Abstract—Personalized learning systems use learning and content analytics to trace student concept knowledge over time. We propose Trace-Memory, a machine-learning-based framework for learning and content analytics that incorporates concepts from cognitive load theory. A key limitation of existing learning algorithms is that student response models oversimplify test performance as a function of only the inherent difficulty of the question and the students mastery of the concepts underlying each question. However, research in cognitive science has shown that students performance can also be affected by their working memory capacity and the cognitive load of the question. In this work, we develop a new dynamic student response model that measures students working memory capacity and each questions cognitive load, and includes these terms in the student response model. Preliminary experimental results using the proposed model on synthetic and real-world data have shown promise, but also suggest that further improvements are needed to improve the models sensitivity to noise. Ultimately, this research has the potential to result in personalized learning systems that can identify students with low working memory capacity and improve their educational experience by customizing learning resource presentation to better cater to their needs.

I. INTRODUCTION

THE educational experience for students today is remarkably similar to that of students one hundred years ago. Knowledge is passed from teacher to students in the classroom using a one-size-fits-all approach where students listen to the same lectures and work on the same homework sets. At the end of the semester, students are given feedback in the form of grades and then advanced to the next course.

This teaching method naturally leads to gaps in student knowledge. In the classroom setting, teachers do not have the time or resources to (1) identify specific concepts that individual students are struggling with and (2) tailor classroom lectures to address individual students needs. Students are also not given specific feedback on how they can optimize their study activities to improve their academic performance. Personalized learning systems have emerged as a potentially powerful tool to address these weaknesses in our current education system.

A personalized learning system (PLS) mines educational data (e.g. students graded responses) and uses machine learning algorithms to automatically generate personalized student feedback and recommend optimal remediation/enrichment actions, in effect closing the learning feedback loop. The primary benefit of a machine learning-based approach to PLS is the minimal level of human effort required. This enables these systems to scale to large student populations and diverse classroom subjects, while still providing highly personalized feedback to individual students. The promise of PLS has led the National Academy of Engineering to include it in their list of Grand Challenges [1].

Two primary research thrusts in personalized learning are learning analytics and content analytics. Learning analytics concerns the analysis of education data to infer how well students understand the educational content and to track their progress in understanding over time. Content analytics examines the course content (e.g. lecture videos, textbooks) and quantifies the materials difficulty and quality. A complete PLS would use both learning and content analytics in order to accurately model the student-content interaction and determine an optimal schedule for students to progress through the material.

Given that the students are the central part of the PLS, it is crucial that personalized learning systems accurately model human behavior. Furthermore, these models need to have high interpretability in order to provide useful insights to teachers on students underlying knowledge states. A major limitation of current student response models is that they have hardly incorporated known cognitive science principles for effective learning.
As a result, these models have limited interpretability as measures of student knowledge and may not be capturing the primary features that characterize student behavior. See [2] for an example of previous work where the incorporation of cognitive science principles (e.g. spaced retrieval practice) into a PLS improved student performance.

A. Contributions

Working memory, another concept in cognitive science research (detailed in Section II-B), has also been shown to be linked to student performance. In this paper I attempt to further bridge the gap between learning analytics models and cognitive science principles by designing a new dynamic student response model that has a built in parameter for working memory and cognitive load.

II. BACKGROUND

A. Evolution of Learning Analytics Models

Learning analytics concerns the modeling of student responses - the probability that a given student answered a given question correctly. The input to these models is student gradebook data (binary matrices indicating which student answered which assessment question correctly). Most learning analytics developed to date can be categorized under two families, item response theory (IRT) and knowledge tracing (KT).

One of the simplest IRT models is the Rasch model [3], which treats student performance as a function of the problem difficulty and the learner’s overall ability. While this model has enjoyed great success due to its simplicity, it is also limited in terms of both the model accuracy and interpretability. Multidimensional item response theory (MIRT) captures further details about the student by modeling the students knowledge in each topic.

The limitation of IRT in general is that student concept knowledge is assumed to be static over time. Naturally this is unrealistic as students’ concept knowledge varies after solving problems and interacting with educational content (e.g. watching a lecture). The next dimension in learning analytics models came from KT models, which treat student knowledge as a time dependent variable. KT models are still somewhat over simplistic as questions are modeled to have only a single concept, and student concept knowledge is treated as a binary entity (i.e. students either do or do not understand the concept).

SPARFA-Trace [4] has recently been introduced as a solution to the limited interpretability of IRT and KT models. The SPARFA-Trace model jointly performs time-varying content and learning analytics, and automatically learns i) a mapping that indicates the extent to which each concept is manifested in each assessment, ii) the quality and difficulty of each learning resource, and iii) a dynamic learning profile for each student that estimates their knowledge of each concept over time. SPARFA-Trace uses the following probability model:

\[
Y_{jt} \sim \text{Ber}(\Phi(Z_{jt}))(t, j) \in \Omega_{\text{obs}}
\]

\[
Z_j^t = w^T x_j^t + \mu_j(t), \forall t, j
\]

\[
p(Y_{jt} | c_j^{(t)}) = \Phi((2Y_{jt} - 1)Z_j^t)
\]

In this model, \(Y_{jt}\) denotes the binary graded student responses (\(Y_{jt} = 1\) if student \(j\)’s answer at time \(t\) is correct and \(0\) if incorrect). \(Y\) is a \(Q \times N\) matrix where \(Q\) is the total number of test questions and \(N\) is the total number of learners. Note that there are typically unobserved entries in \(Y\) as students do not answer all of the questions. The matrix \(w\) is a \(Q \times K\) mapping of the extent to which course concepts are included in each question (\(K\) is the total number of concepts in the course). Note, the notation \(i_j^t\) represents the question \(i\) answered by student \(j\) at time \(t\). The matrix \(c\) is the time-varying student concept knowledge matrix of dimension \(K \times N\) that characterizes how well each student understands each concept in the course at each time instant. The inherent difficulty of each question is represented by \(\mu\). Note the \(\Phi(\cdot)\) can represent either an inverse probit or inverse logit link function (for computational efficiency, we use the inverse logit link function in this paper).

There are a few key assumptions built into this model. We assume the number of concepts is typically much smaller than both the number of students and the number of assessment questions, i.e. \(K \ll N, Q\). Furthermore the matrix, \(w\), is constrained to be both sparse and non-negative. Refer to assumptions (A1 - A6) in [5] and [4] for further details on the constraints built into SPARFA-Trace.

SPARFA-Trace models students progress as a linear dynamical system and traces a single students concept knowledge over time using an approximate Kalman filter. The Kalman filter is one part of the overall EM framework, which alternates between estimating the probability distribution of the student concept knowledge states and updating the question-dependent model parameters until convergence is reached.

SPARFA-Trace assumes that the students concept knowledge states are characterized by Markov transi-
tions:
\[ p(c_j^{(t)} | c_j^{(t-1)}) = \mathcal{N}(c_j^{(t-1)} + d_j^{(t-1)}, \Gamma_j^{(t-1)}) \]  

(2)

where \( d \) and \( \gamma \) are latent parameters that characterize question \( i \) the student \( j \) worked on at time \( t \).

B. Working Memory Model

This section and Section II-C provides an overview of the psychology concepts built into the Trace-Memory model.

The working memory model, first proposed in 1974 by Baddeley and Hitch, defines working memory as the set of cognitive resources in the brain used for short-term mental storage and information processing. To further expand on this notion, working memory is the multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction [6]. Working memory is essentially the center stage for cognitive processing that activates relevant information from long term memory, manipulates information held in short term memory storage, and focuses attention on the processing task at hand [7].

Working memory capacity (WMC), the amount of cognitive resources devoted to working memory, has been found to vary between individuals [8]. The ability to measure working memory is important in education settings as numerous studies have found a link between working memory scores and intelligence (e.g. [6]). In addition studies have demonstrated links between working memory and problem solving skills [9] [10], and have suggested that errors in deductive reasoning may be partially explained by limitations of WMC [11].

Although there is no method to directly measure working memory, numerous studies have shown that working memory capacity tasks (further described in Section 4.2) are valid relative measures of working memory (e.g. [13]).

C. Cognitive Load Theory

Cognitive Load Theory builds upon working memory in how it relates to learning. According to cognitive load theory (CLT), every learning task requires a certain amount of working memory, i.e. cognitive load. When a learning task's cognitive load exceeds a student's working memory capacity, the student experiences cognitive overload, and learning is diminished [14]. The learning tasks to which cognitive load theory can be applied include both the lecture material and assessment problems.

A fundamental concept behind CLT is that the total cognitive load for a learning task is the sum of three types of load: intrinsic, extrinsic, and germane load [15].

- Intrinsic load: inherent difficulty of the problem based on level of interactivity between elements in the question as well as the students concept mastery.
- Extrinsic load: the load imposed by purely by the instructional material rather than the concepts themselves. Manipulations in the design of instructional materials can therefore reduce extrinsic load, but the intrinsic load will remain constant.
- Germane load: the depth of processing experienced by the student to address the intrinsic load of the problem. Students who are more engaged during the learning experience, or who, for example, are presented with a richer variety of learning content which engages deeper thinking, will experience higher germane load.

The primary work in cognitive load research has been identifying methods to reduce extrinsic load in teaching material (e.g. [14]). While there have been studies demonstrating that manipulations of the cognitive load of instructional material can positively affect student performance (e.g. see [16]), no direct methods exist to measure cognitive load. Current measurement schemes for cognitive load include self reported ratings from subjects of their perceived amount of mental effort [17] and the use of physiological measures such as heart rate and blink rate [18].

III. TRACE-MEMORY

Motivated by the central role that working memory plays in student performance, we propose Trace-Memory, a variant of SPARFA-Trace that augments the student response model of SPARFA-Trace with concepts from the working memory model and cognitive load theory. Our new model includes parameters that directly relate to the WMC of each student and the total cognitive load of each question. The motivation for our new model is that by providing separate parameters for the students concept knowledge states and their working memory capacity, the updated model can determine whether a student is performing poorly due to poor concept understanding (which corresponds to information stored in long term memory), or due to their working memory being overloaded (due to the difficulty of answering the question given the subjects working memory).
The proposed model to describe the likelihood of an observation is as follows:

\[
p(Y_j^{(t)} | c_j^{(t)}) = \Phi((2Y_j^{(t)} - 1)Z_j^{t})
\]

\[
Z_j^{t} = w_{ij}^{T}c_j^{(t)} + \mu_i + \min(V_j^{t}, 0))
\]

\[
V_j^{t} = \alpha_j - \frac{\beta_j}{1 + \exp(w_{ij}^{T}c_j^{(t)} + \mu_i)}
\]

In this expression, \(\alpha_j\) relates to the working memory capacity of student \(j\), and \(\beta_j\) is related to the cognitive load of question \(i\) answered by student \(j\) at time \(t\). Note that for simplicity, the model does not distinguish between the different types of load (intrinsic, extrinsic, and germane). The key idea here, based on the overload concept in cognitive load theory, is that overload negatively affects the probability that a student correctly answers a question. Mathematically, overload occurs in the model when \(\alpha_j\) is less than the total load expression (a weighted version of \(\beta_j\)). This causes the expression in the minimum function to evaluate to a negative value.

On the contrary, if the student is not overloaded, the total load equals zero and the students working memory and question cognitive load do not affect the probability the student answers the question correctly. Given that the effects of intrinsic load is reduced if the student has higher prior knowledge [15], the weight on \(\beta\) decreases as the students concept knowledge increases and increases as the difficulty of the question increases.

### A. Laplace Approximation

The relationship between students’ concept knowledge and the observed graded responses cannot be directly represented as a Gaussian variable because the observation variables in the SPARFA-Trace and Trace-Memory models are binary-valued student grades [4]. To use the Kalman filtering and smoothing methods, we must therefore determine a reasonable normal distribution that approximates the true distribution.

We use Laplace approximation in Trace-Memory rather than the expectation propagation approach employed in SPARFA-Trace due to the additional non-linearity imposed on the Trace-Memory model by the addition of the working memory and cognitive load terms. Laplace approximation is an iterative algorithm that relates the mean and covariance of the approximated Gaussian distribution to the mode and the Hessian at the mode of the true forward message passing distribution, \(p(c^{(t)} | y^{(1)}, ..., y^{(t)})\) (see Section 3.4.1 of [19]). Note that the Laplace approximation works best when the true distribution is single modal.

We first approximate the log of the posterior distribution with a second order Taylor expansion. We determine the argmax of this expression with respect to \(c\) (i.e. the distribution mode).

From Bayes rule,

\[
p(c|c^{-1}, y) = \frac{p(y|c) \cdot p(c|c^{-1})}{p(y|c^{-1})}
\]

We take the natural logarithm, and substitute for \(\log p(c|c^{-1})\) using (2) (see [4] for further details on the terms used in this expression). Note that throughout this derivation any terms not including \(c\) are discarded since we will maximize the expression with respect to \(c\).

\[
\Psi(c) = \log p(c|c^{-1}, y) = \log p(y|c) + \log p(c|c^{-1})
\]

\[
= \log p(y|c) - \frac{1}{2}(c - \tilde{m})\tilde{V}^{-1}(c - \tilde{m})
\]

The first and second partial derivatives have different expressions depending on whether the student is overloaded.

\[
\nabla_c \Psi(c) = \nabla_c \log p(y|c) - \tilde{V}^{-1}(c - \tilde{m})
\]

\[
\nabla_c \nabla_c \Psi(c) = \nabla_c \nabla_c \log p(y|c) - \tilde{V}^{-1}
\]

If not overload has not occurred, (i.e. \(c - \tilde{m} > \beta \frac{1}{1+\exp(z)} \geq 0\), let \(z = (2y - 1)(w^Tc + \mu)\)

\[
\nabla_c \log p(y|c) = \frac{(2y - 1)w}{1 + \exp(wc + \mu)}
\]

\[
\nabla_c \nabla_c \log p(y|c) = (\nabla_c \log p(y|c))^2 \exp(z)
\]

If overloading has occurred, let \(z = (2y - 1)(w^Tc + \mu + \alpha - \beta \frac{1}{1+\exp(wc + \mu)})\)

\[
T_1 = \frac{\beta \exp(w^Tc + \mu)}{(1 + \exp(wc + \mu))^2}
\]

\[
T_2 = \frac{(2y - 1)w}{1 + \exp(w^Tc + \mu)}
\]

\[
\nabla_c \log p(y|c) = T_2(1 + T_1)
\]

\[
T_3 = (T_2)(w)(T_1)(1 - \frac{2 \exp(w^Tc + \mu)}{1 + \exp(w^Tc + \mu)})
\]

\[
\nabla_c \nabla_c \log p(y|c) = (\nabla_c \log p(y|c))^2 \exp(z)(1 + T_1) + T_3
\]

The non-linearity of \(p(y|c)\) prevents us from directly solving for the location of the distribution mode, so we use Newton’s method.

\[
R = -\nabla_c \nabla_c \log p(y|c)
\]

\[
c^{\text{new}} = c - (\nabla_c \nabla_c \Psi(c))^{-1} \nabla_c \Psi
\]

\[
= (R + \tilde{V}^{-1})^{-1}(Rc + \tilde{V}\tilde{m} + \nabla_c \log p(y|c))
\]
After the updates for $c$ have converged to a new estimate, we calculate the mean and variance of the approximated normal distribution centered at $c$.

$$\bar{m} = c^{new}$$

$$\bar{V} = (R + \bar{V}^{-1})^{-1}$$

B. Parameter estimation

Trace-Memory follows the EM framework of the original SPARFA-Trace model, which alternatives between estimating the concept knowledge states and estimating the model parameters (see Section 4 of [4]). In the Trace-Memory model, the $\alpha$ and $\beta$ parameters are also estimated as part of the M step.

The $\alpha$ and $\beta$ estimates are determined using a back-tracking line search, which was chosen for its simplicity and reasonable convergence time.

1) $\alpha$ estimation: The estimate for $\alpha$ is calculated by minimizing the negative log likelihood distribution. For interpretability, we impose an $\ell_2$ norm penalty on $\alpha$ and require that $\alpha > 0$. This leads to the following optimization problem:

$$\min_{\alpha_j; \alpha_j > 0} \sum_{(i,j) \in Q'} \mathbb{E}_{\epsilon_j}[-\log \Phi((2Y^{(t)}_{j} - 1)Z_j^{(t)})] + \lambda \|\alpha_j\|_2^2$$

$Z_j^{(t)}$ is defined per (3). Following the method used to estimate $w$, we use the Unscented Transform (UT) sigma vectors to approximate the expectation function (see Section 4.4 of [4]). For each round of the backtracking line search, we take an initial step along the gradient of (3) with respect to $\alpha$ using the UT points for $c$. After computing the initial cost, we backtrack until reaching a point where the cost exceeds the initial cost.

2) $\beta$ estimation: The procedure for estimating $\beta$ is identical to the method detailed for $\alpha$. The optimization problem for $\beta$ is:

$$\min_{\beta_j; \beta_j > 0} \sum_{(i,j) \in \mathcal{N}'} \mathbb{E}_{\epsilon_j}[-\log \Phi((2Y^{(t)}_{j} - 1)Z_j^{(t)})] + \lambda \|\beta_j\|_2^2$$

IV. EXPERIMENT

A. Experiments with synthetic data

We first analyze the performance of Trace-Memory using synthetic datasets. We conduct these experiments to verify the Trace-Memory algorithms are correct and to characterize the performance of the model when the sample size is varied. In Section IV-A2 we compare the parameters used to generate the data (i.e. the ”ground truth”) with the parameter values measured by Trace-Memory. Further experiments comparing the results from the synthetic data experiments with real-world data experiments are detailed in Section IV-C.

<table>
<thead>
<tr>
<th>Table I</th>
<th>Dimensions of synthetic datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Students</td>
</tr>
<tr>
<td>1.</td>
<td>DataSet1</td>
</tr>
<tr>
<td>2.</td>
<td>DataSet2</td>
</tr>
<tr>
<td>3.</td>
<td>DataSet3</td>
</tr>
</tbody>
</table>

1) Dataset: The datasets used for the synthetic experiment are generated using the Trace-Memory probability model per (3). The values for $\alpha$ (working memory capacity) are drawn from a uniform distribution $U[0.3,1]$, and the values for $\beta$ (question cognitive load) are drawn from $U[2,4]$. The limits for these distributions are empirically determined such that the overload condition affects roughly half of the values for $Y$ in the gradebook matrix. The values for $w$, $m_0$, and $V_0$ are also drawn from uniform distributions. The knowledge transition parameter, $d$, is fixed to a sequence uniformly increasing from [0,6] over the total assessment time. $\Gamma$ was drawn from a normal distribution, $\mathcal{N}(0,0.01)$. The initial concept knowledge state per student was drawn from a normal distribution specified by $m_0$ and $V_0$, and subsequent states were calculated using $d$ and $\Gamma$ per (2). Each value of $Y$ was determined by taking a majority vote of 20 draws from the probabilistic distribution specified by by (3).

The datasets used in the experiment are detailed in Table I. Dataset 1 is approximately of the same dimensions as the real-world data described in Section IV-B. Dataset 2 has a larger set of questions, and Dataset 3 has a larger set of both questions and students.  

2) Results: We use Kendall’s tau correlation coefficient to compare the true and estimated parameters. Kendall’s tau is a nonparametric measure of how similar two datasets are based on the rank ordering of the data.

For two datasets $(X, Y)$ of length $n$, the Kendall’s tau algorithm compares datapoints at one index in both sets, $(x_i, y_i)$, with another index $(x_j, y_j)$. If $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, the pair is called concordant. If the inequality signs between $x$ and $y$ are opposite, the pair is discordant, and if either pair is equal they are neither discordant nor concordant. Essentially Kendall’s tau measures the probability of finding concordant and discordant pairs using the following formula:

$$\frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{\frac{n}{2}(n - 1)}$$

Values close to 1 indicate that both datasets are ranked in identical order, and values close to -1 indicate the data is ranked in inverse order [20]. Kendall’s tau was chosen for this application as it is unbiased estimator that is
insensitive to outliers and non-linearity. Given that the true relationship between the actual and measured values for WMC is unknown, the Kendall’s tau coefficient provides a reasonable estimate of the similarity between the two datasets. For consistency, we apply Kendall’s tau for all similarity measurements used in this paper.

For all measurement presented here and in Section IV-C, we also use 5-fold cross validation. Five trials of the model are run for each dataset, and for each trial 20% of the data is held out of the training set to measure prediction accuracy. The final measurements were determined by averaging across the five trials.

a) Fixing all Parameters and Estimate α: To understand the model’s accuracy in measuring a single parameter, we modify the Trace-Memory model such that all values except for α are fixed to the ’ground truth’ value used to generate the synthetic data. The model only estimates the values for α. The results are shown in Table II.

This experiment demonstrates the current model is sensitive to noise. The accuracy of Trace-Memory reduces as the number of questions given to each user decreases.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Kendall’s Tau (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data Set1</td>
<td>0.271</td>
</tr>
<tr>
<td>2. Data Set2</td>
<td>0.551</td>
</tr>
</tbody>
</table>

b) Estimating Model Parameters: We now run the full version of Trace-Memory where the model estimates all parameter values. We calculate the Kendalls’ tau correlation for all parameters as documented in Table III.

While α improves with the increase in the dataset size, β actually decreases. This suggests that there may be identifiability issues affecting model performance. With multiple terms in the probability expression, the addition of the minimum function may not be sufficient to allow the model to differentiate between the original terms in the expression, \( w^T c + \mu \) and the new working memory terms.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Kendall’s Tau (α)</th>
<th>Kendall’s Tau (β)</th>
<th>Kendall’s Tau (w)</th>
<th>Kendall’s Tau (µ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data Set1</td>
<td>0.032</td>
<td>0.188</td>
<td>0.325</td>
<td>0.705</td>
</tr>
<tr>
<td>2. Data Set3</td>
<td>0.192</td>
<td>0.083</td>
<td>0.582</td>
<td>0.787</td>
</tr>
</tbody>
</table>

B. Experiments with real-world data

We obtain a real dataset where participants working memory could affect performance. In this experiment, participants complete two tests that are administered online. In the first test, participants are given three working memory capacity tests (See Section IV-B2) to provide an estimate of their working memory capacity, which we use as ground truth values. At a later date, participants read an introductory chapter on digital logic, then took a quiz. The quiz questions are designed to impose different levels of cognitive load on the participants, (see Section IV-B3), and the number of logic operators in each question serves as an estimate of the ground truth cognitive load.

The data collected from participants are fed into the new SPARFA-Trace model. Section IV-C details the results and analysis from the experiment. Section IV-D examines the limitations of this study and identifies potential reasons for the discrepancy between the model-derived and real-world parameters.

1) Participants: The participants for this experiment are Amazon Mechanical Turk workers. In total 46 participants completed the study, with ages ranging from 21 to 56, and educational backgrounds ranging from high school to professional degrees. Participants are required to be U.S. citizens and fluent English speakers. To ensure they were learning new material, participants are requested to not participate in the study if they are familiar with digital logic.

Through the Mechanical Turk interface, participants are given a maximum of 70 minutes to complete each test. Mechanical Turk workers who score a zero on the working memory assessment, indicating that they had most likely simply clicked through the test without trying to answer the questions, are excluded from the results.

2) Working Memory Capacity Tests: Participants complete three working memory capacity tests: operation span, symmetry span, and reading span. The tests are administered online, and follow similar protocols to those detailed in [21] [22]. The reading and symmetry span tasks use the same questions administered in these studies, and the operation span follows the same formula for generating math problems.

The three WMC tests have identical procedures, but cover different subject areas. Combining the scores across three subject domain areas removes effects due to participants domain-specific expertise (e.g. a participant more proficient in literature than arithmetic may perform better on the reading task than the operation task).

The general goal of these tests is to obtain a measure of how well participants are able to store items in memory without being able to transfer this information to long
term memory through rehearsal. This is accomplished by having participants switch between being presented an object to recall and solving a timed secondary processing task. After completing several task-object pairs, participants are asked to recall the objects they were presented in order by selecting from a grid of answer choices. After submitting their selections they receive feedback on their task performance and object recall performance.

The entire sequence of steps outlined above constitute a single item. The item size is the number of task-object pairs given for that round of testing. The idea here is that participants with higher working memory capacity are able to correctly recall objects from the larger item sizes. For each span test, multiple items of different sizes are given to determine at what number of task-object pairs the participant is no longer able to recall the objects they are shown. Although participants’ working memory is only based on their performance on the object recall test, to ensure that they are fully engaged in the secondary task they are encouraged throughout the test to keep their accuracy on the secondary task above 85%. They are also given a time limit on each secondary task to ensure that they do not have time to complete the secondary task and rehearse the recall object.

a) Operation Span: In the operation span task, participants are given a math expression to simplify for the secondary processing task consisting of basic arithmetic operations, and a potential answer. Participants are asked whether the answer is correct or incorrect. The objects they are asked to recall are letters (F, H, J, K, L, N, P, Q, R, S, T, Y). Participants complete two sets each of item sizes 4-8.

b) Reading Span: In the reading span task, participants are given sentences for the secondary processing task (e.g. Andy was stopped by the police because he crossed the yellow heaven) and asked to indicate whether the sentence makes sense. Sentences are changed from sensible to nonsensical by altering one word in the sentence. The objects they are asked to recall are letters (F, H, J, K, L, N, P, Q, R, S, T, Y). Participants complete two sets each of item sizes 4-8.

c) Symmetry Span: In the symmetry span task, participants are given a black and white 8x8 grid (an example is shown below), and asked to indicate whether the image is vertically symmetrical. The objects they are asked to recall are square positions. Participants are shown a 4 x 4 grid where a single square is highlighted in green. For the recall task, participants selected the highlighted squares in the order they were presented. Participants complete two sets each of item sizes 3-5.

Participants’ working memory scores are determined by calculating the unit partial credit score for each test (see [6] for formula details). The overall WMC score is the average of the three span scores.

3) Digital Logic Quiz: Participants complete the digital logic quiz at a separate time from the WMC test to prevent cognitive fatigue. Digital logic was chosen as the topic for the assessment since few outside of electrical engineering are familiar with the subject. Furthermore, the use of logic operators makes it easier to quantify the difficulty level posed by different questions, as demonstrated below. The entire assessment is completed online using Qualtrics software.

The goal of this second session is to obtain test scores from participants that could be run through the Trace-Memory model. The test is designed with the objective to create questions where quiz performance was mostly dependent on participants’ WMC, and thereby reduce the noise in the dataset from other factors such as poor concept knowledge.

Participants first complete an introductory lesson on the fundamentals of digital logic, which teaches Boolean expressions and logic circuits consisting of AND, OR, and NOT logic operators. The lesson is designed to follow guidelines described by [14] [23] to reduce extraneous cognitive load. For example, the lesson begins with basic expressions (e.g. 1 AND 0) and then gradually builds to more complex expressions. The lesson also provides multiple worked examples for each topic that is covered.

To ensure that participants complete the lesson, they are required to view each page for at least thirty seconds. The total lesson time is at minimum five minutes, and participants are told to spend no more than thirty minutes on the lesson. The website allows participants to go back and review previous sections if desired, but they are not able to skip ahead in the lesson. The goal for this setup is to reduce modality effects (e.g. see [24]). Participants are also told they may take notes during the lesson but they must discard them prior to taking the quiz.

After completing the lesson, participants are given 20 questions uniformly sampled from a set of 30 questions. The questions are presented in random order. Of the 30 total questions, 15 are on simplifying Boolean expressions, and 15 are on determining the output of a digital logic circuit. Each question contains a different number of logic operators, which should roughly correspond to the ’intrinsic load’ of the question. The lowest load problem for both the logic circuit and Boolean algebra problems has 2 logic operators, the highest load problem has 23 logic operators, and all other problems range in between.

The use of both digital logic circuits and Boolean algebra in the quiz problem allows for comparisons of
Fig. 1. Sample digital logic problems. The problems in the top row are examples of low cognitive load (2 logic operators), and the problems in the bottom row are examples of high cognitive load problems (10 logic operators).

the impact on cognitive load across problem domains. In the same way that different topic WMC tests are administered to account for domain-specific expertise effects, presenting two different types of questions with the same cognitive load (i.e. the same number of logic operators) allows us to control for domain-specific effects such as perceptual load and spatial ability (see [25] for an example how spatial ability can impact cognitive load).

The quiz is designed to maximize the likelihood that participants become overloaded. To prevent ‘offloading’, participants are told to not write anything down during the test. In addition, they are given a maximum of 45 seconds to complete each problem to prevent problem rehearsal. An example of problem rehearsal is if a participant repeatedly attempts to determine the output of a circuit, by recalculating portions of the circuit the participant may memorize the output of subsections of the circuit. By constraining participants to a short response time, participants are roughly given sufficient time to review the entire circuit only once.

A few other features included in the quiz are worth noting. To prevent participants from simply clicking through questions in order to complete the quiz quickly, participants are required to spend at least 15 seconds looking at the problem before they are allowed to submit their answer. In addition, given that participants were evaluating logic expressions and circuits, the only two possible answers are 0 and 1. To prevent random guessing (which would give participants a quiz score of at least 50%), an ‘I don’t know’ option is included in the answer choices. Participants are asked to select ‘I don’t know’ only if they are planning to randomly guess the answer, and this option is set as the default answer choice. Based on the responses, participants in general selected this option for the more difficult questions.

The use of immediate feedback is another feature added to the quiz. The Trace-Memory model assumes that learners concept knowledge roughly increases after completing each problem. To ensure that participants learn from their mistakes, they are shown the correct answer after each problem and a worked out, step-by-step solution.

C. Results and Analysis

Table IV shows the mean and standard deviations for the three WMC scores, the correlation coefficient between the three WMC scores, and the overall WMC score (calculated by determining the average WMC score for each participant with equal weighting between tests). The correlation between WMC scores is somewhat lower than what has been reported in previous studies (e.g. see [22]), but this may simply be due to the small sample size used in this study.

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Operation</td>
<td>0.80</td>
<td>0.17</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2. Reading</td>
<td>0.74</td>
<td>0.18</td>
<td>0.70</td>
<td>1.00</td>
</tr>
<tr>
<td>3. Symmetry</td>
<td>0.70</td>
<td>0.25</td>
<td>0.42</td>
<td>0.62</td>
</tr>
<tr>
<td>4. Overall</td>
<td>0.75</td>
<td>0.17</td>
<td>0.79</td>
<td>0.90</td>
</tr>
</tbody>
</table>

The Kendall’s tau rank correlation coefficient for $\alpha$ and $\beta$, working memory and cognitive load, are summarized in the Table V. The ‘true’ cognitive load parameters are approximated as the number of logic operators in each problem and the true working memory values are the participants’ WMC score. Based on the result, it is clear that the estimates for working memory and cognitive load have low correlation with the ground truth values.

| Kendall’s Tau for 1. Estimated $\alpha$ against WMC score, 2. Estimated $\beta$ for Boolean expression questions against the true number of logic operators, 3. Estimated $\beta$ for logic circuit questions against the true number of logic operators. Note, standard deviation for each estimate is also shown. |
|-----------------|-----------------|-----------------|-----------------|
|                | K-$\tau$ | SD |
| 1. Working Memory | 0.09 | 0.08 |
| 2. Cognitive Load - Boolean Expressions | 0.34 | 0.15 |
| 3. Cognitive Load - Logic Circuits | 0.51 | 0.17 |

We also compare Trace-Memory against SPARFA-Trace to further characterize the changes in model performance induced by the overload expression.

In Table VI, we compare the prediction accuracy of SPARFA-Trace against Trace-Memory using both
the synthetic and the real-world data sets. In terms of prediction accuracy the two models are roughly similar.

TABLE VI
Prediction Accuracy on 5 fold cross validation

<table>
<thead>
<tr>
<th></th>
<th>Trace-Memory</th>
<th>SPARFA-Trace</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DataSet1</td>
<td>0.876</td>
<td>0.887</td>
</tr>
<tr>
<td>2. DataSet2</td>
<td>0.923</td>
<td>0.923</td>
</tr>
<tr>
<td>3. RealData</td>
<td>0.698</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Using the same datasets we also compare the \( w \) and \( \mu \) estimates from Trace-Memory and SPARFA-Trace expressions using Kendall’s tau to understand if adding overload affects the calculations for the original model parameters. The results suggest the model parameters are still highly correlated even with the addition of new terms for load.

TABLE VII
Kendall’s Tau between Trace-Memory and SPARFA-Trace for \( w \) and \( \mu \).

<table>
<thead>
<tr>
<th></th>
<th>(K-\tau(w))</th>
<th>(K-\tau(\mu))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DataSet1</td>
<td>0.692</td>
<td>0.848</td>
</tr>
<tr>
<td>2. DataSet2</td>
<td>0.717</td>
<td>0.913</td>
</tr>
<tr>
<td>3. RealData</td>
<td>0.848</td>
<td>0.807</td>
</tr>
</tbody>
</table>

In general, while the new model is able to somewhat capture the trends due to cognitive load, it is highly sensitive to the noise in the dataset and the training data size. Given that in a real-world education settings students can implement strategies to offload information, we expect that the effects of overload will be subtle. Therefore with the present model we do not expect to be able to accurately characterize student’s working memory.

Results from both the synthetic and real-world experiments indicate that identifiability issues may partly account for the poor correlation values we observe. While the \( w \) and \( \mu \) values are similar, the Trace-Memory model may simply be capturing the same information in the \( \alpha \) and \( \beta \) parameters. Further work is required to design a new model that better captures the impact of working memory and cognitive load.

D. Limitations of this study

This section notes concerns and limitations posed by this experimental setup that should be addressed through future studies.

1) One inherent challenge of the current set up is controlling for the mental processing speed of participants, as it possible that for some participants the allotted time simply was not sufficient to solve the problem. A potential solution to this problem would be to use high resolution eye tracking algorithms to determine whether participants have finished attempting to solve the problem once.

2) Participants were given a limit of thirty minutes to complete the lesson. It should be taken into consideration whether this was sufficient time to allow participants to fully process the material and commit the information to long term memory before beginning the quiz, since in a typical education students do not begin to solve homework problems immediately after learning a new topic.

3) While we attempted to quantify the cognitive load of each problem by manipulating the number of logic operations, we should note that there are different ways of generating a circuit with the same number of logic operators that may pose varying levels of difficulty. While we believe that the cognitive load of the problem is roughly captured by the number of logic operators (the 2 logic operators problem should have a lower load than the 23 logic operator problem), it is possible problems that are close in the number of logic operators may have varying levels of load.

4) This study does not account for the impact of the problem’s perceptual load. For example, a participant may have not even attempted to solve a problem if it appeared too difficult and they had low interest. This may be accounted for in a future study by using video data to determine effort.

5) One criticism of cognitive load studies posed in [15] was that the participants may not have interest in studying or motivation to study the topic. This certainly may have been the case for this experiment, as participants’ primary motivation was most likely to receive payment for completing the Mechanical Turk study. A future study should account for these effects by either using video data to gage engagement or using assessment data completed by students in a course that was taken for academic credit.

6) While the participants began the study with the same level of expertise in digital logic, it is possible that some participants were able to discover shortcuts to solving the problems. For example, if one of inputs to an AND logic gate is 0, the output of the gate must be 0 regardless of the value of the second input, thereby eliminating half of the problem. By using these types of tricks, participants may have been able to correctly solve a problem regardless of their WMC. Future studies...
should query participants on what strategies they used to solve the problems.

V. CONCLUSION

We have proposed Trace-Memory, a novel machine learning framework that performs learning and content analytics using a student behavior model based on working memory and cognitive load. We have built a model based on our proposed design and tested it using synthetic and real-world data experiments. These preliminary experiments suggest that further improvements to the model are required to reduce the model’s sensitivity to noise and address potential identifiability issues between parameters in the model.

One method to address this and other issues identified in Section IV-D is to record videos of participants taking the study. Using gaze tracking and facial expression recognition algorithms in conjunction on video data and feeding this in to the Trace-Memory model may allow us to better understand how working memory is impacted over time by factors such as stress and engagement. Furthermore, the addition of another set of observations to the model may improve some of the performance issues identified with the model.

ACKNOWLEDGMENT

Many thanks to Richard Baraniuk, Andrew Lan, Phillip Grimaldi, and Debshila Basu Mallick for their guidance and direction, and to Mike Arbelaez for developing the WMC test. Visit our website www.sparfa.com, where you can learn more about the SPARFA project and purchase SPARFA t-shirts and other merchandise.

REFERENCES