Sequence Matters but How Exactly?
A Method for Evaluating Activity Sequences from Data

Shayan Doroudi
Carnegie Mellon University

Kenneth Holstein
Carnegie Mellon University

Vincent Aleven
Carnegie Mellon University

Emma Brunskill
Carnegie Mellon University

In Proceedings of the 9th International Conference on Educational Data Mining
Raleigh, NC, USA
June 29-July 2, 2016

Sequence Matters, But How Exactly?
A Method for Evaluating Activity Sequences from Data

Shayan Doroudi[1], Kenneth Holstein[2], Vincent Aleven[2], Emma Brunskill[1]
[1]Computer Science Department, [2]Human-Computer Interaction Institute
Carnegie Mellon University
{shayand, kjholste, aleven, ebrun}@cs.cmu.edu

ABSTRACT

How should a wide variety of educational activities be sequenced to maximize student learning? Although some experimental studies have addressed this question, educational data mining methods may be able to evaluate a wider range of possibilities and better handle many simultaneous sequencing constraints. We introduce Sequencing Constraint Violation Analysis (SCOVA): a general method for evaluating alternative activity sequences using existing data. SCOVA can be used to explore many complex sequencing constraints, such as prerequisite relationships, blocking, interleaving, and spiraling. We demonstrate SCOVA on data collected from a fractions intelligent tutoring system (ITS). Some of our findings challenge our initial hypotheses regarding sequencing, illustrating the utility and versatility of the method. The method can also be applied to other learning environments, as long as the available data has substantial variability in students' activity sequences.

1. INTRODUCTION

How does the sequencing of pedagogical activities impact student learning? Answers to this question can both contribute to core learning sciences knowledge, as well as have important practical implications for how educational activities should be sequenced in order to maximize learning. As such, there has been significant interest in this issue, and prior research suggests that student learning can be quite sensitive to temporal sequencing (e.g., [16, 1, 15, 17]).

Prior work that tackles this problem mostly falls into either theoretical analyses or empirical studies. Unfortunately, conducting theoretical analyses of the cognitive demands of individual tasks and the interdependencies among multiple tasks [7, 10, 3] can be prohibitively time consuming for large curricula. In addition, such analyses may be particularly vulnerable to various cognitive biases, such as expert blind spots [12]. Considerable experimental research has examined the effects of activity sequencing along various dimensions, including interleaving versus blocking of topics [1, 17] and sequencing of activities according to the degree of scaffolding they provide [15, 8]. However, such classroom experimental studies typically compare only two or three possible conditions, in contrast to the enormous number of orderings possible (at least exponential in the number of activity categories of interest).

An educational data mining approach could allow us to evaluate a much broader range of possible orderings in order to better understand which sequences may be optimal. Moreover, it might be possible to apply such techniques to any datasets that have considerable variation in how they order instructional content for students. These include datasets generated from educational technologies that present activities in a partially or fully randomized order (e.g., [13]), those that adaptively present activities in response to measured student variables (e.g., [4]), and those that provide students with some degree of control over activity selection (e.g., [11]).

We are particularly interested in investigating which orderings over a variety of topics and activity types are most effective for maximizing student learning and performance. Prior educational data mining approaches have focused on examining pairwise dependencies between instructional items (e.g., individual skills, problems, or problem sets) in a curriculum, in order to infer underlying prerequisite structure [5, 21, 18]. The prerequisite structures learned via such methods could be used, for example, to inform adaptive problem selection algorithms that avoid presenting a given item until the student is believed to have mastered its prerequisites [7]. Other methods for detecting ordering effects over instructional items have additionally relied upon the use of fitted Bayesian Knowledge Tracing (BKT) models [13, 19], and have thus depended upon strong assumptions about student learning. Whereas these prior approaches are typically limited to discovering pairwise relationships between items, and have tended to assume that these items are presented in a blocked fashion, we wish to examine the impacts on student learning and performance of more complex (and potentially softer) sequencing constraints.

We investigate the question of optimal topic and activity type sequencing in the context of our fractions intelligent tutoring system (ITS) [6]. Our tutor covers three broad topics (making and naming fractions, fraction equivalence and ordering, and fraction addition) and three different types of activities that correspond to learning mechanisms in the theoretical Knowledge-Learning-Instruction (KLI) framework: sense-making, induction and refinement, and fluency-building processes [9]. While previous experimental work has investigated the optimal sequencing of activity types under the KLI framework [14], there has been little empirical work investigating the optimal sequencing of topics in a fractions curriculum, and no work to our knowledge examining how the optimal sequencing of activity types may vary across topics.

We develop a general-purpose method for leveraging log data to evaluate and compare different ways of sequencing activities. We believe our method for evaluating sequencing constraints can be utilized to discover how to sequence activities in a variety of learning environments. We tested our method on log data from our fractions tutor and found results that countered our initial hypotheses on how to order both topics and activity types. We also found that the optimal ordering over KLI activity types may vary from topic to topic, but that for the most part, these orderings were consistent with what was suggested by prior literature [14].

2. SEQUENCING CONSTRAINT VIOLATION ANALYSIS (SCOVA)

We first describe our general method, and then present the particular instantiations of our method that we used in our analyses in Section 3. Sequencing Constraint Violation Analysis (SCOVA) is a method for analyzing different sequencing constraints and identifying which ones lead to the best student performance. SCOVA takes as input a set of student trajectories (which contains the sequence of problems given to each student and the students' responses to those problems) and a cost function for each set of sequencing constraints that one wants to evaluate. The cost function is a function over student trajectories that specifies how often a particular set of sequencing constraints is violated; in particular, it assigns to each student's sequence a number of violations up to the total length of the sequence.

Many different types of sequencing constraints can be considered. For example, one sequencing constraint could be that a student must be given at least one instance of problem type X before the student is given problem type Y. For this constraint, whenever problem Y is presented to a student before any instance of problem X, that student trajectory incurs one violation. Another constraint could be that problem X should *always* appear immediately before problem Y, so whenever a student sees problem Y without seeing problem X right before it, that sequence incurs a violation. For such constraints, the cost function is simply the number of problems where the constraint is violated. However, another sequencing constraint could suggest that a student's trajectory should match a particular desirable sequence, and our cost function in that case could be the Levenshtein distance[1] between the student's sequence and the desirable sequence. We can also consider sets of more than one sequencing constraints: for example, the constraints could specify that problem X should come before problem Y and problem Y should come before problem X. In this case, the cost function counts every time *any* constraint is violated.

Unlike many existing methods (e.g., [13, 21, 19]), SCOVA is not limited to evaluating pairwise orderings. Indeed, SCOVA can handle much more general constraints on order sequencing, such as blocking, interleaving, and spiraling. SCOVA can also handle constraints that depend not just on the prior history of problems given, but also on the student's performance and interactions (such as performance on prior activities, pretest score, or measures of affect).

Given the cost functions and student trajectories, SCOVA proceeds as follows for each set of sequencing constraints that we want to evaluate. We first use the cost function to compute the proportion of violations for every student's sequence by dividing the cost of the sequence by the length of the sequence. We next use the proportion of violations as an input variable in a linear regression model that predicts some measure of student performance (e.g., within-tutor performance, posttest score, or learning gains), and fit the parameters that maximize the log likelihood of the resulting model.

To evaluate the impact of a particular set of sequencing constraints, we look at two measures. First, we compute the Bayesian Information Criterion (BIC) of the linear regression model fit for violations of those constraints. This provides us with a way to compare different sequencing constraints; a model with a lower BIC score provides a better fit of the student data (as evaluated by log likelihood, adjusted for the number of parameters of the model). However, BIC alone simply measures predictive fit, not whether the sequencing constraints are beneficial for students or harmful. To understand whether the sequencing constraints may have a positive or negative impact on the outcome variable, we look at the sign of the coefficient of the violation variable in the fit linear model. We limit our attention to models where the proportion of violations has a negative coefficient—that is, models where violating the sequencing constraints is associated with worse student performance. Among these models, we can then compare the sequencing constraints by comparing the BICs of their models.

Recall that SCOVA can handle multiple sequencing constraints conjunctively (e.g., example problem X should come before Y and Y before Z). This makes the most sense when the different sequencing constraints are mutually exclusive, i.e., we cannot incur more than one violation on any particular problem. However, we may want to consider different sequencing constraints that can occur simultaneously and perhaps constrain different aspects of student trajectories (e.g., for example one might constrain the ordering of topics and the other might constrain the ordering of activity types). SCOVA can be extended to simultaneously consider the impact of these different sequencing constraints *disjunctively.* To do so, we learn a predictive linear regression model with one input variable for each set of sequencing constraints. When we have more than one set of sequencing constraints in our model, we focus our attention on models that have negative coefficients for *every* predictor corresponding to violations of sequencing constraints. If the BIC of a model

[1] The Levenshtein distance, often referred to as edit distance, is a standard measure of distance between two sequences, measuring the smallest number of insertions, deletions, and substitutions to change one sequence into another. It is a valid cost function since it takes on a value between 0 and the length of the sequences.

that takes two sequencing constraints into account is lower than that of each of the models that consider just one of the sequencing constraints individually, it suggests that both ordering constraints are important but capture different aspects of student performance. We can also compare the relative effects of violating different sequencing constraints by comparing the coefficients within the same model.

3. EVALUATION DOMAIN

As a concrete example, we now describe how we used our proposed approach to evaluate the impact of ordering on student learning and performance when using an online fractions tutor for fourth and fifth grade fractions topics [6]. The tutor covers topics emphasized in the Common Core, a set of non-binding national standards for mathematics education in the US: making and naming fractions on the number line (MN), fraction equivalence and ordering (EQ), and fraction addition (ADD).[2] The tutor was originally developed to investigate the potential benefits of using a broader range of instructional activity types than is typical of an ITS. Tutor activities were designed to promote each of the 3 categories of learning mechanisms posited under the KLI framework [9]: sense-making (SM), induction and refinement (IR), and fluency-building (F). The tutor's curriculum includes activities targeting each of these categories of learning mechanisms, for each of the main topics.

Under KLI, SM processes correspond to "explicit, verbally mediated learning in which students attempt to understand or reason" [9], IR processes are defined as non-verbal learning processes that improve the accuracy of knowledge, and fluency processes are non-verbal processes that strengthen memory and enable students to apply their procedural knowledge faster and more fluently. As such, SM activities in our tutor were designed to promote conceptual understanding through an interleaving of brief instructional videos with exercises intended to support self-explanation. By contrast, IR activities in our fractions tutor were designed to emphasize procedural learning and practice via fine-grained task decomposition and step-level guidance – as is typical of ITSs [20]. Finally, fluency-building activities were designed to promote the development of fluent performance on minimally decomposed problem-solving exercises. A more detailed description of our operationalization of these three activity types can be found in [6].

3.1 Sequencing Constraints

We consider a variety of sequencing constraints over both topics and activity types in our analyses. Since we have three topics and three activity types there are six potential orderings of each. For each of the following constraints (aside from the baselines at the end) we consider them with respect to each of the six possible orderings (for either topics or activity types).

3.1.1 Exposure-Based Constraints

Exposure-based constraints stipulate that students be exposed to (i.e., carry out) one topic/activity type a certain number of times before being exposed to the next. Every time the student receives a problem before being exposed to its "prerequisite" enough times, a violation is incurred. We define two categories: *Exposure-based topic constraints* require that students do at least one problem of a topic before seeing a problem of the next topic. *Exposure-based type constraints* require that within each topic, students should do one problem of an activity type before seeing the next activity type, without constraining the order of topics. Note that we can have the ordering over activity types fixed for every topic, or we can let it vary. If we let it vary, there are $6^3 = 216$ possible exposure-based *varying* type constraints.

3.1.2 Performance-Based Constraints

Performance-based constraints stipulate that students should reach a certain level of within-tutor performance on a topic/activity type before being exposed to the next. Every time the student receives a problem when their recent performance on its "prerequisite" is not beyond some threshold, a violation is incurred. Notice that even though such a constraint may be satisfied for a given student at a certain point in time, it is possible that it will no longer be satisfied later on, if the student's performance drops. *Performance-based topic constraints* require that students' performance on the last 10 steps of the topic should be beyond some topic-specific threshold before they receive problems for the next topic. (These steps may be from one problem or span over several problems.) By contrast, *performance-based type constraints* require that within each topic, students' performance on the last 10 steps on a particular activity type should be beyond some threshold specific to that topic-type pair before they receive problems of the next activity type (for the given topic). As before, in addition to the six type constraints that are fixed per topic, we have 216 possible performance-based *varying* type constraints.

We selected thresholds to detect a basic level of competency with problems of a particular activity type within a topic—a lower bar than mastery. The thresholds shown in Table 1 were obtained by taking the average student performance on the last 10 steps upon doing two problems of the given topic or topic-type pair

3.1.3 Blocking and Interleaving-N Constraints

To show the flexibility of the SCOVA method in considering sequencing constraints beyond straightforward prerequisite relationships, we consider whether topics and activity types should be interleaved or blocked with respect to topics/types. We measure violations in terms of Levenshtein distance from a particular sequence. The *blocking topic constraint* stipulates that for every student, the first third of their sequence (rounding up) should correspond to the first topic, the second third (rounding up) should correspond to the second topic, and the last third should correspond to the last topic. This is not a sequence we would typically be able to assign in practice, because we do not generally know how many problems a student will do ahead of time, but it represents a pure form of blocking while guaranteeing students see all of the activity types. The *interleaving-N topic constraints*, for $N = 1, \ldots, 6$, require sequences that

[2] In the fractions tutor, activities within each of these three broad topics broke down further into multiple subtopics. For example, fraction equivalence and ordering activities included activities on finding common denominators, reducing fractions, and identifying equivalent fractions using number lines, among other subtopics. In addition, individual activities typically targeted a number of finer-grained skills.

MN	EQ	ADD	MN/SM	MN/IR	MN/F	EQ/SM	EQ/IR	EQ/F	ADD/SM	ADD/IR	ADD/F
0.453	0.360	0.206	0.415	0.514	0.125	0.356	0.547	0.308	0.262	0.158	0.269

Table 1: Thresholds used for performance-based topic and type constraints. Notice that for the type constraints, we have distinct thresholds for each topic. The thresholds were obtained by taking the average student performance on the last 10 steps upon doing two problems of the given topic or topic-type pair.

give N problems of the first topic followed by N problems of the second topic followed by N problems of the third topic. However, if a student did less than $3N$ problems in total, we instead use the sequence used for the blocking constraint, in order to check whether they get reasonable exposure to all three topics.

3.1.4 Proportion-Only Baselines

To see if ordering topics or activity types actually matters, we compare to baselines that just use the proportions of topics or activity types in the sequence as predictors to predict within-tutor performance. Note that our two baselines each have two predictors (e.g., for activity types, we have one for proportion of SM and proportion of IR; the proportion of fluency-building activities is linearly dependent on the first two and so it is not needed in the model).

3.2 Hypotheses

We started data analysis with several hypotheses about the best order of topics and activity types. We note however that in order to illustrate our method, the specific hypothesized best order does not matter, although it does matter in illustrating that the method can produce unexpected (but reasonable) results.

3.2.1 Topic Dependencies

Our first hypothesis is that in early fractions learning, topics build on each other in the following way. MN helps students build a basic representation of fractions as numbers that have a magnitude, represented by their place on the number line. This representation is hypothesized to help in building an understanding of the notion of equivalence and the notion that fractions can be compared and ordered in terms of their magnitude. Moreover, equivalence would appear to be a strict prerequisite for addition of fractions with unlike denominators, because fractions with unlike denominators need to be converted to equivalent fractions before they can be added. Thus, the hypothesized best topic order is MN-EQ-ADD. Topics may not need to be fully blocked (i.e., presenting all MN activities before any EQ activities, and all EQ activities before any ADD activities), but it may be better for students to initially be exposed to topics in this order and perhaps continue to see the different topics in an interleaved fashion (as interleaving has been show to be beneficial [1, 17]).

3.2.2 Type Dependencies

As mentioned, the KLI framework distinguishes between three distinct classes of learning mechanisms, SM, IR, and F. It does not, however, make any claims regarding the order in which these processes might be most effective or even whether each class of mechanisms is needed when learning in a complex domain (such as fractions). There has been little prior work investigating how instructional activities targeting each of the KLI activity types can best be sequenced to maximize student learning and performance. However, [14] previously found that presenting students with SM activities before presenting them with fluency-building activities is beneficial when teaching connection making between multiple graphical representations of fractions. Given the dearth of prior work in this area, we do not have very strong expectations regarding the best order of these different activity types within a topic. However, in line with the work by [14], our hypothesis is that SM-targeting activities should come first, then IR-targeting activities, and finally, F-targeting activities. A second reason to expect that it is effective to do IR activities before F activities is that in our tutors, IR activities provide more elaborate scaffolding than F activities. As before, we do not mean to suggest a fully blocked ordering may be best, but also consider orders that interleave activity types with the hypothesized SM-IR-F order strictly observed early on.

3.3 Data

We collected data from 347 students using our ITS (in 20 classrooms across four different schools). The data was initially collected for a randomized control trial comparing three adaptive problem selection policies and two non-adaptive policies. The three adaptive policies had quite a bit of variation in the kinds of trajectories given to students; they thus provide data that is a good fit for SCOVA. However, the non-adaptive policies resulted in trajectories that were identical in how they sequenced topics and activity types, so we did not use data from those policies in our analyses (leaving 211 students). Students were given a pretest, followed by using the tutor for typically four class periods, and were finally given a posttest that was identical to the pretest. Each student worked at their own pace and completed as many problems as they could during the allotted time, resulting in a tail of students who did many more problems than average. This could present a confound in our analysis since students who do many problems are more likely to be high performing students, as well as violating sequencing constraints less than others (because they are likely to do many problems after satisfying all sequencing constraints). We thus limited our analyses to students who did 60 or fewer problems (197 students).

3.4 Modeling

In the SCOVA framework, we fit a linear regression model with predictors corresponding to the proportion of violations of one or more sets of sequencing constraints. The outcome variable we used was the within-tutor performance of students on all problems of the tutor with each topic-type pair having an equal weight (e.g., each student's performance on MN/SM problems has an equal weight to their performance on EQ/F problems). If a student received no problems of a

	Topic Constraints			Type Constraints	
	Exposure	Performance		Exposure	Performance
MN-EQ-ADD	**-236.28**	**-299.69**	SM-IR-F	**-226.16**	**-236.84**
EQ-MN-ADD	**-244.39**	**-319.13**	IR-SM-F	**-208.59**	**-218.94**
MN-ADD-EQ	**-201.04**	**-274.17**	SM-F-IR	-193.39	**-200.89**
EQ-ADD-MN	**-201.26**	**-254.75**	IR-F-SM	**-196.85**	**-217.20**
ADD-MN-EQ	**-193.81**	**-199.80**	F-SM-IR	**-202.91**	**-224.57**
ADD-EQ-MN	-205.73	**-193.84**	F-IR-SM	**-192.97**	**-200.32**
Proportion-Only	-233.48		Proportion-Only	-201.77	

Table 2: Comparison of BICs of individual exposure-based and performance-based constraints as well as proportion-only baselines. Aside from the proportion-only baselines, BICs corresponding to models where the coefficient of the predictor is negative are shown in bold. The smallest BIC in each column is underlined.

	SM-IR-F	IR-SM-F	SM-F-IR	IR-F-SM	F-SM-IR	F-IR-SM
MN-EQ-ADD	**-246.09**	**-232.81**	-232.95	**-231.28**	**-231.11**	-234.97
EQ-MN-ADD	**-249.30**	**-251.63**	-242.24	**-247.12**	**-240.24**	**-239.11**
MN-ADD-EQ	**-224.69**	**-208.31**	-197.71	**-198.37**	**-202.08**	-196.00
EQ-ADD-MN	**-223.54**	**-217.35**	-201.94	**-203.99**	**-200.60**	-197.16
ADD-MN-EQ	**-225.26**	-205.57	-188.94	-191.63	**-197.83**	-188.64
ADD-EQ-MN	-227.92	-219.54	-200.48	-208.98	-210.61	-201.07

Table 3: Comparison of BICs of models combining exposure-based topic and type constraints. BICs corresponding to models where the coefficients of both predictors are negative are shown in bold. The smallest BIC is underlined.

topic-type pair, then the average is only over the topic-type pairs they received. One could also add other predictors to improve the model fits and potentially control for other confounds. We add the student's pretest score as a predictor to all of our models as this improved the model fit.

4. RESULTS

Table 2 shows the BICs of models with only a single ordering constraint predictor corresponding to performance-based and exposure-based topic and type sequencing constraints in addition to BICs of the two proportion-based baselines. First, we notice that the lowest BIC models using exposure-based and performance-based ordering constraints have a better fit than the baseline models, which, as mentioned, only consider the proportion of activities given for either topic or activity type. This suggests that ordering of topics and activity types makes a difference beyond just the frequency with which they appear.

Second, we find that the lowest BICs for the sequencing constraints over topics are lower than the lowest BICs for sequencing constraints over activity types, especially for the performance-based constraints. This suggests that sequencing over topics might be more important than activity type ordering. This is also supported by the coefficients in the fitted linear regression models; for example, the coefficient for the best fitting performance-based topic constraints is -0.37, whereas for the best fitting performance-based type constraints, it is -0.23.

Third, for both the exposure-based and the performance-based constraints, the models for EQ-MN-ADD have the lowest BICs among all the topics models and the models for SM-IR-F have the lowest BICs among all the types models. We also find that the models that put fractions addition first either have the worst BICs or have positive coefficients (i.e., violation of constraints correlates with increased student performance), which makes sense, as we really do not think students should be doing addition (potentially with unlike denominators) before fraction equivalence. Likewise, the models with the best BICs and largest negative coefficients are the ones that put ADD last.

Finally, we find that the performance-based constraints have lower BICs than the exposure-based constraints. This reasonably seems to suggest that students' within-tutor performance can be predicted more accurately when we take into account the extent to which individual students reached a basic level of competence on one topic/type before being exposed to the next topic/type. We must note, however, that for the performance-based metric, the number of violations is impacted by a student's performance, and is thus related to the outcome variable in a confounded way. For example, a student who does very well on the tutor would be more likely to get fewer performance-based violations for any sequence than a student who does poorly on the tutor, partially explaining the lower BICs for performance-based models than exposure-based models. While we cannot conclude that performance-based constraints are better than exposure-based constraints from this analysis, we hypothesize that the relative ranking of different orders of topics/types may not be impacted severely by this confound.

To start to understand the interaction of type and topic ordering constraints on within-tutor student performance, we fit linear regression models that used two prerequisite violation input variables: one for one of the six topic orderings, and one for one of the six type orderings. Table 3 shows

	SM-IR-F	IR-SM-F	SM-F-IR	IR-F-SM	F-SM-IR	F-IR-SM
MN-EQ-ADD	**-319.39**	**-297.22**	**-301.80**	**-298.25**	**-299.90**	**-296.39**
EQ-MN-ADD	**-328.84**	**-330.35**	**-314.33**	**-336.70**	**-330.46**	**-317.38**
MN-ADD-EQ	**-300.02**	**-285.10**	**-270.36**	**-283.20**	**-286.98**	**-269.96**
EQ-ADD-MN	**-269.67**	**-280.47**	**-249.80**	**-279.55**	**-261.14**	**-250.23**
ADD-MN-EQ	**-239.09**	**-215.61**	**-203.15**	**-214.25**	**-220.07**	**-199.02**
ADD-EQ-MN	**-233.34**	**-213.69**	**-196.73**	-211.92	**-219.29**	**-195.28**

Table 4: Comparison of BICs of models combining performance-based topic and type constraints. BICs corresponding to models where the coefficients of both predictors are negative are shown in bold. The smallest BIC is underlined.

	Exposure-Based		Performance-Based	
	Coefficient	p-value	Coefficient	p-value
Intercept	0.37	$< 2 * 10^{-16}$	0.45	$< 2 * 10^{-16}$
Pretest	0.025	$3.45 * 10^{-8}$	0.023	$4.77 * 10^{-10}$
Topic Violations	-0.20	$8.07 * 10^{-7}$	-0.36	$< 2 * 10^{-16}$
Type Violations	-0.17	$4.20 * 10^{-6}$	-0.22	$2.27 * 10^{-10}$
BIC	-260.77		-355.00	
Adjusted r^2	0.39		0.62	

Table 5: Best fitting models incorporating both topic constraints and varying type constraints. The lowest BIC model according to exposure-based constraints suggests IR-SM-F for EQ, SM-IR-F for MN, and F-IR-SM for ADD, and the lowest BIC model according to performance-based constraints suggests the ordering IR-SM-F for EQ, SM-IR-F for MN, and IR-SM-F for ADD.

the BICs for all 36 models that have pairs of violations of exposure-based topic and type constraints as predictors, and Table 4 shows analogous results for pairs of performance-based constraints. We find that both for exposure-based and performance-based constraints, the model with the lowest BIC uses the EQ-MN-ADD ordering over topics, but for exposure-based constraints the ordering over activity types is IR-SM-F, while for performance-based constraints it is IR-F-SM. Note that this is different from the lowest BIC ordering of activity types when using only type constraints (SM-IR-F, see Table 2). However, we find that for many other orderings over topics (e.g., MN-EQ-ADD and MN-ADD-EQ), the model with the lowest BIC is the one with the SM-IR-F ordering over activity types. This suggests that the best ordering over activity types may depend on how we sequence the topics.

Indeed, the best ordering over activity types might vary from topic to topic (e.g., to maximize student performance it may be best to give IR first for EQ but SM first for MN). To test this possibility, we searched for the lowest BIC model with a predictor corresponding to some *varying* type constraints and a predictor for one of the six topic constraints[3]. The lowest BIC model according to exposure-based constraints suggests the ordering IR-SM-F for EQ, SM-IR-F for MN, and F-IR-SM for ADD (although several models were within three BIC points including ones that suggests IR-SM-F for ADD), and the lowest BIC model according to performance-based constraints suggests the ordering IR-SM-F for EQ, SM-IR-F for MN, and IR-SM-F for ADD (although, again, several models were within three BIC points including ones that suggests IR-F-SM for ADD). Table 5 shows the coefficients and fits for both of these lowest BIC models. Notice that the coefficients for the topic constraints have larger magnitudes than those for the varying type constraints (although not much larger in the exposure-based model), suggesting again that sequencing over topics is more important than sequencing over activity types. Moreover, the coefficients of the topic and activity type constraints violation variables in Table 5 are not only highly significant (i.e., significantly different than 0), but also their magnitudes are quite substantial given the outcome variable is bounded between 0 and 1. This suggests that students who receive activities in an order that has a large proportion of sequencing constraint violations would be expected to have considerably worse performance on the tutor problems.

Finally, we turn to models based on blocking and interleaving constraints. Table 6 shows the results comparing interleaving-N constraints and blocking constraints for all six orderings over topics. Again we find that the model corresponding to the EQ-MN-ADD order has the lowest BIC, but interleaved in chunks of four problems. This agrees with our hypothesis that one should not simply present the topics in a blocked fashion. Interestingly, most of the other models, including ones corresponding to fully interleaving or blocking, have equally bad BICs, regardless of the topic order.

5. DISCUSSION

Our novel method for evaluating activity sequences led to a number of interesting findings about sequencing topics and activity types in our tutor, illustrating the utility of the method. We found that all of the models fit using various topic sequencing constraints unanimously suggested that

[3] This results in 1296 models to search over, as there are $6^3 = 216$ different varying type constraints and six different topic constraint orderings

	Interleaving-1	Interleaving-2	Interleaving-3	Interleaving-4	Interleaving-5	Interleaving-6	Blocking
MN-EQ-ADD	**-193.09**	**-201.03**	**-198.85**	**-198.33**	**-197.80**	**-195.52**	**-195.63**
EQ-MN-ADD	**-195.89**	**-197.01**	**-198.89**	**-211.94**	**-202.93**	**-194.93**	**-193.91**
MN-ADD-EQ	**-194.04**	**-193.27**	**-195.00**	-194.00	**-193.01**	**-195.47**	**-193.01**
EQ-ADD-MN	**-194.75**	**-193.81**	**-194.06**	**-196.07**	**-194.03**	**-193.08**	-194.08
ADD-MN-EQ	**-193.49**	**-193.14**	-192.97	-194.11	-194.62	-193.34	-193.50
ADD-EQ-MN	**-193.62**	**-193.04**	**-192.96**	-196.66	-203.76	-197.92	-195.97

Table 6: Comparison of BICs of models with interleaving-N constraints and blocking constraints. BICs corresponding to models where the coefficient of the predictor is negative are shown in bold.

EQ-MN-ADD is the best way to sequence topics (suggesting that students should at least have some exposure to EQ before MN and some exposure to MN before ADD). This challenges our initial hypothesis that MN-EQ-ADD is the optimal ordering for learning. This result seems to indicate that, in contrast to our hypothesis, learning to make and name fractions (MN) on the number line may be facilitated by knowledge and skill regarding fraction equivalence and ordering (EQ), more so than the other way around. This result may suggest that an understanding of relationships between multiple fractions can help with learning about making and naming individual fractions on the number line, to a greater degree than previously realized. However, we cannot rule out alternative explanations. For example, it could be that our tutor activities are not successful in helping students learn knowledge that transfers to other topics. We note that in the MN activities, students used the number line extensively, whereas they did not in the EQ activities; in the latter they almost exclusively used the symbolic notation of fractions. It may be that if both topics had used the number line, the work on making and naming fractions might have facilitated learning about equivalence and ordering more. Thus, our method for evaluating sequences raises questions about tutor design, which, if and when resolved, could potentially lead to a more effective tutor.

The results on sequencing of activity types were not as unequivocal. We found that the best sequence over activity types may well vary for topics, which is itself an interesting result. For MN and EQ, the models suggest SM should precede F. This result agrees with prior literature on how to order sense-making and fluency activities [14]. However, the relative ordering of SM and IR is not as clear, with it possibly being advantageous to give IR activities before SM activities in many cases, challenging our initial hypothesis.

One may wonder if our results can simply be explained in terms of ordering topics and activity types from easiest to hardest. However, this does not seem to be the case. Note that the performance thresholds in Table 1 provide a measure of difficulty for each topic and each topic-type pair. Based on this measure of difficulty, MN would be classified as easier as EQ, but we saw that our models suggest EQ should come before MN. Furthermore, according to this measure of difficulty, ADD/IR problems would be classified as the most difficult for fraction addition; however, our lowest BIC types models suggest that IR should either come first or second for fraction addition.

Despite the strengths of our method over some prior approaches, the current analysis has several limitations that should be taken into consideration. First, when adaptive problem selection algorithms assign problems to students based on their performance on past problems, the student's performance can itself impact the proportion of violations of sequencing constraints; thus, SCOVA provides correlational, not necessarily causal, information about the impact of orderings. We can avoid this confound by using data with randomized sequences of problems rather than sequences generated from adaptive policies. However, in many cases (as was the case here) we may not have access to randomly generated sequences, and randomized data can often be difficult to collect ethically if we believe that a random sequence could have negative effects on student learning. To test the degree to which this confounds affects our results, we checked if student's pretest scores are correlated with the proportion of violations of various sequencing constraints, which would indicate that students with more prior knowledge tend to adaptively be assigned problems that either obey or violate certain sequencing constraints more than students with less prior knowledge. While we did find such correlations for certain sequencing constraints, the coefficients of the pretest score variables used to predict sequencing constraint violations were less than 0.05 in magnitude, and seemed to indicate that higher-performing students tended to receive ADD earlier and EQ later than lower performing students, which is contrary to the sequences we found most predictive of within-tutor performance! Thus we do not think this confound had a worrisome impact on our results.

Second, ideally we would like to see how sequencing constraints impact student learning as measured via posttest scores rather than just within-tutor performance. However, we were unable to find strong correlations between the proportion of violations of sequencing constraints and the posttest scores of students. This is likely due to the fact that the posttest was comprised of only 16 items and as a result is only a noisy measure of a student's knowledge and does not capture the diversity of concepts taught on the tutor. Note that this is not however a limitation of SCOVA; in theory, SCOVA could be used to compare how various sequencing constraints impact posttest performance.

6. CONCLUSION

We have shown how SCOVA can be used to test a much broader range of sequencing constraints than existing methods (e.g., [13, 21, 19])—including exposure-based, performance-based, interleaving, and blocking constraints. Furthermore, we have shown that when analyzing all of these results in conjunction with each other, a few trends can emerge that can inform practitioners about how to sequence problems. In the case of our fractions tutor, our results suggest presenting students with fraction equivalence before making and naming on the number line, and presenting the latter before fraction addition. In addition, our results suggest that we should not present the topics in a fully blocked fashion, but rather present four problems of each topic at a time. As for activity types, our results suggest that sense-making should typically come before fluency-building, in agreement with prior literature [14], but that the optimal ordering of activity types may vary for certain fractions topics.

These results suggest just some of the use cases of the SCOVA framework. SCOVA can easily be used to test a broader variety of sequencing constraints, as well as informing old debates about sequencing. For example, prior literature has suggested benefits of interleaving in some cases and of blocking in others [2]. From such results, one may be led to wonder "what is the optimal form of interleaving, and under which circumstances?" While it may be difficult to immediately address such a question in an experimental study, due to the sheer size of the space of sequencing constraints, we can easily analyze such a question using SCOVA.

SCOVA can be of benefit to researchers and practitioners in several ways. First, it can lead to refining hypotheses and determining which questions to test empirically (e.g., testing whether EQ should actually precede MN). Second, it can lead to improving the design of tutor problems (e.g., making EQ problems that use the number line and hence build off of the problems that cover making and naming fractions). Finally, it can help with the construction of adaptive policies (e.g., by determining the order of topics in a mastery learning policy as suggested by performance-based constraints).

7. ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education.

8. REFERENCES

[1] W. Battig. Intratask interference as a source of facilitation in transfer and retention. *Topics in learning and performance*, pages 131–159, 1972.

[2] P. F. Carvalho and R. L. Goldstone. The benefits of interleaved and blocked study: different tasks benefit from different schedules of study. *Psychonomic bulletin & review*, 22(1):281–288, 2015.

[3] R. E. Clark, D. Feldon, J. J. van Merriënboer, K. Yates, and S. Early. Cognitive task analysis. *Handbook of research on educational communications and technology*, 3:577–593, 2008.

[4] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *UMUAI*, 4(4):253–278, 1995.

[5] M. C. Desmarais and X. Pu. A bayesian student model without hidden nodes and its comparison with item response theory. *IJAIED*, 15(4):291–323, 2005.

[6] S. Doroudi, K. Holstein, V. Aleven, and E. Brunskill. Towards understanding how to leverage sense-making, induction and refinement, and fluency to improve robust learning. In *EDM*, pages 376–379, 2015.

[7] J.-C. Falmagne, M. Koppen, M. Villano, J.-P. Doignon, and L. Johannesen. Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97(2):201, 1990.

[8] S. Kalyuga. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4):509–539, 2007.

[9] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[10] K. Korossy. Modeling knowledge as competence and performance. *Knowledge spaces: Theories, empirical research, and applications*, pages 103–132, 1999.

[11] Y. Long and V. Aleven. Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In *AIED*, pages 219–228. Springer, 2013.

[12] M. J. Nathan, K. R. Koedinger, and M. W. Alibali. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proc. of Cognitive Science*, pages 644–648, 2001.

[13] Z. A. Pardos and N. T. Heffernan. Determining the significance of item order in randomized problem sets. 2009.

[14] M. A. Rau, V. Aleven, and N. Rummel. Complementary effects of sense-making and fluency-building support for connection making: A matter of sequence? In *AIED*, 2013.

[15] A. Renkl and R. K. Atkinson. Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational psychologist*, 38(1):15–22, 2003.

[16] F. E. Ritter, J. Nerb, E. Lehtinen, and T. M. O'Shea, editors. *In order to learn: how the sequence of topics influences learning*. Oxford University Press, 2007.

[17] D. Rohrer and K. Taylor. The shuffling of mathematics problems improves learning. *Instructional Science*, 35(6):481–498, 2007.

[18] R. Scheines, E. Silver, and I. Goldin. Discovering prerequisite relationships among knowledge components. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 355–356, 2014.

[19] S. Tang, E. McBride, H. Gogel, and Z. A. Pardos. Item ordering effects with qualitative explanations using online adaptive tutoring data. In *Proc. of L@S*, pages 313–316. ACM, 2015.

[20] K. Vanlehn. The behavior of tutoring systems. *IJAIED*, 16(3):227–265, 2006.

[21] A. Vuong, T. Nixon, and B. Towle. A method for finding prerequisites within a curriculum. In *EDM*, pages 211–216, 2011.