Test-size Reduction via Sparse Factor Analysis

Divyanshu Vats
Rice University
dvats@sparfa.com

Christoph Studer
Cornell University
studer@sparfa.com

Andrew S. Lan
Rice University
mrlan@sparfa.com

Lawrence Carin
Duke University
lcarin@sparfa.com

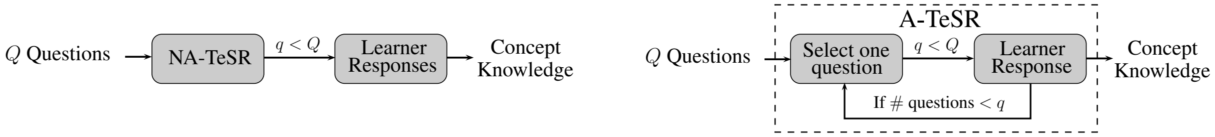
Richard G. Baraniuk
Rice University
richb@sparfa.com

In designing educational tests, instructors often have access to a question bank that contains a large number of questions that test knowledge on the concepts underlying a given course. In this setup, a natural way to design tests is to simply ask learners to respond to the entire set of available questions. This approach, however, is clearly not practical since it involves a significant time commitment from both the learner (in taking the test) and the instructor (in grading the test if it cannot be automatically graded). Hence, in this paper, we consider the problem of *designing efficient and accurate tests* so as to minimize the workload of both the learners and the instructors by substantially reducing the number of questions, or—more colloquially—the *test size*, while still being able to retrieve accurate concept knowledge estimates. We refer to this test design problem as TeSR, short for *Test-size Reduction*. We propose two novel algorithms, a non-adaptive and an adaptive variant, for TeSR using an extended version of the SParse Factor Analysis (SPARFA) framework for modeling learner responses to questions. Our new TeSR algorithms finds fast approximate solutions to a combinatorial optimization problem that involves minimizing the uncertainly in assessing a learner's understanding of concepts. We demonstrate the efficacy of these algorithms using synthetic and real educational data, and we show significant performance improvements over state-of-the-art methods that build upon the popular Rasch model.

1 INTRODUCTION

Testing is a ubiquitous tool used for assessment. In educational scenarios, for example, a test on the prerequisites of a course (or class) can be useful in designing and adapting the course material (Wiggins, 1998; Benson, 2008), and/or for recommending remediation/enrichment for concepts each learner has weak/strong knowledge of (Hartley and Davies, 1976). In self-assessment scenarios, a test can allow learners to effectively plan a course of study in preparing for standardized tests, such as the SAT, ACT, GRE, or MCAT (Loken et al., 2004). In psychological scenarios, a test can be useful in informing a psychologist about characteristics of the testee that pertain to human behavior (Anastasi and Urbina, 1997).

In this paper, we consider the problem of *designing efficient and accurate tests*. In educational scenarios, when given a large database of questions that test learners knowledge on multiple concepts, we are interested in selecting a *small* subset of "good" questions to *accurately* assess the testee's knowledge. Such tests can be useful in reducing the time spent by a testee, whom we refer to as a learner throughout the paper, while still enable accurate assessment of his/her concept knowledge. In psychological scenarios, a smaller list of questions can



(a) Non-adaptive TeSR (NA-TeSR)

(b) Adaptive TeSR (A-TeSR)

Figure 1: An illustration of the two proposed test-size reduction (TeSR) algorithms.

be useful in quickly determining the psychological construct of a testee. In what follows, we refer to such design problems as TeSR, short for *Test-Size Reduction*.

1.1 Contributions

Going beyond the traditional ability-based statistical model (Rasch, 1960) (see Section 1.2 for a detailed discussion), we develop an extended version of the *SPARse Factor Analysis* (SPARFA) framework proposed in (Lan et al., 2014) to model learner responses to multiple-choice questions that test multiple concepts simultaneously. Specifically, while the conventional SPARFA framework associates a learner with a multidimensional vector of parameters that corresponds to their understanding in various concepts, *extended SPARFA* (eSPARFA) in addition associates an ability parameter with each learner. Given the eSPARFA framework, we leverage the theory of maximum-likelihood estimators (MLEs) to formulate TeSR as a combinatorial optimization problem that minimizes the uncertainty of the asymptotic error in estimating both the concept knowledge as well as the ability of each learner.

We propose two TeSR algorithms, one non-adaptive and one adaptive, that approximate the resulting combinatorial optimization problem at low computational complexity. The non-adaptive TeSR algorithm, referred to as *NA-TeSR* (see Figure 1(a) for an illustration of the working principle), reduces the test-size for the traditional setting where all learners are given the same test and the answers to the questions are submitted roughly at the same time. NA-TeSR can be used to rank questions in order of their importance in accurately estimating the concept knowledge of learners. Such a ranking of questions cannot only be used as an aid by instructors when designing questions, but can also be used to improve the quality of questions that are ranked poorly. The adaptive TeSR algorithm, referred to as *A-TeSR* (see Figure 1(b) for an illustration of the working principle), adapts the test questions to *each individual learner*, based on their previous responses to questions. The A-TeSR algorithm is capable of selecting an initial set of questions so that the MLE of the concept knowledge can be computed using as few questions as possible. To demonstrate the efficacy of the proposed TeSR algorithms, we show results for a range of experiments with synthetic data (that enables a comparison to a known ground truth) as well as with two real educational datasets. Experiment results on a real educational dataset show that TeSR can reduce the test size by 40%, without sacrificing predictive performance on unobserved learner responses. See Section 6.4 for more details.

1.2 Related Work

Existing algorithms for selecting small subsets of questions primarily model the learner's responses to questions using item response theory (IRT) (Lord, 1980; Chang and Ying, 1996; Buyske, 2005; van der Linden and Pashley, 2010; Graßhoff et al., 2012). A comprehensive theoretical analysis of the corresponding algorithms has been carried out in (Chang and Ying, 2009). One prominent model used in IRT is the *Rasch model* (Rasch, 1960), where the probability of a learner answering a question correctly is modeled using a scalar ability parameter and a scalar question difficulty parameter. In contrast, the extended SPARFA (eSPARFA) model developed in this paper models the learner's responses to questions using a *multidimensional vector* of not only the ability of a learner, but also the learner's knowledge in multiple concepts that are being tested in the given question set. In this way, eSPARFA more accurately models educational scenarios of tests comprising multiple concepts. Moreover, we show that the proposed TeSR algorithms lead to small tests, where the concept understanding of learners can be measured more accurately when compared to tests designed via the Rasch based model.

The eSPARFA framework performs *factor analysis* (Harman, 1976) on binary-valued graded learner response matrices. Previous factor analysis methods in the educational data mining literature include the Q-matrix method (Barnes, 2005; Desmarais, 2011), learning factors analysis (Cen et al., 2006), multi-way matrix factorization (Thai-Nghe et al., 2011), the instructional factor analysis (Chi et al., 2011), and collaborative filtering item response theory (Bergner et al., 2012). While these methods sometimes achieves good performance in predicting unobserved learner responses, there were no effort on trying to *interpret* the meaning of the estimated factors. In contrary, eSPARFA relies on several unique model assumptions on the factors, enabling the estimation of the learners' concept knowledge.

Some attempts have been made to use *multidimensional item response theory* (MIRT) for designing tests (Luecht, 1996; Segall, 1996; Wang et al., 2011). MIRT typically models learner responses to questions using a multidimensional ability parameter (Reckase, 2009; Ackerman, 1994). However, it has been shown that MIRT models have a highly undesirable property where a learner's ability may decrease after having answered a question correctly (Hooker et al., 2009; Jordan and Spiess, 2012). Thus, the questions selected through an MIRT-based approach are not necessarily useful for estimating the concept knowledge of learners. Furthermore, past work in selecting questions, both using IRT and MIRT, has mainly focused on adaptive methods, where future questions are selected based on prior responses. Our nonadaptive method for selecting questions is novel and appropriate for settings where all learners answer questions at the same time. This is the case, for example, in various massive open online courses (MOOCs) (Martin, 2012; Knox et al., 2012).

Finally, a related, but slightly different, problem to TeSR is that of designing intelligent tutoring systems (ITSs). In ITSs, the main goal is to provide instruction and feedback to learners using a computerized system without any human intervention (Anderson et al., 1982; Brusilovsky and Peylo, 2003; Stamper et al., 2007; Koedinger et al., 2012). One form of ITSs, employed in systems such as the Algebra Tutor (Ritter et al., 1998), the Andes Physics Tutoring System (VanLehn et al., 2005), and the ASSISTment (Feng and Heffernan, 2006), is to ask learners to answer questions associated with a concept, provide feedback, and iterate with different questions until the system believes that the learner understands the concept. Knowledge tracing (Corbett and Anderson, 1994), and its numerous variants (Baker et al., 2008; Pardos and Heffernan, 2011), are popular tools used in an ITS to track learner performance after questions are answered. Although ITSs perform some form of adaptive testing, the main goal in designing questions in an ITS is to teach a learner concepts through a series of questions and associated feedback. In contrast, the main objective in TeSR is to design a test with as few questions as possible, which allow one to accurately assess the concept knowledge of a learner. Nevertheless, we believe that the TeSR methods can be incorporated into ITS models in order to improve their performance

Table 1: Main parameters of the extended SPARFA (eSPARFA) model.

Notation	Description
$\mathbf{W}$	Sparse matrix with non-negative entries that characterizes the relationship between questions and knowledge components
$\boldsymbol{\mu}$	Vector that specifies the intrinsic difficulty of each question
$\mathbf{c}^*$	Vector that represents a learner's ability in each knowledge component
a^*	Scalar that measures a learner's overall ability

using an extension of the methods in (Pardos and Heffernan, 2011).

1.3 Paper organization

The remainder of the paper is organized as follows. Section 2 summarizes the extended SPARFA framework and formulates the TeSR problem. Section 3 describes our data driven approach to approximate the TeSR problem. Sections 4 and 5 detail the non-adaptive and adaptive TeSR algorithms, respectively. Section 6 presents experimental results, and Section 7 concludes the paper.

2 Problem Formulation

In this Section, we formulate the test-size reduction (TeSR) problem, where we select a subset of questions such that the selected questions can accurately assess the concept knowledge of a learner. Although we formulate TeSR in the educational context, TeSR also applies in more general settings such as psychological surveys.

Section 2.1 summarizes the extended sparse factor analysis (eSPARFA) framework, which we use to model learner responses to questions. Section 2.2 formulates the TeSR problem using the eSPARFA framework.

2.1 Extended SPARFA model

Suppose a question set contains Q questions that test knowledge from K knowledge concepts. For example, in a high-school mathematics course, questions could test knowledge from concepts like quadratic equations, trigonometric identities, or functions on a graph. Following the terminology put forward in (Corbett and Anderson, 1994), we refer to these concepts as *knowledge components*.

The original SPARFA framework introduced in (Lan et al., 2014) associates two sets of parameters with each question. The first set of parameters is a column vector $\mathbf{w}_i \in \mathbb{R}_+^K$, where $\mathbb{R}_+$ is the set of non-negative real numbers. The vector $\mathbf{w}_i$ models the association of question i to all K knowledge components. Note that, each question can be linked to multiple knowledge components. For example, solving the equation $x^2 - \cos^2(x) = \sin^2(x) + x$ for $x \in \mathbb{R}$ involves knowledge of both quadratic equations and trigonometric identities. To model this, the j^{th} entry in $\mathbf{w}_i$, which we denote by w_{ij}, measures the association of question i to knowledge component j. The SPARFA model assumes that this association cannot be negative, i.e., $w_{ij} \geq 0$, which means that solving question i cannot reduce the understanding of knowledge component j. Furthermore, if question i does not test any skill from knowledge component j, then $w_{ij} = 0$. To succinctly represent the *question–knowledge component* interactions among all Q questions, we concatenate the column vectors $\mathbf{w}_i$, $i = 1, \ldots, Q$, to form the $Q \times K$ matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_Q]^T$, where the superscript ^{T} stands for the transpose. From the assumptions on $\mathbf{w}_i$ above, we see that $\mathbf{W}$ is, in general, a *sparse matrix* with *non-negative entries*. The second parameter associated with each question is a scalar $\mu_i \in \mathbb{R}$ that represents the intrinsic difficulty of the i^{th} question; the vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_Q]^T$ contains the intrinsic difficulties for each question. In what follows, a larger (smaller) μ_i designates an easier (harder) question.

Next, we define the parameters associated with a learner answering questions. It is these parameters that we are interested in estimating using a small subset of the Q questions. In the original SPARFA model (Lan et al., 2014), the authors assumed that a learner can be modeled using a $K \times 1$ column vector $\mathbf{c}^* \in \mathbb{R}^K$, that measures the ability of a learner in the K knowledge components. In the *extended SPARFA* (eSPARFA) model used in this paper, a learner is modeled not only by the concept knowledge vector $\mathbf{c}^*$, but also by a scalar ability parameter $a^* \in \mathbb{R}$. The properties and advantages of this additional ability parameter in eSPARFA are discussed in Remarks 1 and 3 below. See Table 1 for a summary of the parameters associated with the eSPARFA model.

To model the interplay between $\mathbf{w}_i$, $\boldsymbol{\mu}$, $\mathbf{c}^*$, and a^*, let Y_i be a random variable that denotes the graded response of the learner to question i. If we assume that $Y_i \in \{0, 1\}$, which denotes whether a learner provides a correct (corresponding to 1) or incorrect (corresponding to 0) response, then eSPARFA models the graded response Y_i as

$$\mathbb{P}(Y_i = 1 \,|\, \mathbf{w}_i, \mu_i, \mathbf{c}^*, a^*) = \Phi(\mathbf{w}_i^T \mathbf{c}^* + a^* + \mu_i), \tag{1}$$

where $\Phi(x)$ is the inverse logistic link function defined as $\Phi(x) = (1 + \exp(-x))^{-1}$, $\mathbf{w}_i \in \mathbb{R}_+^K$, $\mu_i, a^* \in \mathbb{R}$, and $\mathbf{c}^* \in \mathbb{R}^K$. We note that the eSPARFA framework can be modified to consider the inverse probit link function, ordinal graded response data (e.g., from tests with partial credit), or categorical responses (e.g., from surveys); see (Lan et al., 2013) for the details. Before formulating the TeSR problem based on eSPARFA, we make some important remarks.

Remark 1 (Rasch model)**.** We point out that eSPARFA corresponds to a generalization of item response theory (IRT) building upon the Rasch model (Rasch, 1960). In particular, if $K = 0$, then the eSPARFA model (1) reduces to the Rasch model, i.e., the probability of answering a question correctly solely depends on a student's ability and the intrinsic difficulty of a question. Several extensions of the Rasch model, also known as the 1PL model, have been proposed in the literature (Baker and Kim, 2004). The 2PL model assumes that questions can be modeled by a difficulty and a discrimination parameter. This discrimination parameter is the degree to which a question discriminates between learners with varying abilities. The 3PL model includes, in addition, a guessing parameter with every question signifying the extent to which learners will make a guess when answering that question. The extension of the eSPARFA framework to both the 2PL and the 3PL model is straightforward, which is why we focus mainly on the 1PL model. The eSPARFA model is also related to cognitive diagnosis model (Templin and Henson, 2006). In particular, the $\mathbf{W}$ matrix in cognitive diagnosis models has binary or categorical entries, while the entries of $\mathbf{W}$ in the eSPARFA model has real-values entries.

Remark 2 (Interpretability of eSPARFA)**.** The key assumption in the eSPARFA model, which was introduced in (Lan et al., 2014), is that the matrix $\mathbf{W}$ is sparse with non-negative entries. The sparsity assumption says that the questions do, in general, not test knowledge from all knowledge components, but only a few of the knowledge components. The non-negativity assumption allows for the knowledge component vector $\mathbf{c}^*$ to be interpretable. In particular, if $\mathbf{c}_j^*$ is large and positive (small and negative), and a learner answers a question that only tests knowledge from knowledge component j, then the probability of answering the question will likely be closer to one (zero).

Remark 3 (Ability parameter)**.** The eSPARFA framework extends SPARFA in (Lan et al., 2014) by adding the ability parameter a^*. In the literature, the introduction of a^* is sometimes referred to as a random effect (Kreft and de Leeuw, 1998). In practice, the need for using a^* depends on the data available for parameter estimation and/or the number of concepts associated with the questions. Some motivations for introducing this additional parameter are given as follows:

(i) If $\mathbf{w}_i$ is estimated to be a vector of zeros (see Remark 4 for how $\mathbf{w}_i$ is estimated), then question i does not test knowledge from any of the knowledge components. In such cases, SPARFA deems the question irrelevant, since the probability of answering the question correctly will not depend on the learner-dependent parameters but only on the intrinsic difficulty. This situation, however, is evidently not desirable in a statistical model, as the provided responses naturally depend on the learner's abilities.

(ii) The eSPARFA framework characterizes the overall ability of a learner across all knowledge components. Such information is not necessarily conveyed in the concept knowledge vector of the original SPARFA framework. For example, consider a test containing only difficult questions testing knowledge from three knowledge components. If a learner answers a *small* number of questions incorrectly, all from a single knowledge component, then SPARFA would estimate the learner's concept knowledge in this component to be relatively weak when compared to the learner's concept knowledge in other components. However, the information that the learner's overall ability is high (since they answered most of the hard questions correctly) is lost when extracting only the concept knowledge vectors. In contrast, the eSPARFA framework is able to characterize both, the overall ability as well as the individual concept knowledge.

We emphasize that the ability parameter may not be needed in some settings and this can be tested when performing parameter estimation (see Remark 4). In such cases, all the algorithms we introduce for test-size reduction will still apply.

Remark 4 (Identifiability and parameter estimation)**.** Given graded response data from multiple learners, the parameters $\mathbf{W}$ and $\boldsymbol{\mu}$ in the eSPARFA model can be estimated using suitably modified versions of the SPARFA-M or SPARFA-B algorithms proposed in (Lan et al., 2014). In all our simulations, we use the SPARFA-M algorithm that estimates $\mathbf{W}$ and $\boldsymbol{\mu}$ using regularized maximum likelihood estimation. In practice, a set of graded responses for estimating $\mathbf{W}$ and $\boldsymbol{\mu}$ can be obtained from a previous offering of a course. Finally, we note that the eSPARFA model is clearly not identifiable. We refer to (Lan et al., 2014) for a discussion of how some identifiability problems can be avoided by appropriate regularization of some parameters. Furthermore, it is clear that eSPARFA depends on choosing a suitable number of knowledge components, i.e., the value K. There are several ways in which K can be chosen appropriately. For example, K can be set using cross-validation, using Bayesian methods as in (Fronczyk et al., 2013), or using prior information about the course content. A full analysis of the specifics of the eSPARFA model is not within the scope of this paper, which is why we assume that the parameters $\mathbf{W}$ and $\boldsymbol{\mu}$ are *known* throughout this paper.

2.2 TeSR: Test-size reduction

We now formulate the test-size reduction (TeSR) problem of selecting an appropriate subset of Q given questions that enable us to obtain accurate estimates for the learner dependent parameters $\mathbf{c}^*$ and a^*. Suppose, that we select a subset $\mathcal{I}$ of $|\mathcal{I}| = q < Q$ questions, and we are given the corresponding graded response vector $\mathbf{y}_\mathcal{I}$. Following the model in (1), and by assuming that all random variables $Y_1, \ldots, Y_Q$ are independent given the parameters $\mathbf{W}$, $\boldsymbol{\mu}$, $\mathbf{c}^*$, and a^*, the joint probability distribution of $\mathbf{Y}_\mathcal{I}$ is given by

$$\mathbb{P}(\mathbf{Y}_\mathcal{I} = \mathbf{y}_\mathcal{I} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}^*, a^*) = \prod_{i \in \mathcal{I}} \frac{\exp\left(y_i(\mathbf{w}_i^T \mathbf{c}^* + a^* + \mu_i)\right)}{1 + \exp(\mathbf{w}_i^T \mathbf{c}^* + a^* + \mu_i)}. \tag{2}$$

Here, the vector $\mathbf{y}_\mathcal{I}$ contains the responses of the learner to the questions $\mathcal{I}$. To see if the independence assumption in (2) is reasonable, for any $i \neq i'$, consider the conditional probability $\mathbb{P}(Y_i = 1 \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}^*, a^*, Y_{i'})$. Since the student parameters are known, it is likely that the response of the student to question i' will not influence the response of the student to question i. This intuition validates the independence assumption. The maximum likelihood estimate (MLE) of the knowledge component vector $\mathbf{c}^*$ and the ability parameter a^* can be written as follows:

$$\begin{aligned} \{\widehat{\mathbf{c}}, \widehat{a}\} &= \underset{\mathbf{c} \in \mathbb{R}^K,\, a \in \mathbb{R}}{\arg\max}\ \log \mathbb{P}(\mathbf{Y}_\mathcal{I} = \mathbf{y}_\mathcal{I} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}, a) \\ &= \underset{\mathbf{c} \in \mathbb{R}^K,\, a \in \mathbb{R}}{\arg\max} \sum_{i \in \mathcal{I}} \left[y_i(\mathbf{w}_i^T \mathbf{c} + a + \mu_i) - \log\left(1 + \exp(\mathbf{w}_i^T \mathbf{c} + a + \mu_i)\right) \right]. \end{aligned} \tag{3}$$

Given $\mathbf{y}_\mathcal{I}$, $\mathbf{W}$, and $\boldsymbol{\mu}$, the problem (3) can be solved via standard convex optimization methods (see, e.g., (Boyd and Vandenberghe, 2004)). The main objective in TeSR is to find an appropriate subset $\mathcal{I}$ such that the estimates $\widehat{\mathbf{c}}$ and $\widehat{a}$ are as close as possible to the true unknown parameters $\mathbf{c}^*$ and a^*, respectively. In order to analytically formulate the TeSR problem, we make use of the fundamental asymptotic normality property of MLEs (see, e.g., (Fahrmeir and Kaufmann, 1985) for more details). Before stating the Theorem, we define the Fisher information matrix by

$$\mathbf{F}_\mathcal{I} = \sum_{i \in \mathcal{I}} \frac{\exp(\mathbf{w}_i^T \mathbf{c}^* + a^* + \mu_i)}{\left(1 + \exp(\mathbf{w}^T \mathbf{c}^* + a^* + \mu_i)\right)^2} [\mathbf{w}_i^T, 1]^T [\mathbf{w}_i^T, 1], \tag{4}$$

where $[\mathbf{w}_i^T, 1]^T$ designates a column vector consisting of $\mathbf{w}_i$ and the scalar 1.

Theorem 1 (Asymptotic normality property (Fahrmeir and Kaufmann, 1985))**.** *Suppose the Fisher information matrix* $\mathbf{F}_\mathcal{I}$ *in* (4) *is invertible for all subsets* $\mathcal{I}$ *such that* $q = |\mathcal{I}| \geq K + 1$, *and let* $\mathbf{e} = q^{1/2}\left([\widehat{\mathbf{c}}, \widehat{a}]^T - [\mathbf{c}^*, a^*]^T\right)$ *be the scaled error in estimating the learner-dependent parameters. Then, as* $q \to \infty$, *the scaled error* $\mathbf{e}$ *converges in distribution to a multivariate normal vector with mean zero and covariance* $\mathbf{F}_\mathcal{I}^{-1}$, *i.e., we have* $\mathbf{e} \xrightarrow{d} \mathcal{N}(0, \mathbf{F}_\mathcal{I}^{-1})$, *where* $\xrightarrow{d}$ *designates convergence in distribution.*

Note that $\mathbf{F}_{\mathcal{I}}$ is a $K+1 \times K+1$ matrix, and so we need at least $K+1$ questions for $\mathbf{F}_{\mathcal{I}}$ to be invertible. Theorem 1 states that as the number of questions q grows, the probability distribution of the error vector $\mathbf{e} = q^{1/2}\left([\widehat{\mathbf{c}}\,,\, \widehat{a}]^T - [\mathbf{c}^*, a^*]^T\right)$ converges to a multivariate normal distribution with mean zero and covariance given by the inverse of the Fisher information matrix. The main assumption in Theorem 1 is for the Fisher information matrix $\mathbf{F}_{\mathcal{I}}$ to be invertible for all choices of the set of questions $\mathcal{I}$. Since $\mathbf{F}_{\mathcal{I}}$ depends on $\mathbf{W}$, the invertibility of $\mathbf{F}_{\mathcal{I}}$ implicitly imposes assumptions on the question–knowledge component matrix $\mathbf{W}$.

As mentioned earlier, the main goal in TeSR is to select the subset of questions $\mathcal{I}$ so that the error $\mathbf{e}$ is as small as possible. Since we have an approximation of the distribution of $\mathbf{e}$, one way of selecting $\mathcal{I}$ is to ensure that the uncertainty in the random vector $\mathbf{e}$ is minimal. A natural way of measuring the uncertainty in a random vector is the differential entropy (Cover and Thomas, 2012), which, for a multivariate normal random vector with mean zero and covariance $\mathbf{\Sigma}$, is given by $\log\big((2\pi e)^q \det(\mathbf{\Sigma})\big)$. Consequently, we define the TeSR optimization problem as

$$\text{(TeSR)} \qquad \widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \; \log\det(\mathbf{F}_{\mathcal{I}})\,.$$

There are two main challenges in finding the solution to the TeSR problem:

(i) The objective function, in general, cannot be computed exactly, as it depends on the (typically) unknown learner-dependent parameters $\mathbf{c}^*$ and a^*.

(ii) The optimization problem is combinatorial in nature, as it involves an exhaustive search over all $\binom{Q}{q}$ subsets of questions.

In Section 3, we address the first problem by approximating the objective function in (TeSR) by means of prior data available on the learner-dependent parameters. Subsequently, in Section 4, we address the second problem by approximating the solution to the combinatorial optimization problem using greedy methods.

3 Approximating the TeSR Problem using Prior Data

In this section, we show how the TeSR objective function, which cannot be evaluated exactly (because it depends on unknown parameters), can be approximated using prior data from multiple learners answering questions. Recall that the random variable Y_i denotes the graded response of a learner to the i^{th} question. Using the probability distribution of Y_i in (1), we see that the scalar term in the summation of (4) corresponds to the variance of the random variable Y_i, i.e., the following relation holds:

$$\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*, a^*] = \frac{\exp(\mathbf{w}_i^T\mathbf{c}^* + a^* + \mu_i)}{\big(1+\exp(\mathbf{w}^T\mathbf{c}^* + a^* + \mu_i)\big)^2}\,. \tag{5}$$

The variance $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*, a^*]$ captures the variability of the learner's graded response in answering the i^{th} question. By defining $\mathbf{V}$ as a $Q \times Q$ diagonal matrix with entries $v_{ii} = \mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*, a^*]$ on the main diagonal, the TeSR problem can be rewritten as

$$\widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \; \log\det\Big(\overline{\mathbf{W}}_{\mathcal{I}}^T\,\mathbf{V}_{\mathcal{I}}\,\overline{\mathbf{W}}_{\mathcal{I}}\Big) \quad \text{with} \quad \overline{\mathbf{W}} = [\mathbf{1},\,\mathbf{W}]\,, \tag{6}$$

where $\overline{\mathbf{W}}_\mathcal{I}$ is the $q \times K+1$ matrix corresponding to the rows of $\overline{\mathbf{W}}$ that are indexed by $\mathcal{I}$ and $\mathbf{1}$ is an all-ones column vector of appropriate dimension. Note that the matrix $\mathbf{V}_\mathcal{I}$ is *unknown*, in general, as it depends on the unknown learner-dependent parameters $\mathbf{c}^*$ and a^*. To approximate the objective function in (6), we use an estimate of the variance $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*, a^*]$. To this end, let $\widetilde{\mathbf{Y}}$ be a $Q \times N$ graded response matrix, e.g., obtained from a previous offering of the same course. This matrix can be built from the same data used to estimate the parameters $\mathbf{W}$ and $\boldsymbol{\mu}$ as mentioned in Remark 4. With this response matrix, we now define the empirical variance of each row of $\widetilde{\mathbf{Y}}$ as

$$\widetilde{v}_{ii} = \frac{1}{N}\sum_{j=1}^{N}\left(\widetilde{Y}_{ij} - \frac{1}{N}\sum_{j=1}^{N}\widetilde{Y}_{ij}\right). \tag{7}$$

Let $\widetilde{\mathbf{V}}$ be a diagonal matrix with the diagonal entries given by $\widetilde{v}_{ii}$. Using $\widetilde{\mathbf{V}}$ as a proxy for the true variances contained in $\mathbf{V}$, a solution to (5) can be approximated by

$$\widetilde{\mathcal{I}} = \operatorname*{arg\,max}_{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q} \log\det\left(\overline{\mathbf{W}}_\mathcal{I}^T\,\widetilde{\mathbf{V}}_\mathcal{I}\,\overline{\mathbf{W}}_\mathcal{I}\right). \tag{8}$$

The rationale behind this approximation is that the responses in $\widetilde{\mathbf{Y}}$ are assumed to be from learners with the same parameters $\mathbf{c}^*$ and a^*. In light of no other available information about the learner, the above approximation seems reasonable—we will see numerical simulations in Section 6 showing the efficacy of using the above approximation in practice. In particular, we compare our proposed approximation to another approximation that completely ignores the variance term, i.e., assumes that the diagonal entries $\widetilde{v}_{ii}$ of $\widetilde{\mathbf{V}}_\mathcal{I}$ are the same for all $i = 1, \ldots, Q$. Finally, since (8) is independent of the learner-dependent parameters, it can be used to extract a subset of questions for multiple learners in a class so that all learners receive the same set of questions. In the next section, we propose a greedy algorithm for finding an approximate solution to the combinatorial optimization problem in (8).

4 Non-adaptive Test-size Reduction (NA-TeSR)

In this section, we develop an algorithm for *non-adaptive test-size reduction*, referred to as NA-TeSR. As illustrated in Figure 1(a), we will design an algorithm that selects q questions from a database of Q questions and then, use the selected questions to assess the concept knowledge of learners. We proceed by solving the optimization problem in (8) using methods that resemble those for *sensor selection* (Joshi and Boyd, 2009; Shamaiah et al., 2010) in signal processing, where it is desirable to select a small number of sensors from a large collection of sensors monitoring an environment. Although the statistical model for sensor measurements differs significantly from the eSPARFA model in (1), the problem formulation of TeSR in (8) is similar to the problem formulation of sensor selection. There are two prominent approaches for sensor selection in the literature. The first approach is based on convex optimization (Joshi and Boyd, 2009) and the second on greedy methods (Shamaiah et al., 2010). The greedy method has advantages over the convex optimization approach in terms of computational complexity and has been shown to lead to superior empirical performance (Shamaiah et al., 2010). For this reason, we solely focus on a greedy approach to find an approximate solution to (8).

Algorithm 1 summarizes the steps of the proposed non-adaptive TeSR (NA-TeSR) algorithm.

Algorithm 1: Nonadaptive test-size reduction (NA-TeSR)

Step 1: For each knowledge component j, where $j \in \{1, \ldots, K\}$, select a question i (from the set of all unselected questions) such that $\widetilde{v}_{ii} w_{ij}^2$ is maximum.

Step 2: Select the $(K+1)^{\text{th}}$ question by solving

$$\widetilde{\mathcal{I}}_{K+1} = \arg\max_{i \in [Q] \setminus \widetilde{\mathcal{I}}} \log\det\left(\overline{\mathbf{W}}^T_{\widetilde{\mathcal{I}}_{[K]} \cup i} \widetilde{\mathbf{V}}_{\widetilde{\mathcal{I}}_K \cup i, \widetilde{\mathcal{I}}_K \cup i} \overline{\mathbf{W}}_{\widetilde{\mathcal{I}}_{[K]} \cup i}\right),$$

where $\widetilde{\mathcal{I}}_{[K]}$ denotes the set of first K questions selected in Step 1.

Step 3: For $\ell = K+1, \ldots, q-1$, select the $(\ell+1)^{\text{th}}$ question by solving

$$\widetilde{\mathcal{I}}_{\ell+1} = \arg\max_{i \in [Q] \setminus \widetilde{\mathcal{I}}_{[\ell]}} \widetilde{v}_{ii}\, [\mathbf{w}_i^T,\, 1] \left(\overline{\mathbf{W}}^T_{\widetilde{\mathcal{I}}_{[\ell]}} \widetilde{\mathbf{V}}_{\widetilde{\mathcal{I}}_{[\ell]}, \widetilde{\mathcal{I}}_{[\ell]}} \overline{\mathbf{W}}_{\widetilde{\mathcal{I}}_{[\ell]}}\right)^{-1} [\mathbf{w}_i^T,\, 1]^T,$$

where $\widetilde{\mathcal{I}}_{[\ell]}$ denotes the set of first ℓ questions selected.

For a set $\mathcal{I}$, let $\mathcal{I}_{[\ell]}$ be the first ℓ elements of $\mathcal{I}$ and let $\mathcal{I}_\ell$ be the ℓ^{th} element of $\mathcal{I}$. Note that $\mathbf{F}_{\mathcal{I}}$ is a $K+1 \times K+1$ matrix. Thus, to obtain accurate estimates of $\mathbf{c}^*$ and a^*, we need to select at least $K+1$ questions. We now elaborate on the three steps of Algorithm 1.

1) The first step in NA-TeSR selects a set of K questions $\widetilde{\mathcal{I}}_{[K]}$ that contains one question from every knowledge component. To do so, note that the Fisher information of the parameter $\mathbf{c}_j^*$ is given by $\sum_{i \in \mathcal{I}} \mathbb{V}\text{ar}[Y_i | \mathbf{c}^*, a^*]\, w_{ij}^2$, where w_{ij} is the $(i,j)^{\text{th}}$ entry of $\mathbf{W}$. Thus, to select the most informative question for every knowledge component in a greedy manner, we want to select a question i so that $w_{ij}\, \mathbb{V}\text{ar}[Y_i | \mathbf{c}^*, a^*]$ is maximized. Substituting the approximation $\widetilde{v}_{ii}$ in lieu of the unknown variance $\mathbb{V}\text{ar}[Y_i | \mathbf{c}^*, a^*]$, we obtain the strategy for selecting question i so that $\widetilde{v}_{ii} w_{i,j}^2$ is maximized.

2) The second step in NA-TeSR selects the $(K+1)^{\text{th}}$ question so that the objective

$$\det(\overline{\mathbf{W}}^T_{\mathcal{I}_{[K+1]}} \widetilde{\mathbf{V}}_{\mathcal{I}_{[K+1]}, \mathcal{I}_{[K+1]}} \overline{\mathbf{W}}_{\mathcal{I}_{[K+1]}}),$$

is maximized, where $\mathcal{I}_{[K]} = \widetilde{\mathcal{I}}_{[K]}$. This maximization can easily be achieved by searching over all the remaining questions to select the $(K+1)^{\text{th}}$ question.

3) The third step selects the remaining questions in a greedy manner using a step that is similar to Step 2 except that a simple trick, motivated from (Shamaiah et al., 2010), is used to simplify the computations. In particular, if $\widetilde{\mathcal{I}}_{[\ell]}$ are the ℓ questions selected, where $\ell \geq K+1$, then the $(\ell+1)^{\text{th}}$ question can be selected by maximizing $\det(\overline{\mathbf{W}}^T_{\mathcal{I}_{[\ell+1]}} \widetilde{\mathbf{V}}_{\mathcal{I}_{[\ell+1]}, \mathcal{I}_{[\ell+1]}} \overline{\mathbf{W}}_{\mathcal{I}_{[\ell+1]}})$, where $\mathcal{I}_{[\ell]} = \widetilde{\mathcal{I}}_{[\ell]}$. Using the well-known identity $\det(\mathbf{X} + \mathbf{b}\mathbf{b}^T) = \det(\mathbf{X})(1 + \mathbf{b}^T \mathbf{X}^{-1} \mathbf{b})$, where $\mathbf{X}$ is a square matrix and $\mathbf{b}$ is a column vector, we have that

$$\det\left(\overline{\mathbf{W}}^T_{\mathcal{I}_{[\ell+1]}}\, \widetilde{\mathbf{V}}_{\mathcal{I}_{[\ell+1]}, \mathcal{I}_{[\ell+1]}}\, \overline{\mathbf{W}}_{\mathcal{I}_{[\ell+1]}}\right) = \det(\mathbf{M})\, \underbrace{\left(1 + \widetilde{v}_{\mathcal{I}_{\ell+1}, \mathcal{I}_{\ell+1}} [\mathbf{w}^T_{\mathcal{I}_{\ell+1}}, 1]\, \mathbf{M}^{-1}\, [\mathbf{w}^T_{\mathcal{I}_{\ell+1}}, 1]^T\right)}_{(a)}, \tag{9}$$

where $\mathbf{M} = \overline{\mathbf{W}}^T_{\widetilde{\mathcal{I}}_{[\ell]}} \widetilde{\mathbf{V}}_{\widetilde{\mathcal{I}}_{[\ell]},\widetilde{\mathcal{I}}_{[\ell]}} \overline{\mathbf{W}}_{\widetilde{\mathcal{I}}_{[\ell]}}$. Since $\mathbf{M}$ is known, (9) can be maximized by searching for an appropriate question $\mathcal{I}_{\ell+1}$ so that the expression in (a) is maximized.

In summary, NA-TeSR solves the TeSR problem (8) in a greedy manner by selecting a locally optimal question in each iteration.

Remark 5 (Comparison to the Rasch model)**.** As mentioned in Remark 1, the eSPARFA model reduces to the Rasch model when $K = 0$. In this case, it is easy to see that the TeSR problem reduces to choosing a set $\mathcal{I}$ that maximizes $\sum_{i \in \mathcal{I}} \mathbb{V}\mathrm{ar}[Y_i | a^*]$, where Y_i is the random variable representing the graded response to the i^{th} question. Since the Rasch model ignores the question–knowledge component relationship, the TeSR problem, in this particular case, is no longer computationally challenging, and each question can be selected independently. On the other hand, since eSPARFA models question–knowledge component relationships, as we see in Algorithm 1, the questions can no longer selected independently. Furthermore, we refer to Section 6 for numerical results on synthetic and real data that show the benefits of using NA-TeSR versus Rasch-based methods when the questions test knowledge on multiple knowledge components.

5 Adaptive Test-size Reduction (A-TeSR)

In this section, we develop an algorithm for *adaptive test-size reduction*, referred to as A-TeSR. In 4, we introduced the non-adaptive TeSR algorithm, where all the questions are selected at the same time before learners submit their responses to the questions. However, in many settings, tests are computerized, and the questions can be selected in an adaptive manner. In such cases of adaptive testing, the individual response history of learners can be used to adaptively select the "next best" question (in terms of minimizing each learner's estimation error). Adaptive tests are popularly employed when learners take standardized tests such as the SAT, ACT, or GRE (van der Linden and Glas, 2000).

From the perspective of the TeSR problem formulation, the response history of a learner allows for an alternative approach to approximate the TeSR objective function using the parameters estimated from the response history instead of prior data from other learners. This appealing property of adaptive testing can potentially allow adaptive tests to ask fewer questions to assess concept knowledge of learners. However, in order to implement an adaptive testing algorithm, it is important to be able to estimate the intermediate knowledge component parameters computed after a learner responds to a question. Although these parameters can be estimated using maximum likelihood, the maximum likelihood estimator (MLE) may not exist for certain choices of the questions. Thus, it is important to understand the specifics of where the MLE may not exist and devise an algorithm to avoid such situations. In Section 5.1, we discuss conditions under which the MLE exists. In Section 5.2, we use these conditions to develop a strategy to adaptively select questions.

5.1 Existence of the maximum-likelihood estimator (MLE)

In an adaptive testing scenario, the learner-dependent parameters, $\mathbf{c}^*$ and a^*, are re-estimated after each question is answered. In particular, if $\mathbf{y}_{\mathcal{I}}$ is the graded learner response to the questions indexed by $\mathcal{I}$, then the MLE of $\mathbf{c}^*$ and a^* is given by

$$\{\widehat{\mathbf{c}}, \widehat{a}\} = \underset{\mathbf{c}\in\mathbb{R}^K,\, a\in\mathbb{R}}{\arg\max} \sum_{i\in\mathcal{I}} \left[y_i(\mathbf{w}_i^T\mathbf{c} + a + \mu_i) - \log\bigl(1 + \exp(\mathbf{w}_i^T\mathbf{c} + a + \mu_i)\bigr)\right]. \tag{10}$$

Although the objective function in (10) is convex, it is not strictly convex. For this reason, the MLE may diverge to infinity. In this case, we say that the MLE does not exist. To avoid this situation, it is important to carefully select the next best question. To this end, we make use of the following existence theorem, whose proof is given in Appendix A.

Theorem 2. *Suppose the graded responses $\mathbf{y}_\mathcal{I}$ from questions $\mathcal{I}$ follow a distribution given the eSPARFA model in* (1). *Further, for the knowledge component j, let S_j be the indices for which $w_{ik} > 0$. If $\mathbf{y}_{S_j\cap\mathcal{I}} = 0$ or if $\mathbf{y}_{S_j\cap\mathcal{I}} = 1$, then the MLE of $\mathbf{c}_j^*$ does not exist. Further, if $\mathbf{y}_\mathcal{I} = 0$ or if $\mathbf{y}_\mathcal{I} = 1$, then the MLE of a^* does not exist.*

Informally, Theorem 2 states that if the graded responses to the questions associated with a knowledge component are all incorrect (indicated by 0) or all correct (indicated by 1), then the MLE of the parameters associated with that knowledge component does not exist. In addition, Theorem 2 states that if all responses to the questions are either incorrect or correct, then the MLE of the ability parameter does not exist. As an example, consider the following matrix $\mathbf{W}$ and the graded response vector $\mathbf{y}$:

$$\mathbf{W} = \begin{bmatrix} 0.2 & 0 & 0.3 \\ 0 & 1.4 & 0.7 \\ 0.1 & 0 & 0 \\ 0 & 0.2 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{11}$$

According to the notation in Theorem 2, we have $\mathcal{I} = \{1,2,3,4\}$. Further, $S_1 = \{1,3\}$ is the set of all questions associated with the first knowledge component. To check whether the MLE for $\mathbf{c}_1^*$ exists, we inspect $y_{S_1\cap\mathcal{I}} = [y_1, y_3]$. Since both these entries are zero, according to Theorem 2, the MLE of $\mathbf{c}_1^*$ does not exist. When confronted with such a situation, we will see in Section 5.2, that the proposed A-TeSR algorithm will select a suitable question that tests knowledge from the first knowledge component.

Remark 6 (Necessary conditions for MLE existence)**.** Note that the condition in Theorem 2 is only a sufficient condition for the MLE to exist. In other words, even if the condition in Theorem 2 holds, it is not guaranteed that the MLE will exist. The necessary and sufficient conditions, as shown in (Albert and Anderson, 1984) for generalized linear models, depends on the rows in the matrix $\mathbf{W}$. An open research problem is to design a method that can completely avoid the conditions for which the MLE does not exist using as few questions as possible. As detailed in Section 5.2, in the event that the MLE does not exist, and the condition in Theorem 2 is not satisfied, we use NA-TeSR to select select the next question.

5.2 Adaptive test-size reduction (A-TeSR) algorithm

A-TeSR (see Algorithm 2) summarizes the steps involved in the proposed adaptive test-size reduction algorithm. We start by selecting $K+1$ questions using the first two steps of NA-TeSR (Algorithm 1) and acquire graded responses from a learner. Note that this step of the algorithm is independent of the learner parameters and is also non-adaptive. The selection of the remaining

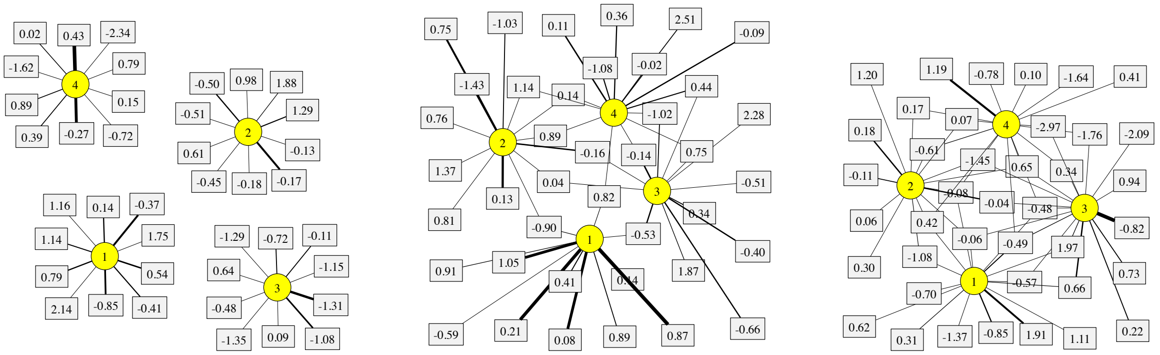
Algorithm 2: Adaptive test-size reduction (A-TeSR)

```
Select $K+1$ questions $\widetilde{\mathcal{I}}_{[K+1]}$ using Steps 1 and 2 of Algorithm 1.
Acquire graded learner responses $\mathbf{y}_{\widetilde{\mathcal{I}}_{[K+1]}}$.
for $j = K+1, \ldots, q-1$ do
    if condition in Theorem 2 is satisfied then
        Let $\mathcal{Q}$ be all questions that map to knowledge components that satisfy the
        condition in Theorem 2.
        Select the $(j+1)^{\text{th}}$ question, $\widetilde{\mathcal{I}}_{j+1}$, using Step 3 of Algorithm 1 such that the
        maximum is over the questions $\mathcal{Q} \backslash \widetilde{\mathcal{I}}_{[j]}$.
    if condition in Theorem 2 is not satisfied then
        Compute the MLEs $\widehat{\mathbf{c}}$ and $\widehat{a}$ using $\mathbf{y}_{\widetilde{\mathcal{I}}_{[j]}}$.
        if $\widehat{\mathbf{c}}$ and $\widehat{a}$ exists then
            Find $\widetilde{\mathcal{I}}_{j+1}$ using Step 3 of Algorithm 1 by replacing $\widehat{\mathbf{V}}_{i,i}$ with $\mathbb{V}\mathrm{ar}[Y_i|\widehat{\mathbf{c}}, \widehat{a}]$,
            which is computed using (5).
        else
            Find $\widetilde{\mathcal{I}}_{j+1}$ using Step 3 of Algorithm 1.
    Acquire graded learner responses $\mathbf{y}_{\widetilde{\mathcal{I}}_{j+1}}$.
```

questions depends on whether the learner parameters can be estimated given graded responses or not. In particular, if the condition in Theorem 2 is satisfied for a knowledge component, then the MLE of that knowledge component parameter diverges to infinity. In this case, we find all questions, say the set $\mathcal{Q}$, that are associated with a knowledge component that satisfies the condition in Theorem 2. Next, we select a question from the predefined set $\mathcal{Q}$ that maximizes the objective in Step 3 of the NA-TeSR algorithm.

If the condition in Theorem 2 is not satisfied for any knowledge component, then the MLE of the learner parameters may exist. This existence can be checked in practice using methods in (Konis, 2007). If the MLE exists, we no longer need to use the approximation of the TeSR problem in (8). Instead, we substitute $\mathbf{c}^*$ with $\widehat{\mathbf{c}}$ and a^* with $\widehat{a}$ to find a new approximation of the TeSR objective function. If the MLE does not exist, we simply perform Step 3 of NA-TeSR to select the next question.

Remark 7. Just as in the case of the non-adaptive algorithm, when $K = 0$, Algorithm 2 reduces to an adaptive Rasch model-based method; see (Chang and Ying, 2009) for examples of such algorithms. As highlighted before, the eSPARFA model used to formulate TeSR takes into account the dependencies among questions, while the Rasch model does not account for such dependencies. Note that Rasch model-based adaptive testing has the natural interpretation that every question selected is such that it's difficulty matches the learner's ability (or estimated ability). This is because such a question choice maximizes the variance of the learner response, i.e., the variance of the random variable Y_i, defined in (1), conditioned on the question and student parameters. In contrast, the A-TeSR based method, that depends on the eSPARFA model with $K > 0$, does not have such an interpretation. Regardless, our numerical simulations clearly show the benefits of using the eSPARFA model for adaptive testing in situations where questions test knowledge on multiple concepts.



(a) $\alpha = 0$

(b) $\alpha = 0.1$

(c) $\alpha = 0.2$

Figure 2: Examples of synthetic $\mathbf{W}$ matrices generated for different values of the sparsity parameter α. See Section 6.2 for a description of α and how $\mathbf{W}$ is generated. In the graphs shown, the rectangles are the question labels with their intrinsic difficulty. A link between a question and concept means that the questions tests knowledge on that concept. In general, a lower (higher) α corresponds to a sparser (denser) $\mathbf{W}$ matrix.

6 Experimental Results

In this section, we assess the performance of NA-TeSR and A-TeSR for test-size reduction on synthetic and real educational data. Section 6.1 describes the simulation setup. Sections 6.2 and 6.3 discuss synthetic results for NA-TeSR and A-TeSR, respectively. Section 6.4 provides results for both A-TeSR and NA-TeSR with real educational data.

6.1 Simulation setup

Generating synthetic W: To generate a matrix $\mathbf{W}$, we assume that most of the questions test knowledge from only one knowledge component and only some questions test knowledge from multiple knowledge components. With this structure of the questions in mind, we generate $\mathbf{W}$ as follows:

- Partition the questions into Q/K groups and map each group to a different knowledge component. The strength of the mapping, w_{ij}, is sampled independently from an exponential distribution with parameter $\lambda = 1$.
- Note that in the matrix $\mathbf{W}$ generated so far, $Q(K-1)$ entries have not yet been assigned. We randomly choose a fraction α of these entries and assign them to non-zero values sampled from an exponential distribution with parameter $\lambda = 1$. In what follows, we refer to α as the *sparsity parameter*.
- In the generation of $\mathbf{W}$ so far, suppose the first question maps to two knowledge components, say 1 and 2, and the second question maps only to 1. If $w_{1,1} = w_{2,1}$, then for a learner with knowledge component parameter $\mathbf{c}^*$, we have $\mathbf{w}_1^T\mathbf{c}^* > \mathbf{w}_2^T\mathbf{c}^*$. This means that, just because the first question tests knowledge from more than one concept, a learner is more likely to get that question correct. To avoid such a situation, the final step in gen-

erating $\mathbf{W}$ is a normalization so that each row in $\mathbf{W}$ is divided by the number of non-zero entries in that row.

Three examples of matrices $\mathbf{W}$ generated using the above approach are visualized in Figure 2 for different values of the sparsity parameter α.

Performance measures: We use three measures to assess the performance of TeSR:

1. The root mean-square error (RMSE) for the knowledge component estimate, defined as $\mathsf{RMSE}_c = \|\hat{\mathbf{c}} - \mathbf{c}^*\|_2/\sqrt{K}$, where $\hat{\mathbf{c}}$ is the estimate delivered by each method and $\mathbf{c}^*$ is the (known) ground truth.

2. The RMSE for the ability parameter, $\mathsf{RMSE}_a = |\hat{a} - a^*|$, where $\hat{a}$ is the estimate delivered by each method and a^* is the (known) ground truth.

3. The negative log-likelihood (NLL) over a hold-out set of questions $\mathcal{H}$

$$\mathsf{NLL} = -\sum_{i \in \mathcal{H}} \left(y_i(\mathbf{w}_i^T \hat{\mathbf{c}} + \hat{a} + \mu_i) - \log(1 + \exp(\mathbf{w}_i^T \hat{\mathbf{c}} + \hat{a} + \mu_i)) \right),$$

   where $\hat{\mathbf{c}}$ and $\hat{a}$ are the estimates generated by the TeSR method under consideration. The set $\mathcal{H}$ is randomly chosen in each trial from the set of all questions Q.

For synthetic data, we only use the RMSE, since the ground truth parameters, $\mathbf{c}^*$ and a^*, are known. For real data, we use both the RMSE and the NLL. Since the ground truth for real data is, in general, unknown, we approximate the RMSE based measures by assuming that the ground truth corresponds to the parameters estimated from all Q available questions. Note that the NLL does not require knowledge about the ground truth. Evidently, we want all the performance measures to be as small as possible.

Methodology: In all the experiments, we assume that $\mathbf{W}$ and $\boldsymbol{\mu}$ are known. This is specified for the synthetic data and estimated using all Q questions for the real data using a properly modified version of the SPARFA-M algorithm from (Lan et al., 2014) that takes into account the ability parameter in the eSPARFA model. For synthetic experiments, the parameters of a learner, namely the knowledge component and ability parameter, are sampled from a uniform distribution on $[-1, 1]$. Further, as shown in Section 3, the TeSR algorithms use prior learner response data $\widetilde{\mathbf{Y}}$ to approximate the TeSR objective function. In all simulations, we obtain a matrix of student response data $\mathbf{Y}$ of size $Q \times (N + 1)$ from $(N + 1)$ learners answering Q questions. We arbitrary subsample the response $\widetilde{\mathbf{Y}}$ matrix of size $Q \times N$ from $\mathbf{Y}$ and then apply the TeSR methods to design a test for the left out learner. In all simulations, we let $N = 50$, and we report the mean and standard deviation of the performance measures computed over 1000 trails.

MLE convergence: As mentioned in Section 3, the MLE may not exist for certain patterns of the response vectors. In such cases, $\hat{a}$ and the entries in $\hat{\mathbf{c}}$ are either set to $+\infty$ or $-\infty$. To deal with such situations and with situations where the entries are too large (or small), we truncate the learner parameters as follows:

$$\hat{a} = \left\{ \begin{array}{ll} \min\{\hat{a}, a^+\} & \hat{a} \geq 0 \\ \max\{\hat{a}, a^-\} & \hat{a} < 0 \end{array} \right. \quad \text{and} \quad \hat{\mathbf{c}}_i = \left\{ \begin{array}{ll} \min\{\hat{\mathbf{c}}_i, \mathbf{c}_i^+\} & \hat{\mathbf{c}}_i \geq 0 \\ \max\{\hat{\mathbf{c}}_i, \mathbf{c}_i^-\} & \hat{\mathbf{c}}_i < 0 \end{array} \right. , \; i = 1, \ldots, K,$$

where $a^+, a^-, \mathbf{c}^+$, and $\mathbf{c}^-$ are computed using the prior response data $\widetilde{\mathbf{Y}}$. For example, a^+ (a^-) is the maximum (minimum) ability parameter among the N learners in the training data. Furthermore, we assume that $a^+ \geq 0$ and $a^- < 0$. The entries in the vectors $\mathbf{c}^+$ and $\mathbf{c}^-$ are defined in a similar manner. The intuition behind the above truncation is that if a parameter is estimated to be too large (or small), then it is reasonable to estimate that parameter to the best (worst) value obtained among a group of learners who have previously answered questions on the same topic.

6.2 NA-TESR EXPERIMENTS

We now show empirical results on synthetic data comparing the NA-TeSR method detailed in Algorithm 1 to three other TeSR methods:

- *EV*: Recall from Section 3 that NA-TeSR uses prior data to approximate the TeSR objective function by using an estimate of the variance $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{w}_i, \mathbf{c}^*, a^*]$. The EV method, short for *equal variance*, assumes that the variance of each question is the same, i.e, $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{w}_i, \mathbf{c}^*, a^*] = \mathbb{V}\mathrm{ar}[Y_{i'}|\mathbf{w}_{i'}, \mathbf{c}^*, a^*] = v$ with $v > 0$ for all $i, i' = 1, \ldots, Q$. EV is implemented by simply using using the approximation $\widetilde{v}_{ii} = v$ in Algorithm 1.

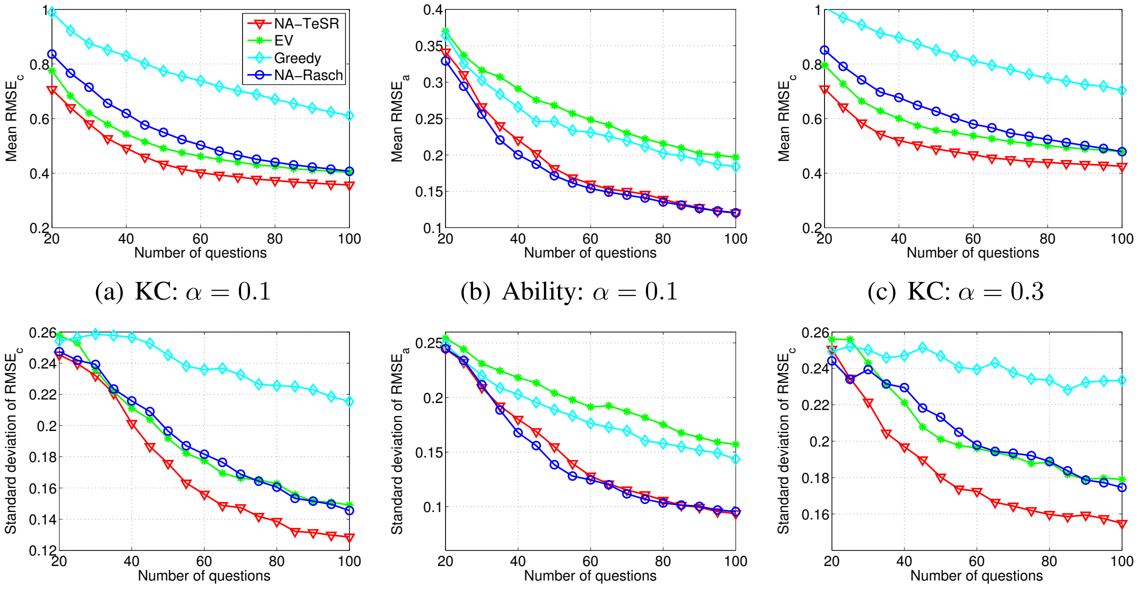
- *NA-Rasch*: The NA-Rasch method ignores the question–knowledge component matrix $\mathbf{W}$ in Algorithm 1 so that the selected questions $\widetilde{\mathcal{I}}$ maximize $\sum_{i \in \widetilde{\mathcal{I}}} \widetilde{v}_{ii}$, where $\widetilde{v}_{ii}$ is the approximation (7). This is equivalent to assuming that the data is being generated from a Rasch model with the difficulty of questions set to $\boldsymbol{\mu}$, or equivalently, corresponding to the eSPARFA model with $K = 0$.

- *Greedy*: The Greedy method iteratively selects a question at random from each concept until the required number of questions has been selected. If all questions from a given concept are exhausted, then Greedy ignores the questions from that knowledge component in subsequent iterations. This method may be considered a straightforward approach adopted by course instructors or domain experts when designing questions.

To generate synthetic data, in each trial of the experiments, we sample the learner parameters and $\mathbf{W}$ as described in Section 6.2 with $Q = 400$. The intrinsic difficulty of each question, μ_i, is sampled from a Gaussian distribution with mean 0 and variance σ^2. Figure 4 plots the mean and standard deviation of the RMSE based performance measures as the number of *selected questions* varies from 20 to 100 and $\sigma^2 = 25.0$. Figure 4 plots the mean and standard deviation of the RMSE based performance measures as the *variance* σ^2 varies from 0.5 to 36.0 and $q = 40$. Some remarks regarding the results are as follows:

- The Greedy method performs worse than NA-TeSR in estimating both $\mathbf{c}^*$ and a^*. This shows that the TeSR problem formulation of minimizing the uncertainly in the estimation of the learner parameters is appropriate for designing small and accurate tests.

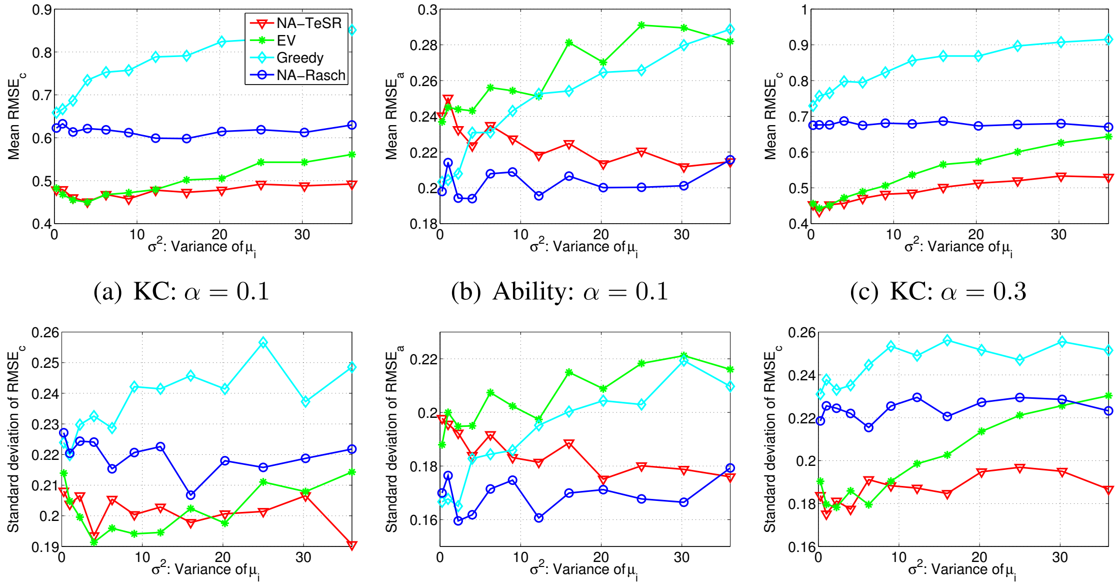
- Although the NA-Rasch method, in general, leads to good estimates of a^* (when compared to other methods), it's estimates of $\mathbf{c}^*$ are, in general, worse than both EV and NA-TeSR. This demonstrates that when questions test knowledge from multiple concepts, the relationship between the questions and the concepts should not be ignored when designing tests.



(a) KC: $\alpha = 0.1$ (b) Ability: $\alpha = 0.1$ (c) KC: $\alpha = 0.3$

(d) KC: $\alpha = 0.1$ (e) Ability: $\alpha = 0.1$ (f) KC: $\alpha = 0.3$

Figure 3: Mean and standard deviation of the RMSE-based performance measures as the number of questions selected varies from 20 to 100; “KC” refers to the knowledge component parameters, and “Ability” refers to the ability parameter.



(a) KC: $\alpha = 0.1$ (b) Ability: $\alpha = 0.1$ (c) KC: $\alpha = 0.3$

(d) KC: $\alpha = 0.1$ (e) Ability: $\alpha = 0.1$ (f) KC: $\alpha = 0.3$

Figure 4: Mean and standard deviation of the RMSE-based performance measures as the variance of the intrinsic difficulty varies from 0.25 to 36.0;

- The EV method is the most competitive to our proposed NA-TeSR method for estimating $\mathbf{c}^*$. In particular, we see in Figure 4 that for small σ^2 (the variance of μ_i), both EV and NA-TeSR yield comparable estimates of $\mathbf{c}^*$. However, as σ^2 increases, NA-TeSR performs substantially better than EV in terms of estimating the knowledge component parameter. The reason for this is that, when σ^2 is small, the difficulty of the questions are likely to be similar. In this case, the variance in answering a question correctly, $\mathbb{V}\text{ar}[Y_i|\mathbf{w}_i, \mathbf{c}^*, a^*]$, is more likely to be similar to the variance of other questions. This validates the EV approximation when σ^2 is small, where all the variances are assumed to be the same.

- The EV method performs the worst when it comes to estimating a^*. Thus, although EV is suitable for designing tests for estimating the knowledge component parameters when σ^2 is small, EV is not appropriate for estimating the ability parameters.

- Figure 4 highlights an additional benefit of NA-TeSR: it enables us to rank questions in order of their importance in assessing mastery of learner parameters. To this end, one can use NA-TeSR with $q = Q$ to obtain an ordering of questions with the property that the error in estimating the learner parameters decreases as the learner answers the questions in a sequential manner from the ordered list of questions. Such an ordering can be useful in identifying questions, mainly towards the end of the order, that are only marginally important in assessing the concept knowledge and ability of learners. Such questions can either be omitted or revised by an instructor or domain expert.

6.3 A-TeSR EXPERIMENTS

We now show the benefits of the adaptive TeSR method, A-TeSR, for designing tests. In addition to comparing A-TeSR to NA-TeSR, we also compare the following two methods:

- *A-Rasch*: The A-Rasch method uses the Rasch model to select questions in an *adaptive* manner based on the prior responses from a learner. We implement A-Rasch using Algorithm 2 with $K = 0$.

- *Oracle*: The Oracle method uses the true underlying (but in practice unknown) knowledge component vector $\mathbf{c}^*$ and ability parameter a^* to compute the TeSR objective, and uses Algorithm 2 to select questions. Note that this algorithm is not practical and is only used to characterize the performance limits of A-TeSR.

Figure 5 compares the RMSE based performance measures as the number of questions vary from 20 to 100 with $\sigma^2 = 4.0$ (the variance of the difficulty μ_i). The rest of the simulation setup is the same as in Section 6.2. To avoid clutter in the presentation of the results, we do not present results for the ability parameter when $\alpha = 0.3$ but note that the results are similar to that of $\alpha = 0.1$, where recall that α is the sparsity parameter that controls the number of non-zero entries in the question–knowledge component matrix $\mathbf{W}$. Furthermore, we do not present results comparing A-TeSR to the other non-adaptive methods outlined in Section 6.2, since NA-TeSR was shown to superior to all other non-adaptive methods. Some remarks regarding the results are as follows:

- A-TeSR outperforms all other methods for estimating the knowledge component parameters. Somewhat surprisingly, A-TeSR's performance in estimating the knowledge component parameter is similar to, and in some cases slightly better, than the Oracle method. The reason for this is that the Oracle method is designed to jointly minimize the error of both the knowledge component and the ability parameter, i.e., the error $\|\widehat{\mathbf{c}} - \mathbf{c}^*\|_2^2 + |\widehat{a} - a^*|^2$, while Figure 5(a) only shows the error from the knowledge component parameter.

- Although the Rasch model based method, A-Rasch, leads to superior results for estimating the ability parameter, it is significantly worse than the A-TeSR method when estimating the knowledge component parameter. Just as in the non-adaptive setting, this shows that using the Rasch model for adaptive testing is not suitable when the questions test knowledge on multiple concepts.

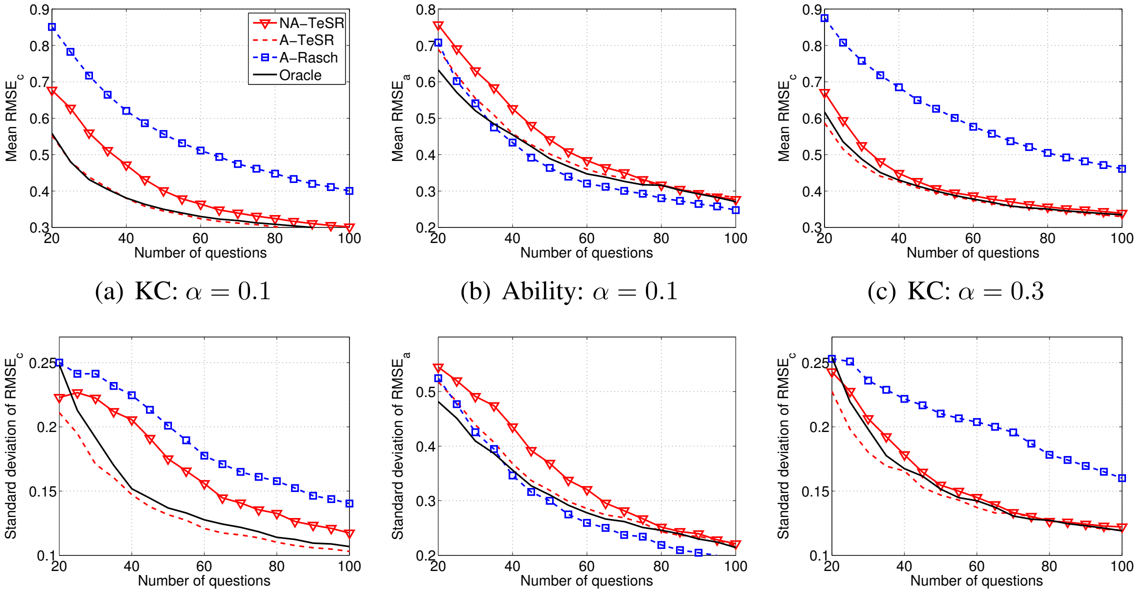
- Comparing Figure 5(a) and Figure 5(c), we see that the performance of NA-TeSR is closer to the performance of A-TeSR when α is larger; recall that larger α corresponds to a matrix $\mathbf{W}$ that is more dense. This behavior suggests that the advantage of using adaptive testing is more significant when there are a smaller number of interactions among the knowledge components. This advantage of A-TeSR can be attributed to the fact that, when $\mathbf{W}$ is sparse, the MLE is less likely to exist for a small number of questions. Thus, since A-TeSR adaptively selects questions using Theorem 2 so that the MLE can be computed in as few number of questions as possible, its performance is superior to NA-TeSR for smaller number of questions. As the sparsity of $\mathbf{W}$ increases, the MLE is more likely to exist for smaller number of questions, and so the performance of NA-TeSR becomes closer to the performance of A-TeSR.

6.4 Real educational data

To assess the performance of the proposed TeSR algorithms under realistic conditions, we carried out experiments using two real educational datasets. The datasets were obtained from exams conducted by a university for admission into their undergraduate program; the learners in these datasets are high-school students. We analyze the data from the exam conducted in 2011 and 2012. Both tests consist of $Q = 60$ questions testing knowledge on physics, chemistry, mathematics, and biology. The exams were graded by negative marking, where a correct response lead to $+3$ points, no response lead to 0 points, and an incorrect response lead to -1 points. For this reason, some learners did not respond to all questions intentionally. In order to fit the data to the eSPARFA model, we treat unanswered questions as missing responses. Note that a more accurate statistical model for this dataset should also model the probability of a learner not answering a question; the development of such a model is part of ongoing research.

For the 2011 data, there are 1714 learners, and for the 2012 data, there are 1567 learners. For both datasets, we use all the data to obtain estimates of $\mathbf{W}$ and $\boldsymbol{\mu}$. In this case, we make use of the tags in each question (i.e., physics, chemistry, mathematics, and biology) to further improve the performance of the SPARFA-M algorithm as described in (Lan et al., 2013). Not surprisingly, the estimated $\mathbf{W}$ matrix maps each question to a single knowledge component.

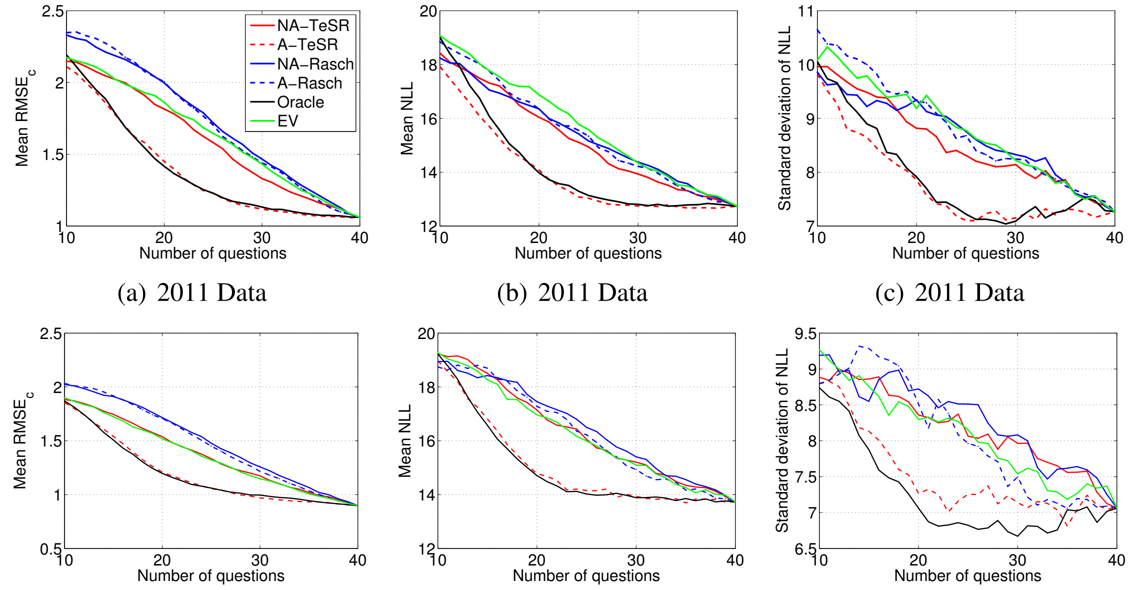
In each trial of the simulation, we randomly select $N = 50$ learners to obtain the prior learning data $\widetilde{\mathbf{Y}}$ (used to approximate the TeSR objective), arbitrary select another learner to test the TeSR methods, and select 20 questions to compute the negative log-likelihood (NLL). Figure 6 plots the mean of the knowledge component RMSE and the mean and the standard deviation of the NLL over 1000 trials for different TeSR methods. To improve readability of the plots, we do not show the results from the Greedy method (the performance was similar to or



(a) KC: $\alpha = 0.1$ (b) Ability: $\alpha = 0.1$ (c) KC: $\alpha = 0.3$

(d) KC: $\alpha = 0.1$ (e) Ability: $\alpha = 0.1$ (f) KC: $\alpha = 0.3$

Figure 5: Performance of the TeSR methods as the number of selected questions vary from 20 to 100. (a)–(c) Mean values of the performance measures over 1000 trials. (d)–(f) Standard deviation of the performance measures over 1000 trials.



(a) 2011 Data (b) 2011 Data (c) 2011 Data

(d) 2012 Data (e) 2012 Data (f) 2012 Data

Figure 6: Results with real data from a University admission test.

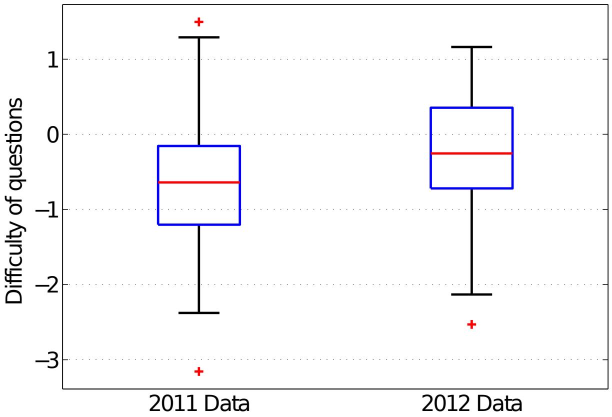


Figure 7: Box-Whisker plot of the difficulty parameters in each real education data set. We clearly see that the 2011 data difficulty parameters have a wider range than the 2012 data difficulty parameters.

worse than Rasch model-based methods). Further, as mentioned earlier, the ground truth for the RMSE and the oracle method is computed using all the available $Q = 60$ questions.

The conclusions drawn from Figure 6 are the same as the conclusions drawn from the synthetic data results. In particular, (i) A-TeSR performs significantly better than all other algorithms, (ii) NA-TeSR performs better than the Rasch-based algorithms, and (iii) A-TeSR performs as good as, and sometimes better than the Oracle method. For example, in the 2011 data, A-TeSR achieves a mean NLL of 14 using approximately 20 questions, whereas all other methods require more than 30 questions to achieve the same NLL.

One particularly interesting aspect of the results is the difference between the performance of NA-TeSR and EV in both data sets. In particular, although NA-TeSR outperforms EV for the 2011 data, the performance of NA-TeSR and EV are nearly the same for the 2012 data. The reason for this can mainly be addressed to the difference in the difficulty of the questions in the 2011 and 2012 data. To illustrate this difference, Figure 7 shows a box plot of the intrinsic difficulty parameters for both data sets. We clearly see that the difficulty parameters for the 2011 data have a wider range than the difficulty parameters for the 2012 data. This implies that the variance of a learner's responses to questions is more likely to be similar for the 2012 data than for the 2011 data. We saw in Section 6.2, this is the main reason why the performance of EV is nearly the same as that of NA-TeSR.

7 Summary

We have proposed two novel methods for test-size reduction (TeSR) that aim to design efficient (small) and accurate tests. Given a question bank containing a large number of questions, TeSR selects a small number of questions such that the selected questions can accurately assess learners. One natural application of TeSR is in designing tests for assessing the knowledge understanding of learners in a course. Yet another application of TeSR is in designing psychological tests. Our methods for solving the TeSR problem used an extended version of the SPARse Factor Analysis (SPARFA) framework proposed in (Lan et al., 2014) to model the relationship between questions and concepts in a course. Subsequently, using theory of maximum likelihood estimators for logistic regression, we formulated the TeSR problem as that of minimizing the uncertainty in the asymptotic error of estimating the concept understanding.

Our first proposed method for TeSR, referred to as non-adaptive TeSR (NA-TeSR), used a data-driven approach to select questions to approximately solve a combinatorial optimization problem in a greedy manner. This approach is suitable in settings where an instructor only has access to the learners responses once all questions have been solved. Our second proposed method, referred to as adaptive TeSR (A-TeSR), is an adaptive algorithm that iteratively suggests questions for each learner individually based on graded responses of learners to prior questions. Our extensive experimental results showed that NA-TeSR and A-TeSR significantly outperform state-of-the-art methods that use the well-established Rasch model (Rasch, 1960). Experimental results on real educational datasets have shown that TeSR can reduce the number of questions needed in a test/assessment by 40%, which significantly reduces learners' workload while still being able to obtain accurate estimates of each learner's concept knowledge.

Our criterion for selecting questions, in both the non-adaptive and the adaptive methods, is based on the Fisher information. Alternative criteria based on Bayesian methods (van der Linden, 1998) and the Kullback-Liebler divergence (Wang et al., 2011) have been proposed in the literature. The framework set forth in this paper can be easily adapted to other methods and comparison of all such methods will be an interesting direction for future research.

While formulating the TeSR problem and developing the proposed algorithms, we have primarily focused on the case where the learner responses are binary, i.e, either correct (1) or incorrect (0). In practice, however, responses can be on an ordinal scale. For example, in educational settings, even if a response is incorrect, a learner may obtain partial credit for showing some understanding of the concepts. The SPARFA model has been extended to handle ordinal data in (Lan et al., 2013). Similar methods can be used for the proposed eSPARFA framework. Subsequently, the TeSR problem can be formulated with respect to the Fisher information of the ordinal model.

Finally, we presented TeSR in the context of selecting q questions out of a database of Q questions. However, by choosing $q = Q$ in the TeSR methods, we can easily output a ranked list of questions such that if question i is ranked higher than question j, then question i is deemed more important/suitable for assessing the knowledge understanding of learners. Such a list can help instructors visualize a ranking of all questions and then select a suitable subsets of questions or revise questions that have been ranked low. We note that a ranking of questions can also be useful when applying TeSR to psychological tests, wherein, a question ranked higher corresponds a question being more suitable for understanding certain aspects of human behavior.

ACKNOWLEDGEMENTS

We are extremely grateful to the Chairman, JAC, IISER Pune, for sharing the educational data. Thanks to Andrew Waters for inspiring discussions about the SPARFA framework and test-size reduction. This work was supported by the National Science Foundation under Cyberlearning grant IIS-1124535, the Air Force Office of Scientific Research under grant FA9550-09-1-0432, the Google Faculty Research Award program, the Swiss National Science Foundation under grant PA00P2-134155, and an Institute for Mathematics and its Applications (IMA) postdoctoral fellowship.

A Proof of Theorem 2

The objective function of the maximum likelihood estimator from (3) can be written as

$$\begin{aligned} L &= \sum_{i \in \mathcal{I}} \big(y_i(\mathbf{w}_i^T \mathbf{c} + \mu_i) - \log\big(1 + \exp(\mathbf{w}_i^T \mathbf{c} + \mu_i)\big)\big) \\ &= \sum_{i \in \mathcal{I}} \left(y_i \sum_{j=1}^{K} w_{ij} c_j + \mu_i y_i - \log(1 + \exp(\mathbf{w}_i^T \mathbf{c} + \mu_i)) \right) . \end{aligned}$$

Consider estimating the j^{th} concept knowledge c_j^*. We can rewrite the likelihood as

$$L = c_j \sum_{i \in \mathcal{I} \cap S_j} y_i w_{ij} + A_{\backslash j} - \sum_{i \in \mathcal{I}} \log(1 + \exp(\mathbf{w}_i^T \mathbf{c} + \mu_i)) \,, \tag{12}$$

where $A_{\backslash j}$ captures the terms that depend on the concepts $c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_K$. When $y_i = 0$ for all $i \in S_j \cap \mathcal{I}$, $L = A_{\backslash j} - \sum_{i \in \mathcal{I}} \log(1 + \exp(w_{ij} c_k) B_{\backslash j})$, where $B_{\backslash j}$ does not depend on c_j. It is clear that $L < A_{\backslash j}$ for all c_j. Moreover, $L \to A_{\backslash k}$ (the upper bound of L) as $c_k \to -\infty$ since the entries in $\mathbf{W}$ are all non-negative. Thus, the ML estimate of c_j^* does not exist. When $y_i = 1$ for all $i \in S_j \cap \mathcal{I}$, then

$$L = c_j \sum_{i \in \mathcal{I} \cap S_j} w_{ij} + A_{\backslash j} - \sum_{i \in \mathcal{I}} \log(1 + \exp(w_{ij} c_k) B_{\backslash j}) < A_{\backslash j} - \widetilde{A}_{\backslash j} \,,$$

where $\widetilde{A}_j$ depends on $B_{\backslash j}$. Further, $L \to A_{\backslash j} - \widetilde{A}_{\backslash j}$ (the upper bound of L) as $c_j \to \infty$. Thus, again the ML estimate does not exist. A similar argument follows for the claim regarding the ability parameter.

References

T. A. Ackerman. 1994. Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education* 7, 4 (1994), 255–278.

A. Albert and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71, 1 (Apr. 1984), 1–10.

A. Anastasi and S. Urbina. 1997. *Psychological testing*. Prentice Hall New Jersey.

J. R. Anderson, C. F. Boyle, and B. J. Reiser. 1982. Intelligent Tutoring Systems. *Science* 228, 4698 (1982), 456–462.

F. B. Baker and S. H. Kim. 2004. *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). Marcel Dekkker Inc.

R. S.J.D. Baker, A. T. Corbett, and V. Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*, Vol. 5091. Springer, 406–415.

T. Barnes. 2005. The Q-matrix Method: Mining Student Response Data for Knowledge. In *Proc. American Association for Artificial Intelligence Workshop on Educational Data Mining*.

D. Benson. 2008. Actively Modifying The Classroom Approach Using Pre-Tests And Recurring Problems. In *American Society for Engineering Education Annual Conf. and Exposition*.

Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. Pritchard. 2012. Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory. In *Proc. of the 5th Intl. Conf. on Educational Data Mining*. 95–102.

S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

P. Brusilovsky and C. Peylo. 2003. Adaptive and Intelligent Web-based Educational Systems. *Intl. Journal of Artificial Intelligence in Education* 13, 2-4 (Apr. 2003), 159–172.

S. Buyske. 2005. Optimal design in educational testing. *Applied Optimal Designs* (2005), 1–19.

H. Cen, K. R. Koedinger, and B. Junker. 2006. Learning Factors Analysis–A General Method for Cognitive Model Evaluation and Improvement. In *Intelligent Tutoring Systems*. Lecture Notes in Computer Science, Vol. 4053. Springer, 164–175.

H. Chang and Z. Ying. 1996. A global information approach to computerized adaptive testing. *Applied Psychological Measurement* 20, 3 (Sep. 1996), 213–229.

H. Chang and Z. Ying. 2009. Nonlinear Sequential Designs for Logistic Item Response Theory Models with Applications to Computerized Adaptive Tests. *The Annals of Statistics* 37, 3 (Jun. 2009), 1466–1488.

M. Chi, K. Koedinger, G. Gordon, and P. Jordan. 2011. Instructional factors analysis: A cognitive model for multiple instructional interventions. In *Proc. 4th Intl. Conf. on EDM*. 61–70.

A. T. Corbett and J. R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (Dec. 1994), 253–278.

T. M. Cover and J. A. Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons.

M. Desmarais. 2011. Conditions for Effectively Deriving a Q-Matrix from Data with Non-negative Matrix Factorization. In *Proc. 4th Intl. Conf. on Educational Data Mining*. 41–50.

L. Fahrmeir and H. Kaufmann. 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* 13, 1 (Mar. 1985), 342–368.

M. Feng and N. T. Heffernan. 2006. Informing teachers live about student learning: Reporting in the ASSISTment system. *Technology Instruction Cognition and Learning* 3, 1/2 (2006), 63.

K. Fronczyk, A. E. Waters, M. Vannucci, M. Guindani, and R. G. Baraniuk. 2013. A Bayesian infinite factor model for learning and content analytics. *submitted to Computational Statistics and Data Analysis* (2013).

U. Graßhoff, H. Holling, and R. Schwabe. 2012. Optimal Designs for the Rasch Model. *Psychometrika* 77, 4 (Oct. 2012), 710–723.

H. H. Harman. 1976. *Modern Factor Analysis*. The University of Chicago Press.

J. Hartley and I. K. Davies. 1976. Preinstructional Strategies: The Role of Pretests, Behavioral Objectives, Overviews and Advance Organizers. *Review of Educational Research* 46, 2 (Jan 1976), 239–265.

G. Hooker, M. Finkelman, and A. Schwartzman. 2009. Paradoxical results in multidimensional item response theory. *Psychometrika* 74, 3 (Sep. 2009), 419–442.

P. Jordan and M. Spiess. 2012. Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika* 77, 1 (Jan. 2012), 127–152.

S. Joshi and S. Boyd. 2009. Sensor selection via convex optimization. *IEEE Trans. on Signal Processing* 57, 2 (Feb. 2009), 451–462.

J. Knox, S. Bayne, H. MacLeod, J. Ross, and C. Sinclair. 2012. MOOC pedagogy: the challenges of developing for Coursera. *Online Newsletter of the Association for Learning Technologies* (Aug. 2012). Issue 28.

K. R. Koedinger, E. A. McLaughlin, and J. C. Stamper. 2012. Automated Student Model Improvement. In *Proc. 5th Intl. Conf. on EDM*. 17–24.

K. Konis. 2007. *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models*. Ph.D. Dissertation. Oxford University.

I. G.G. Kreft and J. de Leeuw. 1998. *Introducing Multilevel Modeling*. Sage.

A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. 2013. Tag Aware Ordinal Sparse Factor Analysis for Learning and Content Analytics. In *Proc. 6th Intl. Conf. on Educational Data Mining*.

A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. 2014. Sparse Factor Analysis for Learning and Content Analytics. *to appear in Journal of Machine Learning Research* (2014).

E. Loken, F. Radlinski, V. H. Crespi, J. Millet, and L. Cushing. 2004. Online study behavior of $100,000$ students preparing for the SAT, ACT, and GRE. *Journal of Educational Computing Research* 30, 3 (May 2004), 255–262.

F. M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates.

R. M. Luecht. 1996. Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement* 20, 4 (Dec. 1996), 389–404.

F. G. Martin. 2012. Will massive open online courses change how we teach? *Commun. ACM* 55, 8 (Aug. 2012), 26–28.

Z. A. Pardos and N. T. Heffernan. 2011. KT-IDEM: introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*. Vol. 6787. Springer, 243–254.

G. Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks paedagogiske Institut.

M. D. Reckase. 2009. *Multidimensional Item Response Theory*. Springer.

S. Ritter, J. Anderson, M. Cytrynowicz, and O. Medvedeva. 1998. Authoring content in the PAT algebra tutor. *Journal of Interactive Media in Education* 98, 9 (Oct. 1998).

D. O. Segall. 1996. Multidimensional adaptive testing. *Psychometrika* 61, 2 (June 1996), 331–354.

M. Shamaiah, S. Banerjee, and H. Vikalo. 2010. Greedy sensor selection: Leveraging submodularity. In *Proc. of 49th IEEE Conf. on Decision and Control*. 2572–2577.

J. C. Stamper, T. Barnes, and M. Croy. 2007. Extracting Student Models for Intelligent Tutoring Systems. In *Proc. of the National Conf. on Artificial Intelligence*, Vol. 22. 113–147.

J. L. Templin and R. A. Henson. 2006. Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods* 11, 3 (Sep. 2006), 287.

N. Thai-Nghe, L. Drumond, T. Horvath, and L. Schmidt-Thieme. 2011. Multi-relational Factorization Models for Predicting Student Performance. *KDD 2011 Workshop on Knowledge Discovery in Educational Data* (Aug. 2011).

W. J. van der Linden. 1998. Bayesian item selection criteria for adaptive testing. *Psychometrika* 63, 2 (June 1998), 201–216.

W. J. van der Linden and C. A. W. Glas. 2000. *Computerized Adaptive Testing: Theory and Practice*. Springer.

W. J. van der Linden and P. J. Pashley. 2010. Item Selection and Ability Estimation in Adaptive Testing. In *Elements of Adaptive Testing*. Springer New York, 3–30.

K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. 2005. The Andes Physics Tutoring System: Lessons Learned. *Intl. Journal of Artificial Intelligence in Education* 15, 3 (2005), 147–204.

C. Wang, H. Chang, and K. A. Boughton. 2011. Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika* 76, 1 (Jan. 2011), 13–39.

G. Wiggins. 1998. *Educative Assessment: Designing Assessments To Inform and Improve Student Performance*. ERIC.