A Bayesian Hierarchical Model for Extracting Individuals' Theory-based Causal Knowledge

Atharva Hans
Graduate Research Assistant
School of Mechanical Engineering
Purdue University
West Lafayette, IN

Ashish M. Chaudhari
Postdoctoral Research Associate
Institute of Data, Systems, and Society
Massachusetts Institute Technology
Cambridge, MA

Ilias Bilionis
Associate Professor
School of Mechanical Engineering
Purdue University
West Lafayette, IN

Jitesh H. Panchal
Professor
School of Mechanical Engineering
Purdue University
West Lafayette, IN

Extracting an individual's scientific knowledge is essential for improving educational assessment and understanding cognitive tasks in engineering activities such as reasoning and decision making. However, knowledge extraction is an almost impossible endeavor if the domain of knowledge and the available observational data are unrestricted. The objective of this paper is to quantify individuals' theory-based causal knowledge from their responses to given questions. Our approach uses directed acyclic graphs (DAGs) to represent causal knowledge for a given theory and a graph-based logistic model that maps individuals' question-specific subgraphs to question responses. We follow a hierarchical Bayesian approach to estimate individuals' DAGs from observations. The method is illustrated using 205 engineering students' responses to questions on fatigue analysis in mechanical parts. In our results, we demonstrate how the developed methodology provides estimates of population-level DAG and DAGs for individual students. This dual representation is essential for remediation since it allows us to identify parts of a theory that a population or individual struggles with and parts they have already mastered. An addendum of the method is that it enables predictions about individuals' responses to new questions based on the inferred individual-specific DAGs. The latter has implications for the descriptive modeling of human problem-solving, a critical ingredient in sociotechnical systems modeling.

Keywords: Knowledge Representation, Theoretical Knowledge, Bayesian Inference, Item Response Theory

1 Introduction

The use of scientific knowledge is prominent in engineering education and design. Engineering students use the theoretical knowledge of mechanics of materials, thermodynamics, and control engineering to devise mechanical and electrical components. Designers use scientific knowledge to extrapolate from experiments to real-world applications. Having the ability to quantify individuals' scientific knowledge can advance both engineering education practices and design research. First, this ability would make it possible to assess students' knowledge accurately [1], and to develop personalized educational support tools [2] [3]. Second, scientific knowledge is an essential ingredient of engineering design expertise necessary for design problem framing and problem-solving [4] [5]. A descriptive decision-making model incorporating prior knowledge can better understand how designers carry out inductive and deductive reasoning tasks [6] [7] [8]. Furthermore, quantifying individuals' knowledge structures is essential for understanding design cognition, expert-novice behaviors, and systems that mimic human problem-solving [9].

The representation of scientific knowledge requires quantifying causal knowledge about specific relationships among the concepts that make up a theory. There is a need for approaches that extract such detailed causal knowledge from individuals' responses. Primary methods in student response modeling (e.g., three-parameter logistic model in item response theory [10]) represent student knowledge using a single node, the so-called "ability" [11] [12] [13]. In such methods, a small number of parent nodes (e.g., ability, educational history, family background) predict whether a student will succeed or fail in an exam. But modeling the

ability using a single parameter is not adequate for situations where a large amount of domain knowledge is required, such as in engineering education. Recent advances in student modeling propose dynamic Bayesian networks to explicitly model prerequisite skill hierarchies and yield meaningful instructional policies [14] [15] [16]. Nevertheless, these approaches still lack the methods for statistical inference of individual-specific Bayesian networks from observed exam responses.

To address this knowledge gap, the objective in this paper is to *quantify individuals' theory-based causal knowledge from individuals' responses to given questions.* As the objective suggests, the paper focuses on a specific type of scientific knowledge, called *theory-based causal knowledge*. Theory-based causal knowledge relies on widely-accepted principles for explaining physical phenomena where relations between physical variables are governed using causal relations. Our approach builds on directed acyclic graphs (DAG), i.e., graphs with directed links with paths that form no cycles, representing causal knowledge. For example, consider the DAG in Fig. 1 (far-left) shows, representing the causal knowledge associated with the distortion energy theory of static failure. There are six physical variables, $X = \{F, G, S_y, M, \sigma, n_y\}$, represented as nodes in the graph. Variable F represents the loading applied to a mechanical component with geometry G. The external loading, F, causes the component to developing internal moment M as shown using a link directed from F to M. The internal moment, M further induces normal stress σ. Normal stress σ and yield strength S_y are used to calculate the yield factor of safety n_y. In Fig. 1, you can also find a schematic of the proposed knowledge extraction methodology. We assume an ideal DAG dictates how different physical variables are interconnected for a given theory. Then, we assume that each individual has a unique DAG, not directly observable, and we model the probability that causal relations are correctly identified (prior). Given an individual's graph, we devise a graph-based logistic (GrL) model that maps question-specific subgraphs of an individual's DAG to the probability of responding correctly to the given question (likelihood). In particular, we assume that the probability of a correct response to a question is proportional to the fraction of question-specific causal links—from the true DAG—that an individual knows correctly. Finally, we use hierarchical Bayesian inference to estimate the posterior over individuals' DAGs conditioned on the observed responses to different questions. The advantage of hierarchical Bayesian inference methodology is that it can infer the population- and individual-level uncertainty in model parameters when few observed responses are available. This paper provides multiple improvements in the Bayesian methodology compared to our previous work in Ref [17]. These improvements are noted throughout the paper.

We illustrate the approach by modeling the DAGs of undergraduate engineering students about the theory of fatigue failure of mechanical components. The dataset includes questions testing the students' knowledge about internal stresses, the endurance limit, adjustments to the endurance limit, and the factor of safety against fatigue failure [18].

The results from the study highlight the merits of our approach for quantification of individual-specific causal knowledge as well as for predictions of individuals' responses to unseen questions. Our approach enables the identification of parts of a theory that a subject struggles with and the features they have already mastered, both essential for providing individual feedback for personalized education. Moreover, our approach makes it possible to draw inferences from an individual's estimated knowledge structure to new situations based on the same theory. Such extrapolation is needed because there are many different problems for a given theory that individuals can solve, and circumstances never repeat themselves perfectly.

The organization of this paper is as follows. In Section 2.2, we provide the necessary background on item response theory. Section 3 provides mathematical details of DAGs, a graph-based logistic model, and the hierarchical Bayesian inference approach. In Section 4, we describe the experimental dataset. In Section 5, we present our results and highlight the salient features of the method. Finally, we discuss the implications of these results from engineering education and design. Section 6 summarizes the key conclusions.

2 Related Work

2.1 Knowledge Representation in Engineering Design

Domain-specific knowledge can be structured in different ways, e.g., causal relations, taxonomies, rules, procedural knowledge, etc. [19–21]. Many studies undertake computational approaches for representing knowledge of design processes and design artifacts, e.g., in the product systems design [22, 23]. The goal of these computational studies is to discover generalized and specialized product knowledge from design databases for supporting tool development for improved analogical design. Dong and Sarakar [24] represent complex products and processes as matrices where nodes are product elements and cells are structural, functional or behavioral relationships between nodes. Then, they derive generalized design knowledge as the macroscopic level information from matrix representations using singular value decomposition. With the goal of quantifying a product's innovativeness in terms of component-level decisions, Rebhuhn et al. [25] represent the product design process as the hierarchy of product, function, and components. They use multi-agent models to propagate novelty scores of products down to the component level. Siddharth et al. [26] develop engineering knowledge graph by aggregating entities and their relationships from a patent database. Fu at al. [27] analyze the US patent database and discover different structural forms such as hierarchy and ring. Despite this development, computational approaches for representing and estimating an individual's theory-driven causal knowledge are lacking. The proposed methodology addresses this gap by modeling theory-specific causal knowledge as a probabilistic causal graph and estimating person-specific causal graphs using Bayesian inference.

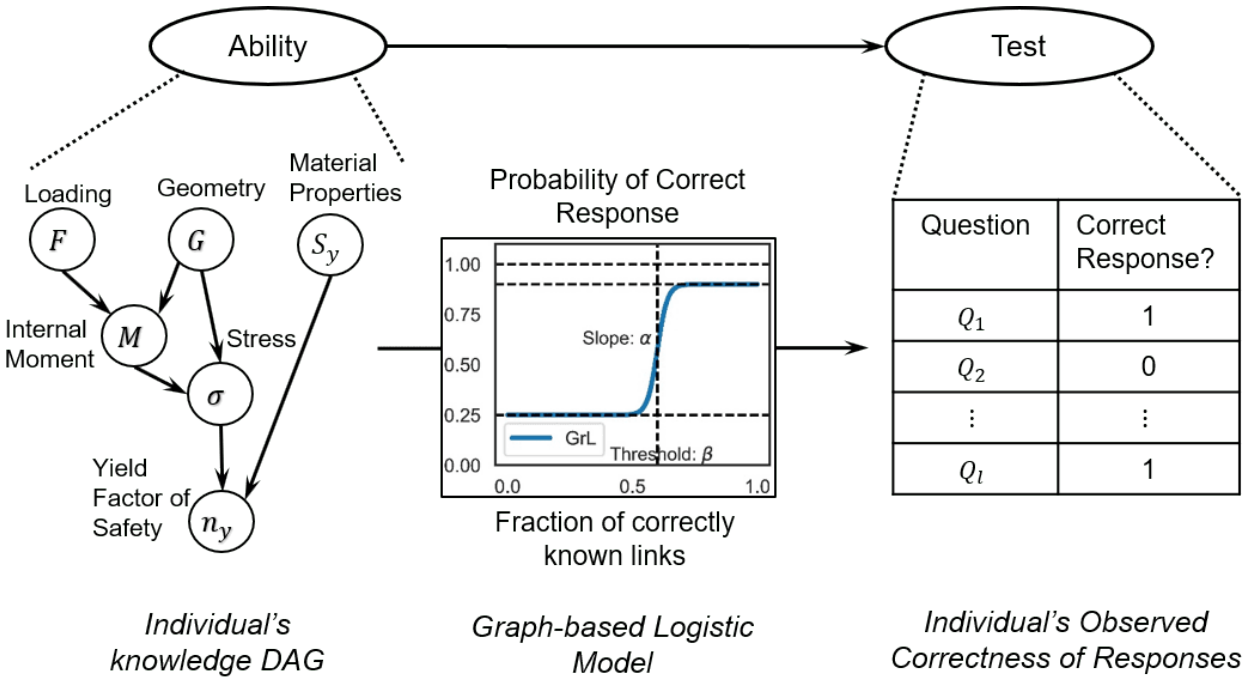


Fig. 1. A schematic for the forward probabilistic graphical method. An example of the question-specific subgraph for calculation of internal moment M comprises of variables (F, G, M).

2.2 Background on Item Response Theory

Item response theory (IRT) describes fundamental principles for formal assessment of individuals' characteristics [28] [10] [29]. IRT is based on the relationship between an individual's performance on test items and the individual's overall ability characteristics, which the test is designed to measure. Several statistical models are used to model individual characteristics and test items. IRT-based models have two standard components: (i) they represent an individual's ability in terms of single- or multi-dimensional latent parameters, and (ii) they represent the probability of correct response as a monotonically increasing function of ability. This function of ability is sometimes called the *item characteristic curve*. An example application of IRT is the force concept inventory which tests the Newtonian concepts along six dimensions such as kinematics, impetus, active force, action-reaction pairs, concatenation of influence, and other effects such as centrifugal forces [30].

Different IRT models are distinguished based on the number of parameters used to define the item characteristic curve. The three-parameter logistic (3PL) model is a basic IRT model for binary response (correct or incorrect). Assuming that R individuals are tested on L questions, the 3PL model has one person-specific ability parameter θ_r, $r = 1, \ldots, R$, and the three question-specific parameters for $l = 1, \ldots L$ are as follows:

1. *Problem discrimination* α_l*:* This measures how the probability of answering a question correctly changes with ability.
2. *Problem difficulty* β_l*:* This measures problem difficulty based on the ability required to get the correct answer. A higher ability to solve a given problem corresponds to greater problem difficulty.
3. *Pseudo-guessing parameter* c_l*:* This accounts for the probability of getting a correct answer by guessing in a multiple-choice question and is one over the number of choices.

Let E_{rl} taking values in $\{0, 1\}$ denote the answer that individual r gives to question l. The probability of a correct answer, i.e., the response likelihood, is:

$$p(E_{rl} = 1|\theta_r, \alpha_l, \beta_l, c_l) = c_l + (1 - c_l)\,\mathrm{sigm}\left(\alpha_l\left(\theta_r - \beta_l\right)\right), \tag{1}$$

where $\mathrm{sigm}(x) = 1/(1 + e^{-x})$ is the sigmoid function. Fig. 2 visualizes the 3PL likelihood function for fixed slope α_l, threshold β, and guessing parameter c_l. The guessing parameter c_l corresponds to the pseudo guessing probability. The slope parameter α_l characterizes the problem discrimination. The threshold parameter β accounts for the problem difficulty. The likelihood of a correct response monotonically increases with a person's ability.

Building on IRT, multidimensional IRT (MIRT) models represent an individual's ability to use more than one dimension are state-of-the-art [31]. In such models, a vector of independent dimensions replaces a unidimensional ability parameter. However, there are limitations to applying MIRT when measuring the interconnected ability characteristics. MIRT models assume that all ability dimensions are required for answering any question correctly, and the probability of a correct response increases with every dimension (*monotonicity assumption*). MIRTs do not allow for the possibility that a question may require only a subset of ability

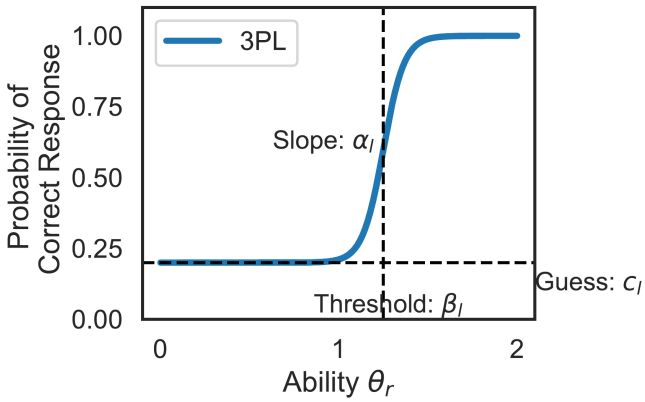


Fig. 2. Graphical representation of the three-parameter logistic (3PL) model given in (1). Slope α_l, threshold β, and guessing parameter c_l represent the problem discrimination, problem difficulty, and the pseudo-guessing probability, respectively.

dimensions to answer correctly. They also assume that the responses of different questions are uncorrelated and based on independent ability dimensions (*local independence assumption*). This assumption does not allow us to make predictions of answers to unseen questions.

IRT models mainly represent unidimensional or independent multidimensional ability parameters. They do not define more complex, interconnected ability characteristics, such as a knowledge graph. The graphical ability representation requires only a subset of the most relevant ability dimensions to explain an individual's performance on a question. Also, current IRT models can only predict responses for test questions used while training the model. The question-specific parameters in the item characteristic curve do not allow predictions to be made on unseen questions. There is also little work in the literature on how we can accurately infer multi-dimensional ability from test responses. The presented approach addresses the above limitations of the existing IRT models in the following manner: (i) it imposes a theory-specific structure on individual-specific ability, which is essential because the nature of human knowledge is best represented through the strong constraints of domain knowledge [32] [8] [33]; (ii) it incorporates Bayesian statistical inference to estimate the individual-specific ability dimensions and represents uncertainty in these estimates. Section 3 presents mathematical details of this approach.

3 Methodology

The proposed method involves the following steps for representing individuals' theory-based causal knowledge: (i) define a structure over the true causal knowledge for a given theory, (ii) model a priori uncertainty in how much individuals know of the true causal scientific knowledge, (iii) defines the relationship between an individual's knowledge and the probability of correct response to theory-related questions, and (iv) characterize the posterior probability over individuals' causal scientific knowledge conditional on the observed question responses.

3.1 Representing Causal Knowledge as Directed Acyclic Graph

We use DAGs to represent causal relationships in a scientific theory. DAGs are graphs consisting of directed links connecting pairs of causally related physical variables with the additional requirement that there are no cyclic paths of directed links. The DAG is an abstraction of structural equations, which may include a diverse set of mathematical models, computer algorithms, etc. [34]. In such structural equations, some variables are inputs, and some are outputs, and the interpretation is that the input variables cause the output variables. These causal relationships are represented with directed links, putting aside the specific equations. For example, Fig. 5 shows a simplified graphical representation of the causal relationships between variables in the stress-based theory of fatigue failure.

An important assumption is that the causal knowledge being modeled is propositional (i.e., the person knows a functional relationship between two variables) rather than procedural (i.e., the person knows a rule) [15]. Further, the physical variables in a DAG can take any real value. Still, the causal links between physical variables are binary, i.e., a link either exists or does not exist. This simple representation still lets us quantify the effects of individuals' knowledge on their responses to theory-related questions.

We also assume that a true knowledge graph, including a set of physical variables and their causal links, is completely known for the theory under study. Subject matter experts can construct such true causal knowledge (e.g., experienced engineers or teachers) or prior knowledge database (e.g., records of theory-based experiments) [12]. Representation of scientific theories in DAGs is feasible because DAGs involve directed links, and their connected paths are acyclic. Every directed link connecting two variables assumes that the starting variable is the cause and the ending variable is the effect. The nonexistence of cyclic paths ensures that a variable cannot be its cause. Note that any feedback loops may be represented by appropriately expanding the DAG in time. Mathematically, let $X = \{x_1, x_2, \ldots, x_N\}$ be the set of physical variables relevant to a given scientific theory. The true knowledge graph for a specific theory is an $N \times N$ binary matrix, $K^{\text{True}} = [k_{ij}^{\text{True}}]$, where k_{ij}^{True} is 1 if the variable x_i is a direct cause of x_j and 0 otherwise.

3.2 Modeling Individuals' DAGs and their Relationship to the Correctness of Responses

We assume that a person's knowledge graph is always a subgraph of the true knowledge graph of the theory. This means that if the theory has no direct link from x_i to x_j, then a person does not makeup such a link. This ensures that a subgraph is acyclic if the true graph is acyclic. We can only test an individual's knowledge in intersection with the true knowledge graph. If the individual has the wrong knowledge graph, they will get the wrong answer. Without the constraint of the true knowledge graph, there will be $N(N-1)/2$ possible links and $2^{N(N-1)/2}$ possible knowledge graphs for N known variables. The inference of the individualistic knowledge graph then becomes intractable even for moderate N.

To quantify a priori uncertainty, we assign a prior probability measure over the space of knowledge graphs. Hierarchical Bayesian modeling further allows us to represent the causal knowledge of each individual and the population in terms of parameters of the prior distribution (hyperparameters). Then, a graph-based logistic model quantifies the effect of individuals' causal knowledge on their responses to theory-related questions. Fig. 3 represents the plate-notation diagram for the proposed hierarchical Bayesian model.

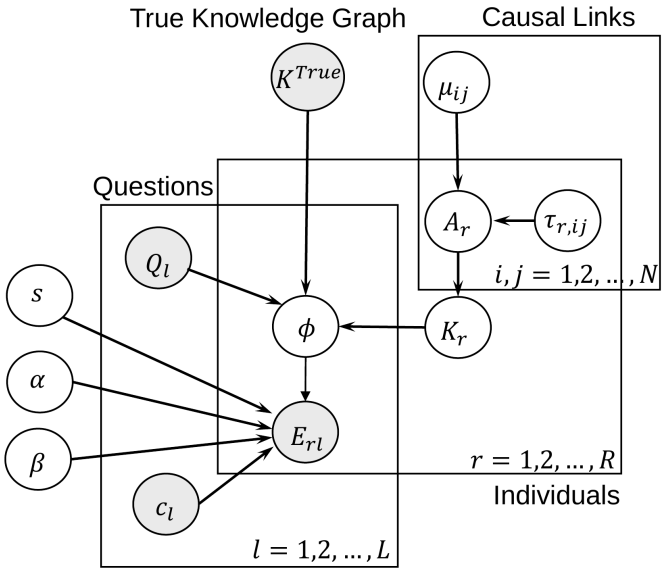


Fig. 3. The plate-notation diagram for the proposed hierarchical Bayesian model. The filled nodes represent the observed variables.

3.2.1 Prior over Knowledge Graphs of Individuals

Let the $N \times N$ matrix $K_r = [k_{r,ij}]$ represent the r-th individual's knowledge about the causal links using the same encoding as in K^{True}. Prior to making any observations, we model our belief that individual r knows about the existence of a true link between x_i and x_j by:

$$p(k_{r,ij} = 1 | k_{ij}^{\text{True}} = 1, a_{r,ij}) = a_{r,ij}, \tag{2}$$

where $a_{r,ij}$ is a hyper-parameter taking values in $[0,1]$. Similarly, the probability that the person does not know that the causal link exists is $p(k_{r,ij} = 0 | k_{ij}^{\text{True}} = 1, a_{r,ij}) = 1 - a_{r,ij}$. Given the matrix of prior link probabilities $A_r = [a_{r,ij}]$, the prior over the causal graph of individual r is:

$$\begin{aligned} p(K_r | K^{\text{True}}, A_r) &= \prod_{i,j:k_{ij}^{\text{True}}=1} p(k_{r,ij} | k_{ij}^{\text{True}} = 1, a_{r,ij}) \\ &= \prod_{i,j:k_{ij}^{\text{True}}=1} a_{r,ij}^{k_{r,ij}} (1 - a_{r,ij})^{1-k_{r,ij}}. \end{aligned} \tag{3}$$

The reader should note that the product is only over the true causal links.

To capture known correlations between different causal links and reduce the number of parameters, we impose an additional assumption on the link probabilities. Namely, we assume that some link probabilities are identical based on an expert belief about whether or not they require the knowledge of the same structural equations. For instance, Fig. 5 represents a knowledge graph in which different subgroups of variables are enclosed in separate boxes. We assume that different directed links connecting variables between two fixed subgroups require the knowledge of the same causal relationships. Then those links are assigned the same link probability. For example, the probability of all links between Marin Factors and variable S_e are equal.

3.2.2 Hyperprior over the Population-level Knowledge Graph

To represent the population's aggregate causal knowledge, we need to assign hyper-priors to the link probability matrices A_r. To that end, we reparameterize the link probability $a_{r,ij}$ using continuous latent variable $\hat{a}_{r,ij}$ in $\mathbb{R}$, which may be interpreted as individual r's link-specific ability. Then, we assign a normal distribution as the hyper-prior over $\hat{a}_{r,ij}$. The mean parameter in $\hat{a}_{r,ij}$'s hyper-prior represents the group means of the population's link-specific ability. Furthermore, rather than sampling an individual's link-specific ability $\hat{a}_{r,ij}$ from a fixed hyper-prior, we represent it using a systematic offset from the group mean. This approach of representing hyperpriors is called "non-centered parameterization," which is an intuitive way of quantifying hierarchical information for large population size and eases exploration of "funnels" in hierarchical models [35]. Let μ_{ij} represent the group means of the population's ability for causal link ij. Let an individual r's offset from the mean ability be $\tau_{r,ij}$. Then, the individual's link-specific ability is defined as $\hat{a}_{r,ij} = \mu_{ij} + \tau_{r,ij}$. Finally, the link probability for individual r is found by passing the latent link-specific ability $\hat{a}_{r,ij}$ through a sigmoid function:

$$a_{r,ij} = \text{sigm}(\mu_{ij} + \tau_{r,ij}), \tag{4}$$

The non-centered reparameterization helps to effectively capture the population-level and individual-specific causal knowledge in terms of model parameters.

The group means μ_{ij} of these hyper-priors are unknown and, also, of potential interest because they represent the population's link-specific abilities. Therefore, we assign prior distributions to them for quantifying uncertainty in their values. The group means μ_{ij} can take any value on the real line and has a normal distribution. The offset parameter $\tau_{r,ij}$ represent an individual-specific quantile in the population-level distribution and is modeled using a normal distribution. The specific shape and scale parameters of the probability distributions are chosen as follows:

$$\begin{aligned} \mu_{ij} &\sim \text{Normal}(0,1), \\ \tau_{r,ij} &\sim \text{Normal}(0,15). \end{aligned} \tag{5}$$

Individual-specific offset $\tau_{r,ij}$ is chosen to have a significant variance to allow a large deviation from the population mean. For a slight variance in $\tau_{r,ij}$, the individual-specific link probability would converge to the population's mean.

This model uses a different prior distribution for every link in the knowledge graph and a non-centered parameterization for representing the hyper-priors. In constrast, our previous work [17] defined same prior distribution over each link and used a centered parameterization for hyper-priors. The rationale for these change are that separate prior distributions allow us to better study the correlation between individuals' link-specific abilities and the non-centered parameterization helps better posterior exploration.

3.2.3 The Likelihood of Correct Responses by Individuals

In contrast to the IRT, our model requires detailed knowledge about the subgraph of the true knowledge graph that each question tests. Each question involves a set of input variables and an output variable to be evaluated. A question using multiple output variables may be divided into separate questions, each with a single output variable. A person answers the question by providing a value of the output variable. The knowledge relevant to answer question l is part of the knowledge graph that connects the input variables to the output variable. Mathematically, we can get the relevant subgraph from the knowledge graph using an $N \times N$ *reduction* matrix Q_l, whose cell value $q_{l,ij}$ is 1 if variable x_i and x_j belongs to the set of relevant input variables and zero otherwise. Then, the true knowledge subgraph for question l is the Hadamard product (elementwise product) of the reduction matrix Q_l and the true knowledge graph K^{True}, denoted as $Q_l \circ K^{\text{True}}$. In matrix $Q_l \circ K^{\text{True}}$ irrelevant variables have been replaced by zeros. Furthermore, we assume that r-th individual's response to question l depends only on the relevant subgraph $Q_l \circ K_r$.

To proceed, we postulate that the probability that an individual's correct response is a function of how close the individual's relevant knowledge subgraph is to the true knowledge subgraph. We propose that the fraction of relevant links that a person correctly identifies represents the person's normalized problem-specific ability. Specifically, for question l, the number of relevant links is $\| Q_l \circ K^{\text{True}} \|_1$, where $\| B \|_1 = \sum_{i,j} |b_{ij}|$ is the sum of absolute cell values of matrix B. Notice since the individual's knowledge graph is a subgraph of the true graph, the number of correctly matched links is simply $\| (Q_l \circ K_r) \circ (Q_l \circ K^{\text{True}}) \|_1$. Thus, the fraction of correctly identified links is:

$$\phi(Q_l \circ K_r, Q_l \circ K^{\text{True}}) := \frac{\| (Q_l \circ K_r) \circ (Q_l \circ K^{\text{True}}) \|_1}{\| (Q_l \circ K^{\text{True}}) \|_1}. \quad (6)$$

This quantity is normalized by the number of all relevant links for a given question.

Similar to the 3PL model in IRT, we have the following parameters to represent the problem-related effects when modeling the probability of correct response:

1. *Slope parameter* α: The sensitivity of the probability of correct answer to the normalized problem-specific ability $\phi(Q_l \circ K_r, Q_l \circ K^{\text{True}})$.
2. *Threshold parameter* β: The minimum fraction of correctly identified links required to answer correctly with probability greater than 0.5. This parameter quantifies the relevance of selected question-specific links for correctly answering a given set of questions.
3. *Pseudo-guessing parameter* c_l: The probability of correct answer by guessing alone. This parameter is dependent on how many choices are available and how responses are evaluated. Generally, the pseudo-guessing probability equals one over the number of possible answers.
4. *Slip parameter* s: This accounts for the possibility that an individual may know the right causal graph, but they may not be able to use it correctly, for reasons such as available information is limited, grading criteria are unknown, or they make numerical errors.

Slope, threshold, and slip parameters are independent of a test question because of how the ability is defined. The problem-specific ability variable $\phi(Q_l \circ K_r, Q_l \circ K^{\text{True}})$ incorporates problem difficulty, unlike the ability variable in the 3PL model. This design allows us to predict responses to unseen questions directly from the individual's knowledge graph and a given pseudo-guessing probability.

Finally, the probability that individual r answers question l correctly is a transformed sigmoid function:

$$\begin{aligned} &p(E_{rl} = 1 \mid Q_l, K_r, c_l, \alpha, \beta, s, K^{\text{True}}) \\ &\quad = a + b \operatorname{sigm}\left(\alpha(\phi(Q_l \circ K_r, Q_l \circ K^{\text{True}}) - \beta)\right). \end{aligned} \quad (7)$$

To calculate scaling factors a and b, we simultaneously solve two constraints that impose a lower bound c_l and an upper bound $1-s$ on the probability of correct response:

$$\begin{aligned} p\left(E_{rl} = 1 \mid \phi(Q_l \circ K_r, Q_l \circ K^{\text{True}}) = 0\right) &= c_l, \\ p\left(E_{rl} = 1 \mid \phi(Q_l \circ K_r, Q_l \circ K^{\text{True}}) = 1\right) &= 1 - s. \end{aligned} \quad (8)$$

Accordingly, we find that the scaling factors are $b = \frac{1-s-c_l}{\operatorname{sigm}(\alpha(1-\beta)) - \operatorname{sigm}(-\alpha\beta)}$ and $a = c_l - b \operatorname{sigm}(-\alpha\beta)$. An individual with zero knowledge relevant to question l might answer correctly, by guessing, with a probability c_l. After accounting for the slip parameter, an individual with complete knowledge will answer correctly with a probability of $1-s$. The previous work in Ref. [17] did not have any parameter (similar to s) to account for individuals' propensity to fail even after having the right knowledge.

We call this likelihood function—the *graph-based logistic* (GrL) model. Parameters α and β are invariant across different questions because each question's explanatory quantity ϕ is normalized. Moreover, global parameters α and β are meant to quantify the suitability of the questions for testing the true knowledge graph, or equivalently, the utility of the true knowledge graph for answering the questions. A large

slope α would signify an abrupt change in the correctness probability and, thus, significant sensitivity to the fraction of correctly identified links. Additionally, the threshold β closer to 1 would imply that individuals need to know all the relevant links to answer the given questions correctly. Finding that α is of the order of 10 and β close to 1 would indicate that the true knowledge graph and given questions are perfectly compatible with each other. For illustration, Fig. 4 visualizes the probability of answering question l correctly as a function of the fraction of perfectly matched links.

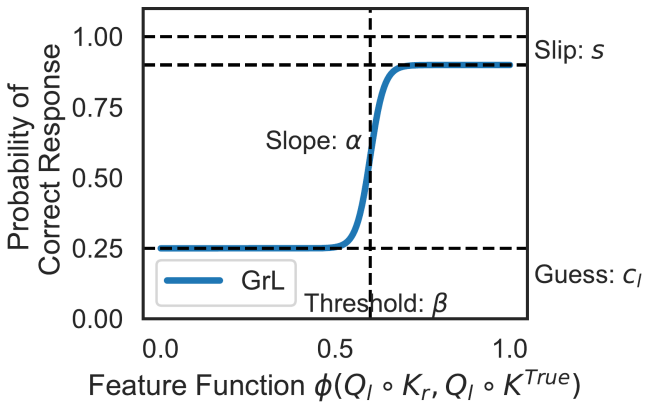


Fig. 4. The graph-based logistic (GrL) model represents the probability of correct response as a function of the normalized problem-specific ability, as defined in (7).

The GrL model assumes that all relevant links have equal importance in answering a question. Other models assign weights to relevant links for quantifying their relative importance [1], which introduces additional model parameters. But the GrL model focuses on quantifying individuals' causal knowledge and does not infer relationships between a scientific theory and given questions. To quantify errors due to incorrect identification of relevant links, the GrL allows for a possibility that knowing a *fraction* of relevant links can result in a correct answer (through threshold β). Therefore, the GrL model is an extension of models requiring all relevant links for a correct answer (the AND-type influence) and models directing at least one link for a correct answer (the OR-type influence) [15].

Like the prior over individuals' knowledge graph, we need to assign priors for model parameters in the likelihood function in (7). For the analysis in this paper, we assume zero chances of pseudo-guessing and take $c_l = 0$ for all questions. The slope α can only take positive values and should be of order ten or higher, implying that the model can effectively differentiate between individuals with different abilities. Since the threshold β can take values between 0 and 1, we assign it a flat distribution. Finally, the slip parameter s should be close to zero. The following are the priors over α, β and s.

$$\begin{aligned} &\alpha \sim \text{Exponential}(0.1), \\ &\beta \sim \text{Beta}(1,1), \\ &s \sim \text{Beta}(0.5,1), \text{and} \\ &c_l = 0. \end{aligned} \tag{9}$$

3.3 Conditioning Individuals' DAGs on Observed Responses

Next, a posterior distribution over causal links in a DAG quantifies the uncertainty about an individual's causal knowledge, given the observed question responses.

3.3.1 Posterior over Individuals' Knowledge Graphs

According to (4), the individual-specific link probabilities are reparameterized using the latent model parameters representing the population-level causal knowledge. The likelihood function in (7) contributes three additional model parameters, slope α, threshold β, and slip factor s (guessing is assumed absent $c_l = 0$). Let $V = \{\mu_{ij}, \tau_{r,ij}\}_{r=1:R;i,j=1:N} \cup \{\alpha, \beta, s\}$ represent the set of these latent model parameters. Then, after observing individuals' responses $E = [E_{rl}]_{r=1:R;l=1:L}$, the following Bayes' rule gives the posterior probability distribution over the model parameters:

$$p(V|E) = \frac{p(E|V)p(V)}{p(E)}. \tag{10}$$

The likelihood of observed question responses $p(E|V) = \prod_{r=1:R;l=1:L} p(E_{rl}|V)$ is defined by the graph-based logistic model from (7). The prior distributions over model parameters $p(V)$ are defined a priori, possibly by assigning independent priors over the model parameters. The model evidence $p(E) = \int_\Omega p(E|V)p(V)\,dV$ is the integration of conditional likelihood probability over the entire parameter space Ω.

3.3.2 Procedure for Sampling from the Posterior

The final step is to sample parameter values from the joint posterior in (10). A common approach for estimating the posterior distribution is Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm. However, sampling the posterior for a hierarchical Bayesian model using Metropolis-Hastings MCMC can be slow [36], especially for large graphical models. Instead, this study employs the No-U-Turn sampler (NUTS) that extends the Hamiltonian Monte Carlo method [37]. The PyMC3 library in a Python environment implements the NUTS [38]. To further speed-up sampling using the NUTS, we reparameterize the binary link variable $k_{r,ij}$ into an unconstrained real variable using a sigmoid function of latent variable $\lambda_{r,ij}$, while ensuring that the link probability $a_{r,ij}$ remains the same. The specifics of the reparameterization are as follows:

$$\begin{aligned} &\lambda_{r,ij}|a_{r,ij} \sim \text{Normal}(\Phi^{-1}(a_{r,ij})), 1), \\ &k_{r,ij} = \text{sigm}(50\lambda_{r,ij}). \end{aligned} \tag{11}$$

The function Φ^{-1} is the inverse cumulative density function of the standard normal distribution. Using a significant value of slope (=50) in the sigmoid function ensures that the reparameterized $k_{r,ij}$ is very close to 0 or 1, and its values can be directly used for likelihood calculations in (6).

4 Dataset

An anonymized dataset for training and testing the proposed model was collected from the responses to questions in a final exam of an undergraduate machine design course. Note that the exam was not explicitly designed for this paper; instead, it was a part of an observational study. The dataset consists of responses to 13 questions by 205 undergraduate mechanical engineering students. The exam tested the students' aggregated knowledge about the concepts of fatigue failure analysis using a circular shaft design problem. The students did not receive monetary incentives for their participation; however, being the final exam, they were motivated to achieve the best possible grade. The exam tested each student's domain-specific knowledge using a total of 13 questions with an overall goal of estimating the factor of safety against fatigue failure. Refer to appendix 6 for the problem statement and a list of questions provided to the students during the exam.

The questions were intended to test the knowledge of causal relationships between variables shown in Fig. 5. Here variable F represents the external loading applied to the steel shaft with geometry G, which is operated at room temperature T. The external loading, F, causes the bar to develop bending moment M. Variable R is the reliability requirement for the bar. The ultimate tensile strength S_{ut} is a material property. The theoretical endurance limit Se' is defined in terms of the ultimate tensile strength S_{ut} using empirical relations [18, sec. 6-7]. The nominal stress σ_o is adjusted by multiplying with the fatigue stress-concentration factor for bending K_f. The adjusted stresses are shown as σ. The endurance limit Se' is adjusted through multiplication by *Marin Factors* for different surface finish conditions, size, loading, temperature, and various factors. This adjusted endurance limit is denoted as Se. Finally, the factor of safety (FOS) is shown as n_f.

Each question included input variables (design parameters) and expected the students to calculate an output variable. Table 1 summarizes the input variables, output variables, and the relevant causal links for all 13 questions. For illustration, consider question 2 and question 9, which are highlighted using loosely spaced dashes and densely spaced dots, respectively in Fig. 5. In question 2, the subjects are required to calculate the bending moment M_{max} in terms of the force F_a. To answer this question correctly, the subjects need to know how the external loading F causes a bar with geometry G to develop internal loads (bending moment) M. Therefore, for question 2, the bending moment M becomes the output variable, and force F and geometry G become the input variables. Similarly, for question 9, nodes Se', k_a, k_b, k_c, k_d, and k_e are the input variables and endurance limit Se becomes the output variable. The input variables cause the output variables and answers to a given question to depend on parent nodes for that question.

Table 1. Causal links required to answer the questions

Question	Design Parameters	Output Parameter	Relevant Causal Links
Q_1	F, G, M	σ_o	(G,M), (G,σ_o), (F,M),(M,σ_o)
Q_2	F, G	M	(G,M), (F,M)
Q_3	G, M, σ_o, K_f	σ	(G,M), (G,σ_o), (M,σ_o), (σ_o,σ), (K_f,σ)
Q_4	S_{ut}	Se'	(S_{ut},Se')
Q_5	G, S_{ut}	k_a	(G,k_a) (S_{ut},k_a)
Q_6	F, G	k_b	(F,k_b),(G,k_b)
Q_7	R	k_e	(R,k_e)
Q_8	F, T	k_c, k_d	(F,k_c), (T,k_d)
Q_9	Se', k_a, k_b, k_c, k_d, k_e	Se	(Se',Se), (k_a,Se), (k_b,Se), (k_c,Se), (k_d,Se), (k_e,Se)
Q_{10}	σ, Se	n_f	(σ,n_f), (Se,n_f)
Q_{11}	F, G	M	(G,M), (F,M)
Q_{12}	G, M, σ_o, K_f	σ	(G,M), (G,σ_o), (M,σ_o), (σ_o,σ), (K_f,σ)
Q_{13}	F, G, M	σ_o	(G,M), (G,σ_o), (F,M),(M,σ_o)

5 Results and Discussion

The results include posterior estimates of model parameters and checks for model accuracy for both the three-parameter logistic (3PL) model and the graph-based based logistic (GrL) model. Specifically, the following four parts constitute the results: i) model checking where we analyze the model fit, ii) comparing the estimates of individual-specific aggregate ability with the observed exam score, iii) evaluating question difficulty in terms of estimated model parameters, and iv) analyzing individuals' causal knowledge in terms of estimated direct acyclic graphs. The analysis uses the training dataset and testing dataset as two partitions of the exam questions to perform model checking. The responses to questions Q_1 to Q_{10} form the training dataset, whereas the answers of questions Q_{11}, Q_{12}, and Q_{13} form the testing dataset. Note that the relevant links for the testing dataset should be a subset of the relevant links for the training dataset.

Similar to the GrL model, this work opts for a Bayesian approach for training the 3PL model. The model parameters in (1) have the following prior distributions, assuming the ability parameter θ_r is a real number, the slope α_l and the threshold β_l are positive, and, $\hat{\mu}$ and $\hat{\sigma}$ are the hyperpriors

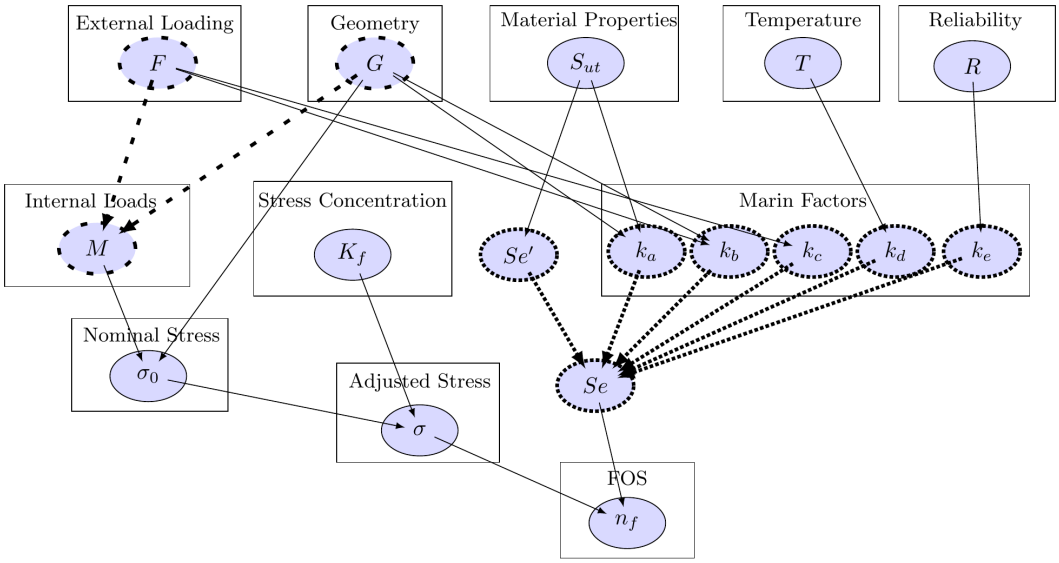


Fig. 5. The true directed-acyclic graph for the scientific knowledge of fatigue failure. The loosely dashed and densely dotted nodes represent the relevant variables for questions 2 and 9, respectively.

over θ_r:

$$\begin{aligned}
\hat{\mu} &\sim \text{Normal}(0,2),\\
\hat{\sigma} &\sim \text{Gamma}(1.5,1),\\
\theta_r &\sim \text{Normal}(\hat{\mu},\hat{\sigma}),\\
\alpha_l &\sim \text{Lognormal}(0,1),\\
\beta_l &\sim \text{Lognormal}(0,1), \text{ and}\\
c_l &= 0.
\end{aligned} \tag{12}$$

The hyperpriors let us estimate the posterior distributions over hyperparameters without the need for manual tuning. The training procedure for hyperparameters is similar to that of the other parameters.

In both the GrL and the 3PL models, we assign value zero to the pseudo-guessing parameter, $c_l = 0$. The pseudo-guessing probability c_l is not estimated from observations, rather it represents the intrinsic uncertainty in randomly guessing the correct answer. Of course, this uncertainty depends on how many choices are available and how responses are evaluated. In our case, the dataset consists of written responses graded by humans. Given the problems can be answered in an arbitrary way, the number of possible answers is infinite and thus the probability of guessing the answer correctly is zero.

We train both models using the NUTS sampler of the PyMC3 library in a Python environment [38]. Posterior parameter samples are computed on Dell compute clusters with two 64-core AMD Epyc 7662 "Rome" processors (128 cores per node) and 256 GB of memory. The computational time for the GrL model, with reparameterization of the binary link variable $k_{r,ij}$, is approximately 180 minutes for 60,000 iterations. This time is significantly less than the case without reparameterization, for which 2000 iterations take approximately 118 minutes. The computational time for the 3PL model is about 16 minutes for 60,000 iterations. The computational times were averaged over four separate runs.

5.1 Checking Model Accuracy

For estimation of predictive accuracy for the 3PL model and GrL model, we use three separate approaches: (i) using an information criterion, precisely Watanabe-Akaike information criterion (WAIC) [39], for finding the in-sample deviance with an adjustment for the number of model parameters, (ii) using posterior predictive checks to perform visual verification of how close the models' predictions are to observed responses (for both training and testing datasets) and calculate test quantities such as a total number of correct answers, and (iii) using prediction accuracy scores which represent the fraction of model predictions that exactly match the observed training and testing data.

The WAIC estimates the expected log pointwise predictive density of observed data $\widehat{\text{lpd}}$ and subtracts a correction term p_{WAIC} based on the effective number of model parameters to adjust for overfitting.

$$\widehat{\text{elpd}}_{\text{WAIC}} = \widehat{\text{lpd}} - \widehat{p}_{\text{WAIC}}, \tag{13}$$

where $\widehat{\text{lpd}}$ is the computed log pointwise predictive density which can be calculated using

$$\begin{aligned}
\widehat{\text{lpd}} &= \sum_{r=1}^{R}\sum_{l=1}^{L} \log p(E_{rl}|E)\\
&= \sum_{r=1}^{R}\sum_{l=1}^{L} \log \int p(E_{rl}|V)p(V|E)dV,
\end{aligned} \tag{14}$$

and $\widehat{p}_{\text{WAIC}}$ is a correction term based on the effective number

of model parameters and is given as

$$p_{\text{WAIC}} = \sum_{r=1}^{R} \sum_{l=1}^{L} \text{var}_{\text{post}} \left(\log p(E_{rl}|V) \right). \tag{15}$$

For further information about predictive information criteria for Bayesian models, refer to [39]. We calculated the WAIC estimates using the PyMC3 library [38].

The in-sample WAIC estimates suggest that the GrL model can better represent the observed training data than the 3PL model. Table 2 presents the values of WAIC, p_{WAIC}, and standard error (SE) for WAIC computations. The lower the WAIC, the better the predictive accuracy. These results indicate that the GrL model has a more significant penalty (has a higher p_{WAIC} value) as compared to the 3PL model, but the overall WAIC for the GrL model (638.86) is still lower than that of the 3PL model (1026.18). This implies that the additional model complexity of the GrL model is justified, at least according to WAIC. Note that in this work, the model fit is better as compared to our previous work [17]. The WAIC score of the former, 638, is lower than the latter, 721, even with a much larger number of model parameters.

According to the posterior predictive checking on the training dataset in Fig. 6, both the GrL model and 3PL model seem to match the observed response patterns. This result is further supported in Fig. 7 where both models adequately explain the total number of correct responses. Bayesian p-values close to 0.5 signify that about half of posterior samples are more significant than the observed test quantity, see Ch. 6 of [40]. For questions Q_3 and Q_{10} in the 3PL model posterior samples (Fig. 6), the prediction accuracy is low. This highlights the 3PL model's ineffectiveness in predicting responses to questions based on a single ability parameter θ.

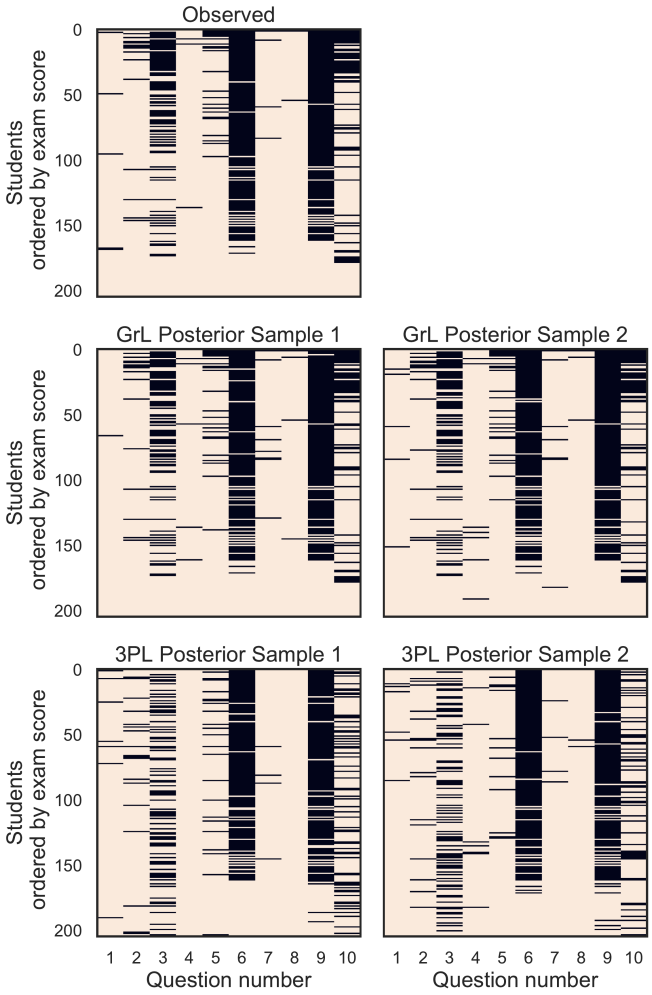


Fig. 6. Posterior predictive checking on the training dataset. Black color represents incorrect responses, and ivory color represents correct responses. Students are ordered by their exam scores.

To investigate the model accuracy at the level of individual students, we look at the predictive accuracy score. Suppose s_l is an individual's response to question l, then the average predictive accuracy score is the fraction of predicted responses that match with the observed response, $\frac{\sum_{i=1}^{1000} \mathbb{1}_{s_l}(\hat{s}_{l,i})}{1000}$. Here $\hat{s}_{l,i}$ is a sample from the posterior and $\mathbb{1}_{s_l}(\hat{s}_{l,i})$ is an indicator function which is 1 if $\hat{s}_{l,i}$ equals to s_l and 0 otherwise. Fig. 8 presents histograms of the student populations' average predictive accuracy score on the training dataset. Under the 3PL model, the average predictive accuracy score is 80% or higher for 84% students. In contrast, under the GrL model, the average predictive accuracy score is 90% or higher for over 95% of the student population.

An essential distinction of the GrL model is its ability to make predictions on unseen questions Q_{11}, Q_{12}, and Q_{13} using the posterior link probabilities of respective relevant links. The 3PL model cannot make such predictions because the question-specific parameters α_l and β_l are unknown for the questions in the testing dataset.

From the results in Fig. 9, the GrL model seems to generally predict the observed patterns correctly except for question Q_{13}. Fig. 10 shows that the number of total correct responses in the testing dataset is lower than the corresponding prediction made using the GrL model. Overall, the average predictive accuracy score for the testing dataset is higher than 90% for approximately 79% of the students, as seen in Fig. 11.

The lower predictive accuracy for some questions in the testing set under the GrL model may be attributed to the inconsistencies in the observed responses. If we assume that any two questions have the same relevant links, then an individual's responses to those questions should be the same (either correct or incorrect). However, this is not always the case in the observed responses. For instance, consider student #34 in Fig. 11 (marked using a star) for whom the average predictive accuracy score is 64%. This student answered training questions Q_1 and Q_2 correctly but answered question Q_3 wrong. Consequently, we should expect a correct response for Q_{11}, an incorrect response for Q_{12}, and a correct response for Q_{13}; because questions Q_1, Q_2, and Q_3 have the same relevant causal links as questions Q_{13}, Q_{11}, and Q_{12} respectively. However, the student's actual response to question Q_{11} is correct, response to Q_{12} is incorrect, and response

Table 2. Model comparison based on Watanabe-Akaike information criterion (WAIC) and the number of actual training parameters

Model	WAIC	SE	P-WAIC	#Parameters
3 Parameter Logistic	1026.18	24.51	58.60	231
Graph Based Logistic	638.86	32.23	245.20	3299

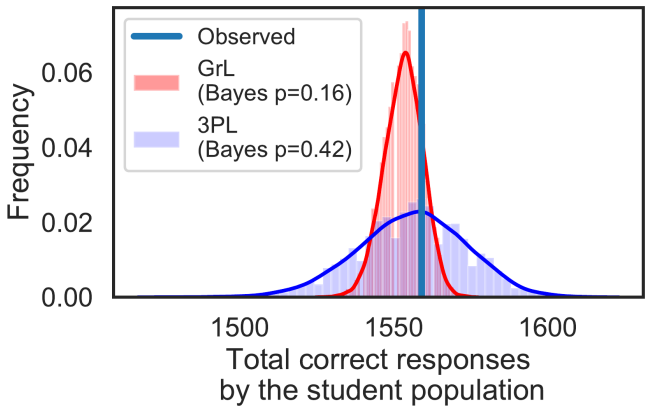


Fig. 7. Both the GrL and the 3PL models adequately explain the total number of correct answers on the training data. Bayes p value close to 0 or 1 would imply that a model is incorrect.

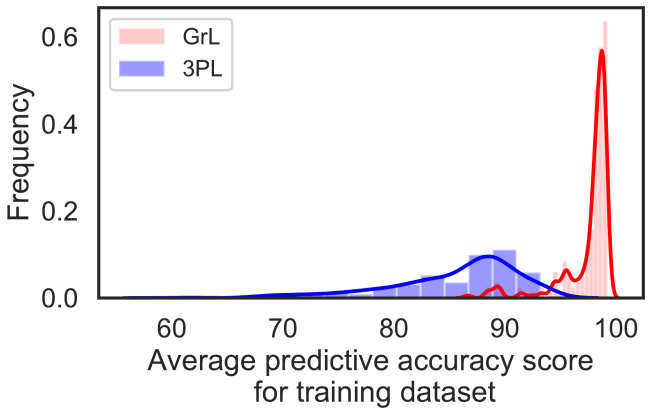


Fig. 8. Predictive accuracy scores of the GrL and 3PL models across the student population. The 3PL model has 80% or higher scores for 84% of the students, whereas the GrL model has 90% or higher scores for over 95% of the students.

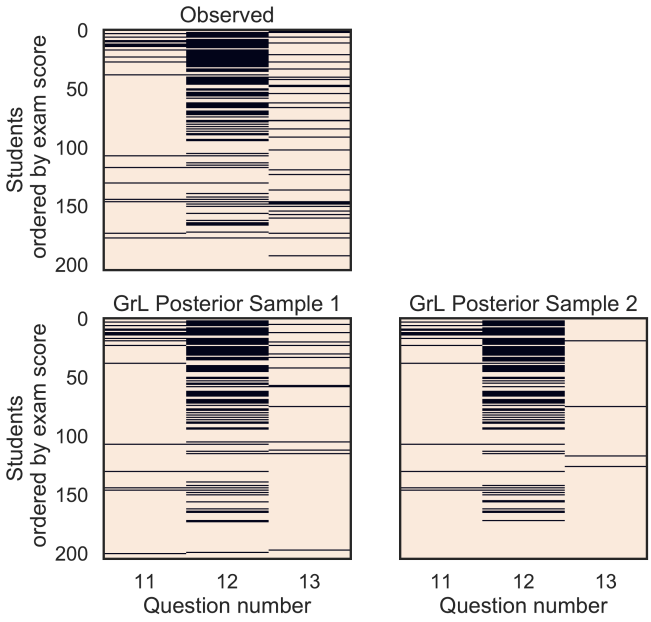


Fig. 9. Posterior predictive checking on the testing data responses. Incorrect responses are highlighted in Black and correct responses in Ivory. The GrL model appears to predict the observed responses correctly, except for question Q_{13}.

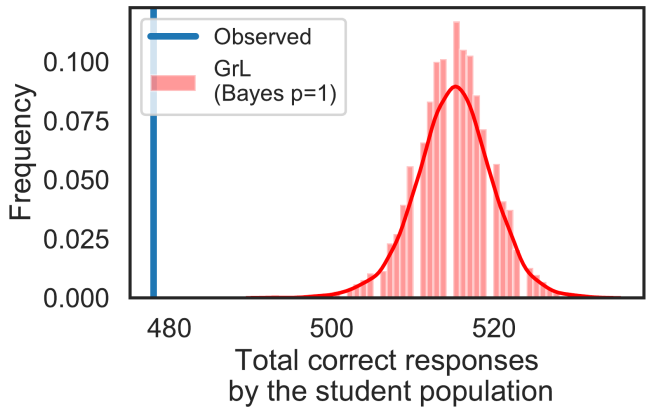


Fig. 10. Posterior predictive checking on the number of total correct responses for the testing data. The observed number of correct answers is lower than the prediction made using the GrL model.

to Q_{13} is incorrect. Total 30 students with scores between 60% and 80% have such inconsistency while answering one of the three testing dataset questions (see Fig. 11). Three students with accuracy scores close to 33% have an inconsistency for two testing dataset questions. Two students with accuracy scores below 30% have inconsistent answers for all testing dataset questions. An exact reason for such errors in the dataset is unclear, but they could arise because students make mistakes even after correctly knowing the causal links.

5.2 Representing Aggregate-level Ability

Since knowledge assessment practices commonly use a single number such as total test score to measure an individual's aggregate ability. We investigate whether model parameters in the 3PL and GrL model can be rearranged to reflect individuals' aggregate ability accurately. In the case of the fatigue questions, the aggregate ability is observed from the students' total exam score, which quantifies overall knowledge of the topic.

In the 3PL model, the threshold θ is by definition di-

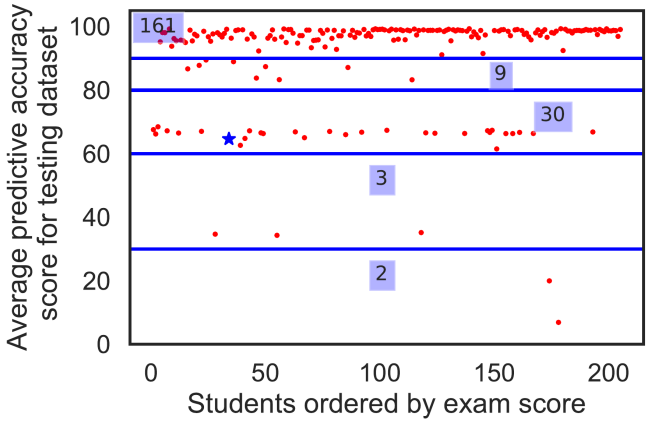
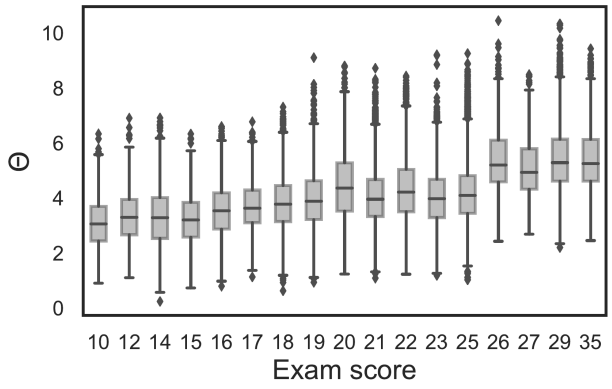


Fig. 11. Posterior predictive accuracy scores for individual students on the testing dataset using the GrL model. The horizontal lines emphasize the 30, 60, 80, and 90 predictive accuracy scores. The highlighted numbers 2, 3, 30, 9, and 161 in the boxes represent the number of students between two horizontal lines or between a horizontal line and maximum (or minimum) predictive accuracy score. The predictive accuracy score is higher than 90% for 161 students.

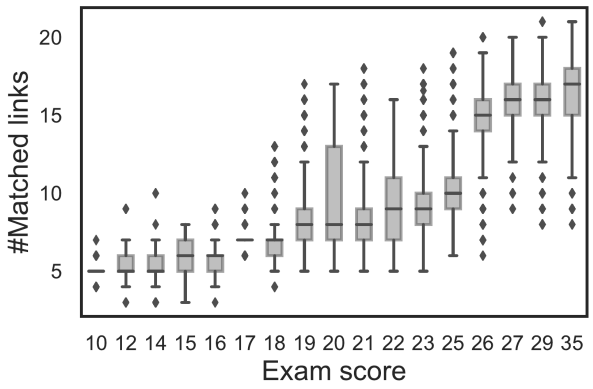
rectly related to an individual's aggregate ability. In the GrL model, a similar aggregate ability measure can be created by taking the intersection of the true and an individual-specific knowledge matrices:

$$\#\text{Matched links} = \| \ (K_r \circ K^{\text{True}}) \ \|_1 \ . \tag{16}$$

Fig. 12a shows that the estimated θ in the 3PL model increases with the exam score. However, this increase is almost linear, and there is considerable uncertainty (variance) in their values. On the other hand, the estimated numbers of matched links in the GrL model (see Fig. 12b) have small uncertainty and cluster the students based on similarity of exam scores while maintaining the positive correlation.



(a) Estimated person-specific ability θ from the 3PL model versus the exam score



(b) Estimated number of matched links from the GrL model versus the exam score

Fig. 12. Representation of students' aggregate-level ability. While both the θ parameter in the 3PL model and the number of matched links in the GrL model have a positive correlation with the exam score, the GrL model appears to strongly distinguish the ability in relation to the exam scores.

5.3 Representing Question Difficulty

Question difficulty is a latent property of different questions used in knowledge elicitation. Because we do not observe question difficulty directly in the given dataset, we analyze comparative question difficulty based on the estimated model parameters.

In the 3PL model, the threshold parameter β_l signifies the difficulty of a question. Based on the posterior estimates of β_l in Fig. 13, we may infer that questions Q_1, Q_2, and Q_4 are easy questions and questions Q_6 and Q_9 are difficult for individuals across the population. This estimation of difficulty is mainly along the lines of the percentage of wrong responses. For instance, high fractions of the students, approximately 73%, get questions 6 and 9 wrong.

The GrL model lacks a specific model parameter to quantify problem difficulty. Instead, the number of relevant links (as listed in Table 1) can proxy for problem difficulty. The higher the number of relevant links required to answer a question correctly, the larger the question can be. However, it is essential to emphasize that problem difficulty varies across different questions and students based on students' abilities. An accurate measure of difficulty would need a precise representation of individuals' question-specific knowledge, achievable by quantifying individuals' knowledge about causal links as discussed in Section 5.4.

The threshold parameter β, in the GrL model, represents a fraction of relevant links required to answer a question pertinent to a given theory. The posterior estimate of threshold β is close to 0.25, as shown in Fig. 14. A posterior β from 1 signifies the partial relevance of selected causal links for predicting a correct answer. This deviation may arise from various external factors. For example, the written responses in the dataset were graded by multiple graders, which may induce variation. Also, a written response can be partially correct, in which case a grader uses their judgment to mark the answer right or incorrect.

The posterior estimates of slip parameter s using the GrL

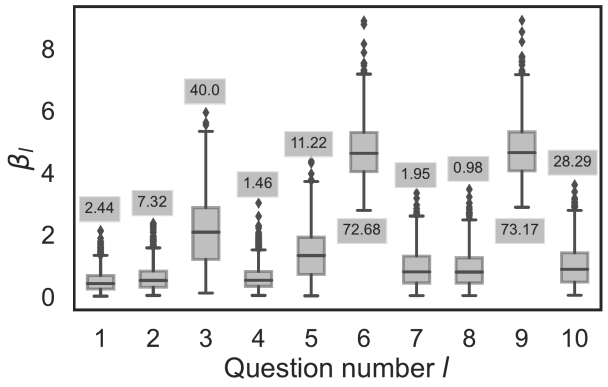


Fig. 13. Posterior estimates of problem difficulty parameter β_l for three-parameter logistic (3PL) model. The boxed numbers highlight the percentage of subjects who answered the related question incorrectly.

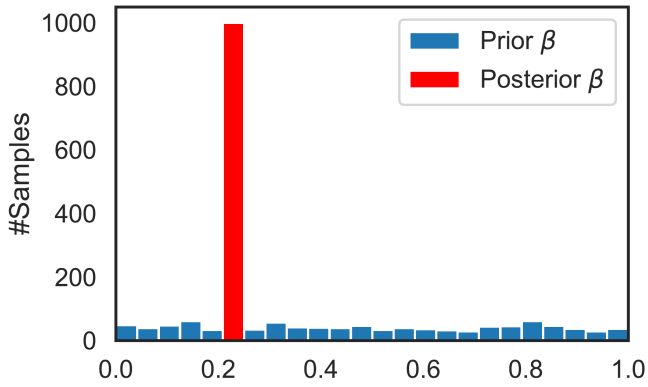


Fig. 14. Comparison of the prior and posterior distribution of the β parameter in the GrL model.

model is of order 10^{-2}, which indicates a low probability for knowledgeable student subjects to answer incorrectly. Further, the posterior estimates of slope parameter α are close to 100, signifying that the model effectively differentiates between student subjects with varying abilities.

5.4 Representing Causal Knowledge

Unlike the 3PL model, the GrL model can quantify causal knowledge in terms of link probabilities. Fig. 15 presents the distributions of estimated link probabilities across different causal links for the entire student population. Here, the x-axis represents the population-level estimate of link probability (colormap corresponds to the x-axis), and the y-axis represents the number of samples. We observe that the students have better knowledge of some links than the others. For example, the students know links (S_{ut}, Se'), (T, k_d), (R, K_e) etc., with high certainty. Conversely, for some causal links, such as (G, k_a) and (G, k_b), probability density is skewed towards 0, signifying poor knowledge of these links. An instructor can utilize this link-specific knowledge to better focus on the concepts that a population might find challenging. Some causal links, such as (K_a, Se) and (K_b, Se), have identical probability density because of the assumption that causal links between fixed sub-groups have equal link probability.

The GrL model is also helpful for categorizing the students based on their knowledge of causal links. Consider two students, a high-scoring student who answered all ten training questions correctly versus a low-scoring student who answered four training questions correctly. The differences in the knowledge structures of these two students are evident from the estimated link probabilities in Fig. 16. The high-scoring student seems to have high knowledge of all relationships such as, (F, k_b), (F, k_c) etc., whereas the low scoring student appears to know only some relationships, such as (S_{ut}, Se'), (T, k_d), (R, k_e), with higher probability. In Fig. 16, it is essential to note that even for the high scoring student, some links, such as (G, M) and (M, σ_o), have low probability. One possible reason for this could be the low threshold β for the GrL model. This might cause the model to assume that a particular link is not required to answer a given question.

Further, for some causal links such as, (Se', Se), (σ, n_f), the model is uncertain at a population level (Fig. 15). The uncertainty in some links could be because they either do not repeat enough times across multiple questions or even when they repeat more than once, they happen to repeat with many other links. For example, links (σ, n_f) and (Se, n_f) appear only in Q_{10}. If an individual gets Q_{10} incorrect, then the model will be uncertain about whether both or one link is weak. On the other hand, the model is quite sure about the link (S_{ut}, Se'). This link happens to be the only link required to answer Q_4 correctly, because of which the model has high certainty about this link. Also, note that the training dataset was a field dataset and was not specifically collected for this model. An ideal set of questions to be designed would have the following features, i) a question tests only one causal link to maximize the learning about the link-specific knowledge, ii) if multiple causal links constitute a single question, they should repeat in other questions so that the trained model has substantial certainty about individual link probabilities.

5.5 Discussion: Implications for Engineering Education and Design

The probabilistic graphical method proposed in this paper has implications for learning and teaching in engineering education. Quantification of students' causal knowledge can support the detailed representation of students' ability [11]. The method provides an in-depth understanding of concepts a given population (or an individual) understands poorly knowledge and the concepts that the people (or an individual) know well. Based on this understanding, instructors and individualized tutoring systems (ITS) [41] [42] can provide personalized feedback to help improve students' knowledge. In the context of adaptive tests, instructors and ITS can potentially use estimated causal knowledge structures to generate new questions with varying difficulty using different combinations of causal links. This would allow instructors to test the same concepts using different questions and help reveal students' knowledge of multiple concepts. The proba-

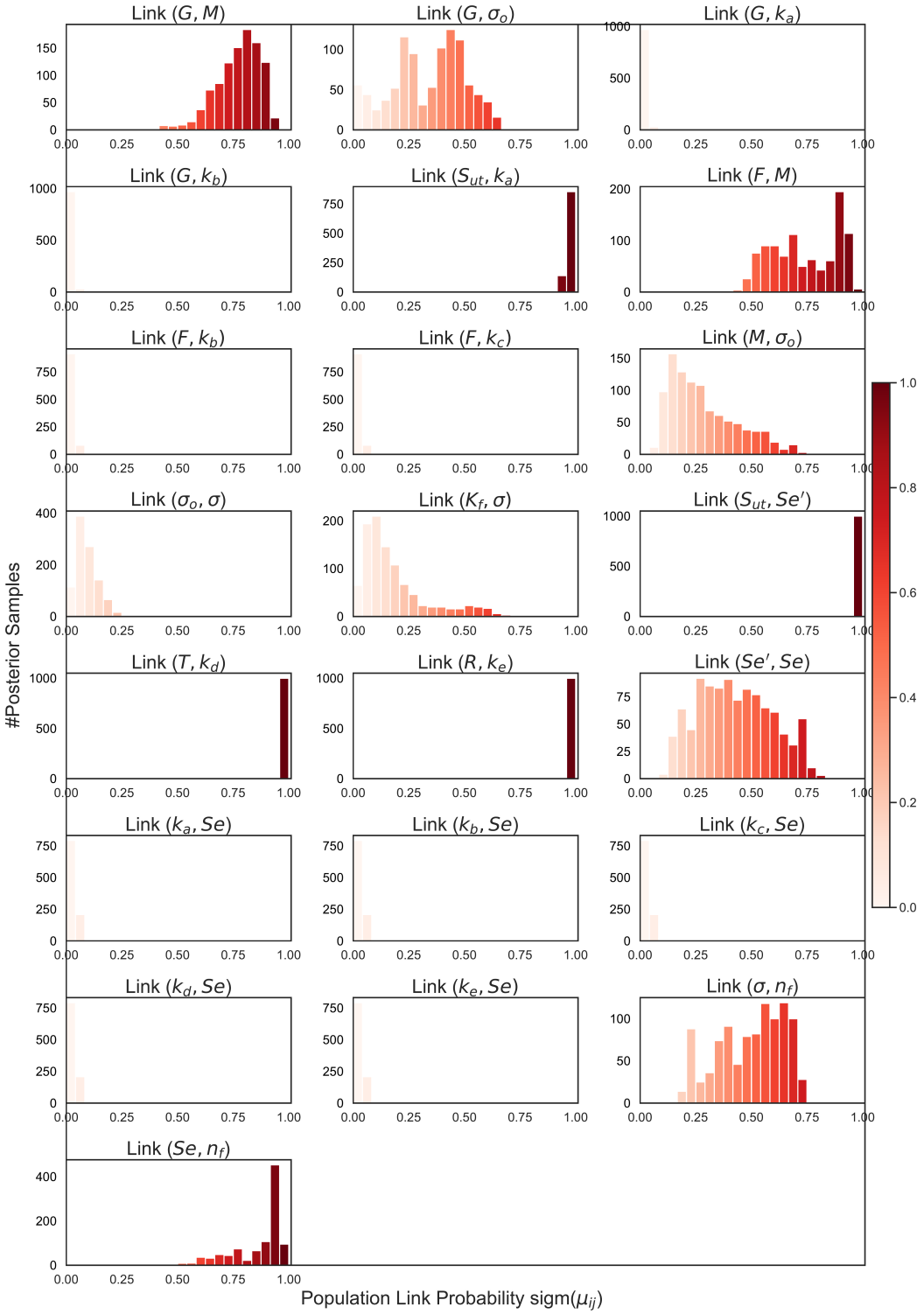
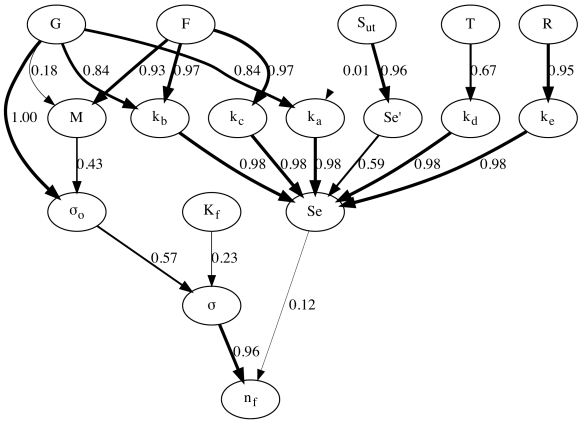


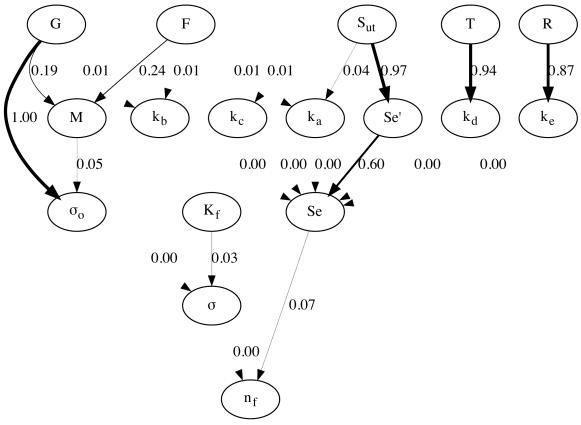
Fig. 15. The posterior distributions of population-level link probabilities using the graph-based logistic (GrL) model.

bilistic graphical method can be utilized for scaffold learning by assisting an instructor in understanding how much support a learner needs to complete learning tasks. The problem-specific scaffolding is achievable by assessing the threshold parameter (β) and the link probabilities (a_{ij}) from the estimated theory-based causal knowledge. This accurate assessment enables the optimal degree of assistance to support the learner's development [43].

Another application of the probabilistic graphical method is modeling the dynamic nature of learning. The method can aid the dynamic Bayesian networks of individuals' learning [44] [45] through quantitative assessment of causal knowledge at different time steps. For example, an individual takes more than one assessment in exams or quizzes



(a) Estimated knowledge graph for a student with ten correct training responses. Larger line thickness represents a greater probability of knowing the link correctly.



(b) Estimated knowledge graph for a student with four correct training responses. A larger line thickness represents a greater probability of knowing the link correctly.

Fig. 16. Comparison of the directed-acyclic graphs representing knowledge structures for high-scoring and low-scoring students. A larger line thickness represents a greater probability of knowing the link correctly. The numbers on the link connecting each node represent the link probability.

for a given course. During these assessments, an individual's estimated DAG can capture the state of individuals' knowledge and indicate how the individual's knowledge increases over time.

The computational modeling of human decision-making in engineering design is another avenue that can benefit from quantifying causal knowledge. Using the estimated DAGs, the expertise research can differentiate experts from novices and test design theories such as novice designers implement situation-independent rules, and experienced designers tend to think in a pattern-based way [46]. For design practitioners, a better understanding of the knowledge structures can help reduce the inefficiencies caused by a poor comprehension of relevant physical variables and their interrelations. A fast inference of an individual's causal knowledge can better design personal support tools that help human designers in decision-making [47] [48] [7] and knowledge-based inductive reasoning [49] [8]. With the quantification of knowledge structures, better human-machine interaction (e.g., corobotics) and improved design of partially automated artificial intelligence (AI) based products and systems that work with humans [9] can be made possible.

To realize the applications above, we have created a tool that implements the proposed method (https://github.com/atharvahans1/knowledge-dag.git). Given a theory-specific true directed acyclic graph (DAG), a set of questions, and individuals' responses to these questions, such a tool would follow the steps in Section 3 and infer DAGs for the population and individuals (similar to Figures 15 and 16). The purpose of such a tool is to augment existing educational tools for knowledge assessment [50] and adaptive tests [51] with predictive functionalities.

6 Conclusions

This paper quantifies individuals' theory-based causal knowledge using an approach based on directed acyclic graphs (DAGs), a graph-based logistic (GrL) model, and hierarchical Bayesian inference. The approach uses relational constraints from a given theory to model individuals' abilities (causal knowledge). It predicts the correct response based on individuals' question-specific causal understanding. This approach is domain-general and can be implemented for any causal theory. In the illustrative study, we tested the approach on engineering students' response data to questions related to fatigue failure. The results suggest that hierarchical Bayesian inference quantifies uncertainty in the population's causal knowledge as well as uncertainty in individual-specific causal knowledge. The posterior estimates of individual-specific DAGs allow us to identify low-knowledge and high-knowledge subjects across different causal links as presented in Fig. 16. Further, the GrL model can leverage the estimated individual-specific DAGs to predict individuals' responses to unseen questions, given that a new question requires the same theoretical knowledge and the pseudo-guessing parameter c_l is pre-defined.

Further work is necessary for validating the performance of the GrL model on multiple-choice questions, for which the responses are likely to have fewer errors from external factors such as variation in grading criteria. Future work should consider improvements in the representation of individuals' causal knowledge, e.g., through modeling knowledge of functional relationships connecting parent variables to a child variable, instead of simply modeling link probabilities. This will also help to better model the complexity of questions. The presented method requires a priori definition of theoretical knowledge in terms of a true knowledge graph. Future extensions could model procedural knowledge by developing novel prior distributions for unconstrained graphs. Additionally, a Bayesian inference tool for causal knowledge representation is required to augment the existing intelligent tutoring systems for educational remediation and existing de-

cision support systems for engineering design.

Acknowledgements

The authors gratefully acknowledge financial support from the National Science Foundation through NSF CMMI Grants No. 1662230 and No. 1728165. Earlier version of this work was presented at ASME IDETC 2020 [17].

Nomenclature

$X = \{x_i\}_{i=1:N}$	A collection of N physical variables relevant to a given theory
$K^{\text{True}} = \{k_{ij}^{\text{True}}\}_{i,j=1:N}$	$N \times N$ binary matrix representing the true knowledge graph for a specific theory
A_r	$N \times N$ matrix of link probabilities representing prior belief about individual r's knowledge
K_r	$N \times N$ matrix representing individual r's knowledge graph with same encoding as K^{True}
Q_l	$N \times N$ binary matrix where ij^{th} cell is 1 if the relationship between variables x_i ans x_j is relevant to question l, or 0 otherwise
E_{rl}	Binary variable denoting whether individual r's response to question l is correct (1) or incorrect (0)
μ_{ij}	The group means of the population's ability for causal link ij
$\tau_{r,ij}$	Individual r's offset from the mean ability μ_{ij}
α	Slope parameter
β	Threshold parameter
c_l	Pseudo-guessing probability for question l
s	Slip parameter
ϕ	A feature function calculating the fraction of correctly identified links, i.e., matching links between K_r and K^{True}

References

[1] Millán, E., DescalçO, L., Castillo, G., Oliveira, P., and Diogo, S., 2013. "Using bayesian networks to improve knowledge assessment". *Computers & Education,* **60**(1), pp. 436–447.

[2] Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., and Ríos, A., 2004. "Siette: A web-based tool for adaptive testing". *International Journal of Artificial Intelligence in Education,* **14**(1), pp. 29–61.

[3] Desmarais, M. C., and Pu, X., 2005. "A bayesian student model without hidden nodes and its comparison with item response theory". *International Journal of Artificial Intelligence in Education,* **15**(4), pp. 291–323.

[4] Dorst, K., 2008. "Design research: a revolution-waiting-to-happen". *Design studies,* **29**(1), pp. 4–11.

[5] Wolmarans, N., 2016. "Inferential reasoning in design: Relations between material product and specialised disciplinary knowledge". *Design Studies,* **45**, pp. 92–115.

[6] Chaudhari, A. M., Bilionis, I., and Panchal, J. H., 2019. "Similarity in engineering design: A knowledge-based approach". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 59278, American Society of Mechanical Engineers, Anaheim, California, USA, p. V007T06A045.

[7] Chaudhari, A. M., Bilionis, I., and Panchal, J. H., 2020. "Descriptive models of sequential decisions in engineering design: An experimental study". *Journal of Mechanical Design,* **142**(8), p. 081704.

[8] Griffiths, T. L., and Tenenbaum, J. B., 2009. "Theory-based causal induction.". *Psychological review,* **116**(4), p. 661.

[9] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J., 2017. "Building machines that learn and think like people". *Behavioral and Brain Sciences,* **40**, p. E253.

[10] De Ayala, R. J., 2013. *The theory and practice of item response theory*. Guilford Publications, New York, NY.

[11] Xenos, M., 2004. "Prediction and assessment of student behaviour in open and distance education in computers using bayesian networks". *Computers & Education,* **43**(4), pp. 345–359.

[12] Conati, C., 2010. *Bayesian Student Modeling*. Springer, Berlin, Heidelberg, pp. 281–299.

[13] Beck, J. E., Chang, K.-m., Mostow, J., and Corbett, A., 2008. "Does help help? introducing the bayesian evaluation and assessment methodology". In International Conference on Intelligent Tutoring Systems, Springer, Berlin, Heidelberg, pp. 383–394.

[14] Millán, E., and Pérez-De-La-Cruz, J. L., 2002. "A bayesian diagnostic algorithm for student modeling and its evaluation". *User Modeling and User-Adapted Interaction,* **12**(2-3), pp. 281–330.

[15] Millán, E., Loboda, T., and Pérez-De-La-Cruz, J. L., 2010. "Bayesian networks for student model engineering". *Computers & Education,* **55**(4), pp. 1663–1683.

[16] Käser, T., Klingler, S., Schwing, A. G., and Gross, M., 2017. "Dynamic bayesian networks for student modeling". *IEEE Transactions on Learning Technologies,* **10**(4), pp. 450–462.

[17] Hans, A., Chaudhari, A. M., Bilionis, I., and Panchal, J. H., 2020. "Quantifying individuals' theory-based knowledge using probabilistic causal graphs: A bayesian hierarchical approach". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 83921, American Society of Mechanical Engineers, Virtual, Online, p. V003T03A014.

[18] Budynas, R. G., Nisbett, J. K., et al., 2008. *Shigley's mechanical engineering design*, Vol. 8. McGraw-Hill New York.

[19] Sriram, R. D., 2012. *Intelligent systems for engineering: a knowledge-based approach*. Springer, London.

[20] Furini, F., Rai, R., Smith, B., Colombo, G., and Krovi, V., 2016. "Development of a Manufacturing Ontology for Functionally Graded Materials". Vol. 1B: 36th Computers and Information in Engineering Conference of *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Me-

chanical Engineers, Charlotte, North Carolina, USA, p. V01BT02A030.

[21] Ming, Z., Nellippallil, A. B., Yan, Y., Wang, G., Goh, C. H., Allen, J. K., and Mistree, F., 2018. "PD-SIDES—A Knowledge-Based Platform for Decision Support in the Design of Engineering Systems". *Journal of Computing and Information Science in Engineering,* **18**(4), 07. 041001.

[22] Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F., and Gao, W., 2013. "The evolution, challenges, and future of knowledge representation in product design systems". *Computer-aided design,* **45**(2), pp. 204–228.

[23] Wu, D., and Gary Wang, G., 2020. "Knowledge-assisted optimization for large-scale design problems: A review and proposition". *Journal of Mechanical Design,* **142**(1), p. 010801.

[24] Dong, A., and Sarkar, S., 2014. "Generalized design knowledge and the higher-order singular value decomposition". In *Design Computing and Cognition'12*. Springer, Dordrecht, pp. 415–432.

[25] Rebhuhn, C., Gilchrist, B., Oman, S., Tumer, I., Stone, R., and Tumer, K., 2014. "A multiagent approach to evaluating innovative component selection". *Design, computing, and cognition*, pp. 227–244.

[26] Siddharth, L., Blessing, L., Wood, K. L., and Luo, J., 2022. "Engineering knowledge graph from patent database". *Journal of Computing and Information Science in Engineering,* **22**(2), p. 021008.

[27] Fu, K., Cagan, J., Kotovsky, K., and Wood, K., 2013. "Discovering structure in design databases through functional and surface based mapping". *Journal of mechanical Design,* **135**(3).

[28] Lord, F. M., and Novick, M. R., 2008. *Statistical theories of mental test scores*. IAP, Charlotte, NC.

[29] Self, J. A., 1994. "Formal approaches to student modelling". In *Student modelling: The key to individualized knowledge-based instruction*. Springer, pp. 295–352.

[30] Hestenes, D., Wells, M., and Swackhamer, G., 1992. "Force concept inventory". *The physics teacher,* **30**(3), pp. 141–158.

[31] Reckase, M. D., 2009. "Multidimensional item response theory models". In *Multidimensional item response theory*. Springer, New York, NY, pp. 79–112.

[32] Tenenbaum, J. B., Griffiths, T. L., and Kemp, C., 2006. "Theory-based bayesian models of inductive learning and reasoning". *Trends in cognitive sciences,* **10**(7), pp. 309–318.

[33] Tenenbaum, J. B., and Griffiths, T. L., 2001. "Generalization, similarity, and bayesian inference". *Behavioral and brain sciences,* **24**, p. 629–640.

[34] Pearl, J., et al., 2009. "Causal inference in statistics: An overview". *Statistics surveys,* **3**, pp. 96–146.

[35] Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., 2003. "Non-centered parameterisations for hierarchical models and data augmentation". In Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, Vol. 307, Oxford University Press, New York, NY.

[36] Goudie, R. J., and Mukherjee, S., 2016. "A gibbs sampler for learning dags". *The Journal of Machine Learning Research,* **17**(1), pp. 1032–1070.

[37] Hoffman, M. D., and Gelman, A., 2014. "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.". *Journal of Machine Learning Research,* **15**(1), pp. 1593–1623.

[38] Salvatier, J., Wiecki, T. V., and Fonnesbeck, C., 2016. "Probabilistic programming in python using pymc3". *PeerJ Computer Science,* **2**, p. e55.

[39] Gelman, A., Hwang, J., and Vehtari, A., 2014. "Understanding predictive information criteria for bayesian models". *Statistics and computing,* **24**(6), pp. 997–1016.

[40] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B., 2013. *Bayesian data analysis*. Chapman and Hall/CRC, Boca Raton, FL. Ch.6.

[41] Bloom, B. S., 1984. "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring". *Educational researcher,* **13**(6), pp. 4–16.

[42] Nwana, H. S., 1990. "Intelligent tutoring systems: an overview". *Artificial Intelligence Review,* **4**(4), pp. 251–277.

[43] Ueno, M., and Miyazawa, Y., 2017. "Irt-based adaptive hints to scaffold learning in programming". *IEEE Transactions on Learning Technologies,* **11**(4), pp. 415–428.

[44] Reye, J., 2004. "Student modelling based on belief networks". *International Journal of Artificial Intelligence in Education,* **14**(1), pp. 63–96.

[45] Manske, M., and Conati, C., 2005. "Modelling learning in an educational game.". In AIED, Vol. 2005, pp. 411–418.

[46] Cross, N., 2004. "Expertise in design: an overview". *Design studies,* **25**(5), pp. 427–441.

[47] Simpson, T. W., Frecker, M., Barton, R. R., and Rothrock, L., 2007. "Graphical and text-based design interfaces for parameter design of an i-beam, desk lamp, aircraft wing, and job shop manufacturing system". *Engineering with Computers,* **23**(2), p. 93.

[48] Chen, W., Hoyle, C., and Wassenaar, H. J., 2012. *Decision-based design: Integrating consumer preferences into engineering design*. Springer Science & Business Media, London.

[49] Heit, E., 2007. "What is induction and why study it". pp. 1–24.

[50] Mislevy, R. J., and Gitomer, D. H., 1995. "The role of probability-based inference in an intelligent tutoring system". *ETS Research Report Series,* **1995**(2), pp. i–27.

[51] Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., and Ríos, A., 2004. "Siette: A web-based tool for adaptive testing". *International Journal of Artificial Intelligence in Education,* **14**(1), pp. 29–61.

A1. Questions for Testing the Knowledge of Fatigue Analysis

Fig. 17 sketches a circular shaft under cyclic loading that the students of a machine design class analyze as part of the final exam. The particular exam includes 13 sub-questions listed in Table 3. Each exam question requires estimating one output variable given the values of the output variable's parent variables. The students have additional information through the following problem statement:

> A circular steel bar is fixed to the floor, as shown in Fig. 17. The bar has an ultimate tensile strength $S_{ut} = 180$ kpsi, a yield strength $S_y = 140$ kpsi, and it has a machined surface. The bar operates at room temperature. The fatigue stress concentration factors for bending and shear at the fillet are known to be $K_f = 2.3$ and $K_{fs} = 1.8$, respectively.

Table 3. Questions on Fatigue Analysis of a Circular Shaft

Question	**Question Statements**
Q_1	For the critical plane at the shoulder identify the critical points on Fig. 17.
Q_2	Expression for the bending moment M_{max} at the critical plane in terms of F_a.
Q_3	Expression for the maximum normal stress adjusted for stress concentration for the critical point (ignore shear stresses due to transverse load) as a function of F_a.
Q_4	Calculate the theoretical endurance limit Se'.
Q_5	Calculate the Marin factor k_a.
Q_6	Calculate the Marin factor k_b.
Q_7	For a reliability of 99% calculate k_e.
Q_8	For the given conditions determine the Marin factors k_c, k_d and k_f.
Q_9	Calculate the endurance strength Se of the bar for a reliability of 99%.
Q_{10}	The magnitude of the load, F_a, in pounds, for which the infinite life fatigue factor of safety at the critical point is $n_f = 1.5$.
Q_{11}	Expression for the bending moment M_{min} at the critical plane in terms of F_a.
Q_{12}	Expression for the minimum normal stress adjusted for stress concentration for the critical point (ignore shear stresses due to transverse load) as a function of F_a.
Q_{13}	Show a plot of stress versus time for Q_3.

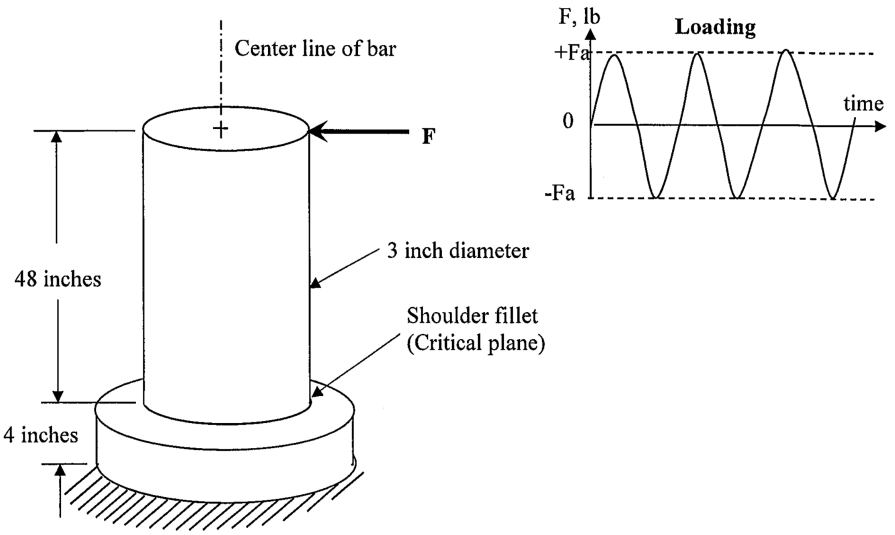


Fig. 17. A circular shaft under dynamic loading F

List of Tables

List of Figures