0	0	0
36	1	1	0	0	0	0	0	0

References

Alonzo, A. C., & Steedle, J. T. (2009). Developing and accessing a force and motion learning progression. *Science Education, 93*(3), 389–421.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Chattergoon, R. (2020). *Using polytomous item response theory models to validate learning progressions*. Univeristy of Colorado at Boulder, Doctoral Dissertation.

Chen, Y., & Wang, S. (2023). Bayesian estimation of attribute hierarchy for cognitive diagnosis models. *Journal of Educaitonal and Behavorial Statistics*. https://doi.org/10.3102/10769986231174918

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–31.

Clement, J. (1987). *The use of analogies and anchoring intuitions to remediate misconceptions in mechanics*. Paper presented at the Annual Meeting of American Educational Research Association, Washington, DC.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Report of the Center on Continuous Instructional Improvement, Teachers College, Columbia University, New York.

Dadey, N. (2017). *Reporting scores from NGSS assessments: Exploring scores & subscores*. Portsmouth, NH: Reidy Interactive Lecture Series.

Dahlgren, M. A., Hult, H., Dahlgren, L. O., af Segerstad, H. H., & Johansson, K. (2006). From senior student to novice worker: Learning trajectories in political science, psychology and mechanical engineering. *Studies in Higher Education,31*(5), 569–586.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). Learning Trajoctories in Mathematics: A Foundation for Standards, Curriculum, Assessment and Instruction. CPRE Research Report #RR-68. New York: Columbia University.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied psychological measurement, 39*(1), 62–79.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2&3), 165–255.

diSessa, A. A. (2014a). A history of conceptual change research: Threads and fault lines. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 88–108). Cambridge: Cambridge University Press.

diSessa, A. A. (2014b). The construction of causal schemes: Learning mechanisms at the knowledge level. *Cognitive science, 38*(5), 795–850.

diSessa, A. A. (2017). Knowledge in pieces: An evolving framework for understanding knowing and learning. In T. G. Amin & O. Levrini (Eds.), *Converging perspectives on conceptual change: Mapping an emerging paradigm in the learning sciences* (pp. 7–16). London: Routledge.

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. Berlin: Springer.

Foisy, L.-M.B., Potvin, P., Riopel, M., & Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics? *Trends in Neuroscience and Education, 4*(1–2), 26–36.

Gu, Y., & Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning, 20*, 1–58.

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79–95.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter. *Journal of Chemical Education, 90*(12), 1602–1608.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Kapon, S., & diSessa, A. A. (2012). Reasoning through instructional analogies. *Cognition and Instruction, 30*(3), 261–310.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kuo, B. C., Chen, C. H., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement, 42*(3), 179–191.

Lanting, A. S. Y. (2000). *An empirical study of a district-wide K–2 performance assessment program: Teacher practices, information gained, and use of assessment results*. University of Illinois at Urbana-Champaign.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research, 54*, 771–794.

Liu, Y., & Culpepper, S. A. (2023). Restricted latent class models for nominal response data: Identifiability and estimation. *Psychometrika*. https://doi.org/10.1007/s11336-023-09940-7

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*(2), 220–240.

Minstrell, J. (1989). Teaching science for understanding. In L. B. Resnick & L. E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 129–149). Alexandria, VA: Association for Supervision and Curriculum Development.

Minstrell, J. (1991). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in Physics Learning: Theoretical Issues and Empirical Studies, Proceedings of an International Workshop, Bremen, Germany*, March 4–8, 1991 (IPN, Kiel Germany, 1992) pp. 110–128.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical Issues and Empirical Studies* (pp. 110–128). Kiel: IPN.

Minstrell, J. (2000). Student Thinking and Related Assessment: Creating a Facet Assessment-based Learning Environment. In Pellegrino, J. , Jones, L., and Mitchell, K. (Eds.) *Grading the Nation's Report Card: Research from the Evaluation of NAEP*. Committee on the Evaluation of National and State Assessment of Educational Progress, Board of Testing and Assessment. Washington D.C.: National Academy Press.

Minstrell, J., Anderson, R., & Li, M. (2015). Diagnostic instruction: Toward an integrated system for classroom assessment. In Richard A. Duschl & Amber S. Bismark (Eds.), *Reconceptualizing STEM education: The central role of practices.*

New York: Routledge.

Morphew, J. W., Mestre, J. P., Kang, H.-A., Chang, H.-H., & Fabry, G. (2018). Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course. *Physical Review Physics Education Research, 14*(2), 020110.

Naumenko, O. (2014). Comparison of various polytomous item response theory modeling approaches for task-based simulation CPA exam data. In *AICPA 2014 summer internship project*. Retrieved from naumenko-polytomous-2014.pdf (aicpa.org).

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Potvin, P., Masson, S., Lafortune, S., & Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference. *International Journal of Science and Mathematics Education, 13*(1), 21–43.

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*(2), 91–104.

Smolleck, L., & Hershberger, V. (2011). Playing with science: An investigation of young children's science conceptions and misconceptions. Current Issues in Education, 14. Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view

Stavy, R., Babai, R., Tsamir, P., Tirosh, D., Lin, F.-L., & McRobbie, C. (2006). Are intuitive rules universal? *International Journal of Science and Mathematics Education, 4*, 417–436.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 46*(6), 699–715.

Su, S., Wang, C., & Weiss, D. (2021). Performance of the S-χ^2 statistic for the multidimensional graded response model. *Educational and Psychological Measurement, 81*(3), 491–522.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*(2), 89–112.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339.

Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The diagnoser project: Combining assessment and learning. *Behavior Research Methods, Instruments, & Computers, 36*(2), 234–240.

Thompson, F., & Logue, S. (2006). An exploration of common student misconceptions in science. *International Education Journal, 7*(4), 553–559.

Tjoe, H., & de la Torre, J. (2014). On recognizing proportionality: Does the ability to solve missing value proportional problems presuppose the conception of proportional reasoning? *The Journal of Mathematical Behavior, 33*, 1–7.

Underwood, S., Posey, L., Herrington, D., Carmel, J., & Cooper, M. (2018). Adapting assessment tasks to support three-dimensional learning. *Journal of Chemical Education, 95*, 207–217.

Vosniadou, S., & Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. *Learning and Instructions, 14*, 445–451.

Wang, C. (2021a). Using penalized EM algorithm to infer learning trajectories in latent transition CDM. *Psychometrika*, *86*(1), 167–189.

Wang, C. (2021b). On interim cognitive diagnostic computerized adaptive testing in learning context. *Applied Psychological Measurement*, *45*, 235–252.

Wang, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics, 46*, 58–84.

Manuscript Received: 29 SEP 2023
Accepted: 20 MAR 2024
Published Online Date: 28 APR 2024