Ecological content sequencing: from simulated students to an effective user study

John Champaign* and Robin Cohen

David R. Cheriton School of Computer Science,
University of Waterloo,
Waterloo, ON, N2L 3G1, Canada
E-mail: jchampai@uwaterloo.ca
E-mail: rcohen@uwaterloo.ca
*Corresponding author

Abstract: In this paper, we present an algorithm for reasoning about the sequencing of content for students in an intelligent tutoring system, influenced by McCalla’s ecological approach. We record with each learning object those students who experienced the object, together with their initial and final states of knowledge, and then use these interactions to reason about the most effective lesson to show future students based on their similarity to previous students. We validate our approach through a novel method of validation, providing details of the model of learning used in the simulation and the results obtained in order to demonstrate the value of our model. Beyond confirmation through simulations of student learning, we report on a study with human users and expand on a previous pilot study. We demonstrate the effectiveness of our algorithms for selection of learning objects to solidify the overall defence of our approach.

Keywords: peer-based intelligent tutoring; simulating students; ecological approach to instructional design; modelling learners.

Reference to this paper should be made as follows: Champaign, J. and Cohen, R. (2013) ‘Ecological content sequencing: from simulated students to an effective user study’, Int. J. Learning Technology, Vol. 8, No. 4, pp.337–361.

Biographical notes: John Champaign completed his PhD degree in 2012 at the University of Waterloo and has since worked as a Lecturer, also at the University of Waterloo, and as a Post-Doctoral Research Fellow at the Massachusetts Institute of Technology. He conducts cross-disciplinary research in the fields of artificial intelligence (AI), human-computer interaction (HCI) and software engineering, with a particular focus on exploring the use of social networking techniques (collaborative filtering and crowdsourcing) to identify and refine content within an intelligent tutoring system.

Robin Cohen is a Professor in the David R. Cheriton School of Computer Science, where she has been a faculty member since 1984. She conducts research in the artificial intelligence areas of multiagent systems, intelligent interaction and user modelling. From 2002–2005, she served as the Associate Dean for Graduate Studies and Research in the Faculty of Mathematics, and held a Faculty Fellowship. She received a Cheriton Faculty Fellowship in Fall 2010.
This paper is a revised and expanded version of a paper entitled ‘Exploring the effects of errors in assessment and time requirements of learning objects in a peer-based intelligent tutoring system’ presented at FLAIRS, Palm Beach, 2011 and ‘A model for content sequencing in intelligent tutoring systems based on the ecological approach and its validation through simulated students’ presented at FLAIRS, Daytona Beach, 2010.

1 Introduction

In this paper, we revisit a model for reasoning about the sequencing of content for students in an intelligent tutoring environment that was first presented at the Flairs conference, intelligent tutoring track in Champaign and Cohen (2010, 2011). Our approach is one of peer-based intelligent tutoring and is motivated by McCalla’s ecological approach, which advocates attaching models of learners to learning objects in a repository in order to be able to subsequently leverage that history towards directing the learning of new students. This vision of education is one where peer experiences should inform the study of students but where it is reasoning about past benefits that will drive the proposed presentation of content [and not, as has been more typically done in peer-based tutoring systems, an environment where peers are advising each other, in real-time (Vassileva, 2008; Brooks et al., 2006; Hausmann et al., 2008)].

In order to determine how to propose content to be presented to students (within any application domain) our solution lies in effectively modelling the similarity of a current student to other peers [not unlike the approach of collaborative filtering recommender systems (Breese et al., 1998; Herlocker et al., 2004)] and in carefully representing the benefits that peers have derived, in order to ultimately select those learning objects which have the greatest potential for yielding improvements to the learning of the current student.

In this work we endeavour to examine the effect of determining, based on past benefit provided by the learning objects and similarities between the students, the best learning object to show to a learner at a specific point in their course of study. In order to examine this effect, we ignore other data about the learning object and effects of the sequence between learning objects.

After presenting our model, we proceed to demonstrate its value in depth through a novel process which we refer to as simulated student learning. To do so, we require a detailed modelling of the target levels of instruction of learning objects and their potential impact on student learning, so that we can measure the success of our algorithms against a metric of the mean average knowledge of all students experiencing learning objects presented according to our methods (and in comparison with other benchmark algorithms). In this paper, we present more elaborate results than have been previously reported.

Also going beyond earlier presentation of our approach and its validation through simulation, we describe a study with human participants where we contrast two variants of our techniques to assign learning objects to humans with two control groups. We do this within the domain of home healthcare, specifically learning to become the primary caregiver for a child with autism spectrum disorder. In so doing, we offer complementary
support for the value of our approach and for its applicability in directing the learning of students within intelligent tutoring environments.

In all, we offer confirmation of the value of an ecological approach to intelligent tutoring, through a detailed articulation of how to select the best learning objects for each student within a repository tagged with previous student histories. We also illustrate the potential of our proposed approach through dual methods of validation:

1. simulated student learning, which we are able to demonstrate as an indispensable approach for measuring the value of tutoring models in large peer-based environments
2. a user study in a truly vital scenario, that of patient-led management of healthcare.

2 Our model for curriculum sequencing

Given a number of learning objects, a history of interactions between the objects and past students and a student ready to be assigned work, which of the possible learning objects should be assigned to her? Our approach to this problem uses a collaborative filtering inspired algorithm to balance the similarity of the current student to past students with the benefit those past students have obtained, in order to create a personalised recommendation.

**Algorithm 1  Collaborative learning algorithm**

```plaintext
Input the current-student-assessment
for each learning object: do
    Initialise currentBenefit to zero
    Initialise sumOfBenefits to zero
    Input all previous interactions between students and this learning object
    for each previous interaction on learning object: do
        similarity = calculateSimilarity(current-student-assessment, interaction-initial-assessment)
        benefit = calculateBenefit(interaction-initial-assessment, interaction-final-assessment)
        sumOfBenefits = sumOfBenefits + similarity * benefit
    end for
    currentBenefit = sumOfBenefits / numberOfPreviousInteraction
    if bestObject.benefit < currentBenefit then bestObject = currentObject
end for
if bestObject.benefit < 0 then
    bestObject = randomObject
end if
```

Our proposed algorithm for determining which learning objects to present to students is presented in Algorithm 1. We assume that we are tracking a set of values, \( v[j, l] \), representing the benefit of the interaction for user \( j \) with learning object \( l \). \( v[j, l] \) is determined by assessing the student before and after the interaction, and the difference in assessed knowledge is the benefit. We also record for each learning object the previous
interactions of students with that object, in terms of their initial and final assessments. We assume that a student’s knowledge is assessed by mapping it to 18 concrete levels: A+, A, A–, .. F+, F, F–, each representing \( \frac{1}{18} \)th of the range of knowledge. This large-grained assessment was used to represent the uncertainty inherent in assessing student knowledge, and only this large-grained assessment is used to reason about the students’ ability in our approach.

The anticipated benefit of a specific learning object \( l \), for the active user, \( a \), under consideration would be:

\[
p[a, l] = \kappa \sum_{j=1}^{n} w(a, j) v(j, l)
\]

where \( w(a, j) \) reflects the similarity \( \in (0, 1] \) between each user \( j \) and the active user, \( a \), and \( \kappa \) is a normalising factor. \( \frac{1}{\text{difference}} \) was used as the value for \( \kappa \) in this work where \( n \) is the number of previous users who have interacted with learning object \( l \). \( w(a, j) \) was set as

\[
\text{difference} = \text{initial assessment} - \text{current-student-assessment},
\]

So the difference of A+ and B– would be 5 and the difference of D+ and C– would be 1. This is shown as the calculate similarity function in Algorithm 1.

\( v(j, l) \) is also computed using a difference, not an absolute difference but an actual difference (between the initial and final assessments). For example, \( v(j, l) \) where \( j \) is initially assessed as A+ and finally assessed at B– would be –5, while where \( j \) is initially assessed at B– and finally assessed at A+ would be 5. This is shown as the calculate benefit function in Algorithm 1.

In the absence of other criteria, a user \( a \) will be assigned the learning object \( l \) that maximises \( p[a, l] \). In the case that the maximum \( p[a, l] \) is a negative anticipated benefit, a random learning object will be assigned to the user.

It is important to note that for the simulations to validate the collaborative learning algorithm we used a multi-dimensional model of knowledge. For clarity we will first discuss similarity and benefit in terms of a single dimension; however, this technique would be most appropriately applied in multiple dimensions, with each axis representing a facet of the overall knowledge. Standard hyper-dimensional geometry allows the extrapolation of these examples.

### 2.1 Example

Here we provide a simplified example for illustration. Suppose we track each learning object, LO with [index, [StudentID, initial assessment], [final assessment]]. After multiple interactions with three students, S1, S2 and S3, the experiences are shown in Table 1.

**Table 1**  
Student experiences interacting with learning objects

<table>
<thead>
<tr>
<th>LO[1; S1(B, C), S3(B, A+)]</th>
<th>LO[4; S1(C, A–), S2(B, B)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO[2; S1(A, A), S3(C, A–)]</td>
<td>LO[5; S3(C+, B)]</td>
</tr>
<tr>
<td>LO[3; S2(B–, A)]</td>
<td></td>
</tr>
</tbody>
</table>
At this point the system is slightly positive on the benefit of LO[1] for B students (because one time it raised a student to A+, and another time it lowered a student to C). It is neutral on LO[2] for A students (the lesson did not change the student’s assessment), and very positive for C students (since it raised a C student to an A–). Similarly LO[3] is good for B– students, LO[4] is very good for C students and neutral for B students, and LO[5] is good for C+ students.

Suppose the system were now considering which lesson to recommend for a student, S4, with current-student-assessment of B+. Per Equation 1, it would consider LO[1] to have a currentBenefit of \( \frac{1}{14} \times 3 + \frac{1}{14} \times 4 \times 2 = 0.25 \), LO[2] a currentBenefit of \( \frac{1}{14} \times 0 + \frac{1}{14} \times 5 \times 2 = 0.67 \), LO[3] a currentBenefit of \( \frac{1}{14} \times 4 = 1.33 \), LO[4] a currentBenefit of \( \frac{1}{14} \times 5 + \frac{1}{14} \times 0 \times 2 = 0.5 \) and LO[5] a currentBenefit of \( \frac{1}{14} \times 2 = 0.5 \).

In this situation, LO[3] would be recommended. After the system’s interaction between S4 and LO[3] there will be more information to reason about with future students. The next B+ student will be assigned to LO[3] if S4 has a positive experience, but will instead be assigned to LO[2] if S4 has a neutral or negative experience with LO[3]. This assumes that no additional students use these learning objects in between S4’s interactions.

3 Simulations to validate our model

We used simulated students to validate our content sequencing approach.

In this discussion below, for clarity, we discuss learning objects with just 1 overall knowledge dimension. Our solution leverages the idea that learning objects should raise a student’s knowledge and improve their learning if their target level is close to the student’s current level of knowledge and if students at a similar level derived significant benefit from this learning object. Let \( LOK[l, k] \) represent some learning object \( l \)’s target instruction level of knowledge \( k \), such that \( LOK[l, k] \in [0, 1] \) where 0 is complete ignorance and 1 is complete mastery within the course of study. For example, the target instruction level might be 0.68 for a 90-minute lab on recursion, since students have completed previous learning but are still gaining an understanding of nuances.

In order to simulate the derived benefit to a student, we model a concept known as impact which can be positive or negative\(^3\). Let \( I[l, k] \in \mathbb{R} \), represent the impact of learning from learning object \( l \) on the knowledge \( k \), that is, in the optimal case how much the learning object can adjust a student’s knowledge \( k \). The impact can be thought of as, for a student at the target level, what is the expected learning benefit of the object. This is information used by our approach to simulate the learning that is occurring.

Let \( UK[j, k] \) represent user \( j \)’s knowledge of \( k \in K \), such that \( UK[j, k] \in [0, 1] \). An example from computer science would be a knowledge of recursion recorded to be at 0.33. This would be interpreted as the student has an understanding of 33% of the course content dealing with recursion. The cardinality of \( K \) would be the number of dimensions making up the knowledge domain being simulated.

After an interaction with an object \( l \), a user \( j \)’s knowledge of \( k \) is changed by:

\[
\Delta UK[j, k] = \frac{I[l, k]}{1 + (UK[j, k] - LOK[l, k])^2}
\]
This has the implication that the impact of a lesson is at a maximum when the student’s knowledge level matches the target level of the learning object. As the two values differ, the impact of the lesson exponentially decreases.

Based on this change, the user’s knowledge in that area is updated as:

$$UK'[j,k] = UK[j,k] + \Delta UK[j,k]$$

The user’s average knowledge can then be calculated as:

$$\bar{UK}[j] = \frac{1}{|K|} \sum_{k=1}^{K} UK[j,k]$$

This perspective on learning is supported by Vygotsky’s ‘zone of proximal development (ZPD)’ as presented in Vygotsky and Cole (1978). The idea behind this educational theory is that the greatest learning occurs in the area between that which the student already understands and can do unaided and that which the student is incapable of doing. The more general pedagogical strategy of ‘scaffolding’ grew out of this theory. This takes the perspective that students need to be supported while learning, and as they achieve mastery the support can be gradually removed, like scaffolding from a building.

### 3.1 Simulation parameters

For this experiment, the parameter values we set are indicated in Table 2.

<table>
<thead>
<tr>
<th>Table 2 Parameters for simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target level of instruction</strong></td>
</tr>
<tr>
<td><strong>Impact values</strong></td>
</tr>
<tr>
<td><strong>Number of knowledge dimensions for each object/student</strong></td>
</tr>
<tr>
<td><strong>For each student, each knowledge dimension</strong></td>
</tr>
<tr>
<td><strong>Number of trials</strong></td>
</tr>
<tr>
<td><strong>Number of iterations per trial</strong></td>
</tr>
</tbody>
</table>

We sought to map, as our metric for effective performance of our algorithm, the mean of the average knowledges of all students after each trial. The simulation was run repeated (number of iterations), and the averaged results were then graphed. Each learning object reflects a different set of knowledges (target level of instruction and impact) to be conveyed to the student and each student arrives with a different level of understanding of each of these knowledge areas. Randomly generated students were distinct from each other, having a different multi-dimensional profile.

Two reference points were created to compare our approaches against. Random association associates each student with a randomly assigned annotation; greedy God chooses the best possible interaction for each student for each trial by pre-calculating the benefit of every annotation.
Three variations of Algorithm 1 were then run. The impact values and target levels of objects are used for the reasoning of the greedy god algorithm. In contrast, for the following three algorithms, our ecological approach is used to select the learning objects to be presented to users (so based on their similarity to previous students who have experienced these objects and on the benefit that these students derived). We assume that as each simulated student is assigned a learning object, that student’s interaction with the object can be used as the ‘previous experience’ to which subsequent students are matched. Using our ecological approach, learning objects presented to students should end up being ones that have an effective combination of impact value and target level (i.e., beneficial to those previous users and at a somewhat similar level of knowledge).4

Raw ecological has each student matched with the learning object best predicted to benefit her knowledge; pilot group has a subset of the students (10%) assigned, as a pilot group, systematically to learning objects – these interactions are used to reason about the best sequence for the remaining 90% of the students; simulated annealing provides a good approximation of the global maxima for a large search space. During the first 1/2 of the trials there is an inverse chance, based on the progress of the trials, that each student would be randomly associated with a lesson; otherwise, the ecological approach was applied. This approach provides a nice balance between exploration and exploitation.

3.2 Results

As seen in Figure 1 the random associations of students with learning objects is clearly and consistently shown to be an inferior approach5 compared to our algorithms for improving the average knowledge of a group of students, as expected. Similarly, an omniscient sequencer using perfect knowledge of students, learning objects and the outcome of a potential interaction (greedy god) can consistently produce the greatest learning benefit.

While Figure 1 seems to show superior performance of the ecological approach with a pilot group, it is important to remember that 10% of the class was used as a pilot group for this experimental condition. These were not included in the average assessed knowledge. Their increased knowledge, which would be roughly equivalent to the lack of increase shown by random associations, is omitted and the improved performance of the remaining students can be viewed as at the expense of the pilot group.

The ‘simulated annealing’ technique was interesting as it underperformed the other two techniques during its ‘cooling period’ but quickly gained ground after the cooling period was complete. This is due to the randomness added during the cooling period leading to a greater exploration of the possible interactions between learning objects and students. This improved understanding of the two groups could then be used when reasoning about which students to match with which learning objects in later trials. In the largest condition (50 students and 100 learning objects), simulated annealing matched the performance of the pilot group condition, without sacrificing 10% of the class. With the correct choice of cooling periods, this technique shows promise for delivering comparable long term performance at the expense of early progress for the entire group instead of no progress for a pilot group.
3.3 Error

Algorithm 1 proposes the selection of learning objects for students based on their similarity to peers and the benefits these peers have obtained from existing learning objects. Both similarity and benefit are determined in terms of the assessment levels of the students (obtained by pre- and post-test of each student, mapped to a level of a letter grade).

Since assessment in the real world is both imprecise and occasionally inaccurate, there is merit in exploring how well the algorithm would perform when validated in a simulated environment where the assessments include an element of error. Our interest is in how well the algorithm makes recommendations using this noisy data.

3.3.1 Approach

Our original hypothesis was that we expected errors in assessment, of the form of introduced noise, to degrade the learning curves observed. That is, we expected that as the noise increased, the slope of each learning curve embodying Algorithm 1 would decrease and they would gradually move away from the ideal greedy god curve and towards the random baseline. It was expected that rather than converging on perfect
knowledge (the 1 value on the y axis), the curves would converge on a lower value (that would drop ever lower as the noise increased). This would happen as the number of interactions between students and learning objects increased (modelled as trials in the x axis).

If the algorithm were robust in the face of error (that is, if the learning curves stay closer to the ideal case), this tells us that our approach can handle errors in assessment and continue to provide worthwhile recommendations to students, even when facing assessment errors. Poor performance, where the slope of the curves would drop quickly towards the random baseline, would tell us that this approach is highly dependent on good assessments (and would thus constrain the environment where it would be appropriate to use this approach).

In order to produce the error, we modified the assessment function in our simulation. Rather than mapping a knowledge level (continuous values in the range \([0, 1]\)) to a discrete level \(\{A^+, A, \ldots F, F^-\}\), we first added a random number, using a Gaussian distribution, with a mean of zero and a standard deviation of 0.05, 0.1, and 0.5. These experiments, taken as a whole, should provide us with an understanding of how increasing levels of noise in the assessment affects the effectiveness of curriculum sequencing performed by this approach. The greedy god and random baselines remained unchanged, since neither relied on assessment.

Figure 2  Comparison for varying standard deviation of error in assessment
3.3.2 Results

We did not see evidence of what we originally expected with the 3 graphs created with standard deviation (0.05, 0.1, 0.5). Instead, all 3 curves looked quite similar to the learning curves obtained using this approach on data without noise added to it, as presented in Section 3.2. In Figures 2(a) to 2(c), all three variations of the algorithm are performing well, in getting close to attaining the ideal average level of knowledge for students (i.e., the greedy god) by the end of the 200 trials. Note, as expected, simulated annealing takes longer to converge, as it is coping with random information at the beginning.

An initial concern was that our experiment might somehow be accidentally determining the appropriate learning objects without relying on the assessment. To test this concern, we replaced the assessment function with a function that randomly provides one of the 18 discrete levels (instead of an assessment, it provides a random grade). Since the three variations on this approach (ecological, ecological with pilot and simulated annealing) all rely on assessments to function, our expectation was that this change would produce three curves that were degraded to the performance of the random baseline. This is the result we saw (see Section 3.5).

In all, these results tell us that this approach is, in fact, highly robust with noisy data. As long as there is a tendency for an assessment to be closer to a correct value than an incorrect value, this approach will steadily improve the curriculum sequence suggestions as more data is obtained. Realistic amounts of noise, which are expected with any assessment, would seem to be acceptable to the functioning of this approach.

While our initial feeling was that a standard deviation of 0.5 was a large amount of noise, we then ran an experiment with a standard deviation of 1.5 [see Figure 2(d)]. The consequence of this is that we’re adding noise which is very likely to move data points anywhere in the range (with a standard deviation of 1.5, there is roughly a 25% chance of a perfect knowledge of 1 being mapped to a F- or, conversely of a complete absence of knowledge of 0 being mapped to an A+). With this massive amount of noise being added, we then saw the degradation we had initially expected (with the ecological condition converging on 0.8 instead of 1.0).

3.4 Variable time of instruction

In the previous approach outlined so far in this paper, which learning object should be assigned to a particular student is dependent on similarity of peers and the previous learning benefit obtained by those peers, alone. We explored a new extension, where we incorporate reasoning about the length of time it takes to complete an interaction with a learning object as well.

Clearly, in real learning situations, learning events can take variable amounts of time. Watching a recording of a lecture might take 76 minutes, while attending a day long seminar might take eight hours. Rather than making the simplifying assumption that each interaction with a learning object will take an equivalent length of time, we can incorporate this concept into our reasoning.
*calculateBenefit* in Algorithm 1 then needs to be modified to incorporate time. Rather than consider the benefit of the learning object, we can think of the proportionate benefit, that is, how much benefit it provides per minute of instruction (assuming a repository where each learning object’s average time to completion is recorded). This can be calculated by dividing the benefit of the learning object by the length of time it takes to complete the interaction for the average student.

We are interested in ensuring that, with this more sophisticated consideration incorporated, the approach outlined in Algorithm 1 continues to provide worthwhile recommendations for curriculum sequencing.

### 3.4.1 Approach

We modified the previous approach (Section 3 and 3.3.1) such that, as well as generating a random set of target instruction levels for each learning object, we also generated a random length of completion (ranging from 30 to 480 minutes). We used 50 students, 100 learning object and three runs – an error of 0.05, 0.1 and 0.5 standard deviation – each time for 20,000 minutes of simulated instruction. As well as the random and greedy god baselines, we again considered the raw ecological, ecological with pilot and simulated annealing variants. These results are displayed in Figures 3(a) to 3(c).

It is worthwhile to note that initially we experimented with about 2,400 minutes of instruction, based on this being roughly the amount of instruction in a typical university course. This was determined to be far too short a length of experiment as the learning curves reflected only the initial part of the graphs shown here. Our conclusion was that we were simply failing to see, yet, the benefit to learning that the students achieve and that either longer lesson times were needed or that it may be valuable to track students over multiple classes.

### 3.4.2 Results

With the increased time provided in Figure 3, we did indeed attain the kind of student learning that we expected (reasonably high average level of knowledge, for students). With the added sophistication of allowing learning objects to require different lengths of times to complete, this approach continues to make worthwhile curriculum recommendations to students. The fact that all three variants on the algorithm are approaching the ideal of the greedy god at the end of the trials is encouraging. As expected, the greater the standard deviation of error introduced, the more challenged each algorithm is to attain appropriate student knowledge levels, but the differences between Figures 3(a) to 3(c) are still relatively minor.

An unusual feature is that the learning curves approach a final knowledge less than 1, whereas in the experiments of Figure 1 they approached one. Initially this was thought to be a consequence of the introduced error; however, considering the curves from the error section above does not support this idea. It is possible that each approach is developing a bias towards short lessons, and is therefore not taking advantage of the full range of learning objects that may help the students approach complete mastery.
The software developed to run the simulations in this work eventually amounted to thousands of lines of source code. While the code was developed carefully and tested during development, bugs are inevitable in a codebase of this size. One of the techniques we performed to validate the source code during development was to run simulations that were different from what the code was written for, ensuring that the results matched the expectation.

One example of this was a simulation where the assessment portion of the source code was replaced with a method that would return a random grade instead. The expectation was that this would degrade the performance of all techniques to match the random baseline. We expected this since our techniques use assessment data to reason about student benefit and similarity to other students, and without this data it cannot provide reasonable recommendations.

The results obtained from this experiment are shown in Figure 4 and they match what was expected.
Figure 4  Effect of random assessment in place of student assessments

Figure 5  Error and time of instruction for a large corpus

3.6 Large corpus

It has been suggested (McCalla, 2004) that ecological approaches, rather than degrading with large amounts of information, improve. Intuitively this makes sense, with more data better recommendations should be possible. In order to investigate this, we considered a student group interacting with a large library of learning objects (5,000 objects). Collecting a massive amount of educational content offers a valuable resource, but also introduces the challenge of navigating a large corpus.

When we consider a simulation with dramatically more learning objects (Figure 5), we see that both the simulated annealing and the ecological with pilot learning curves become steeper. This corresponds with McCalla’s prediction. The ecological with pilot group has a sustained improvement and outperforms the raw ecological more dramatically than in previous experiments. Similarly, the simulated annealing conditions performs well with a larger library. The additional exploration of the corpus available to these algorithms in their initial phases appears to be providing some valuable benefit. Note that these curves approach the ideal average knowledge of the greedy god to a
greater extent with the larger repository (compared to the small repository used in Figure 3), which again provides support for McCalla’s hypothesis.

4 User study

In Section 3 we validated our models using simulations. To learn more about the effectiveness of our approach with real users we also conducted a evaluation with participants in an application domain enabling users to learn about how to care for a child with autism (which may arise as a home healthcare scenario, of interest to projects such as hSITE (Plant, 2008), with which we are involved).

4.1 Overview

Our first step was to assemble a repository of learning objects: the material that our users would learn from. In collaboration with a clinical psychologist specialising in children and autism, we created 20 learning objects (16 text articles and four videos) that each took about five minutes to experience. Also in collaboration with the psychologist we created a ten-question multiple choice assessment, covering material from the learning objects. This was used to carry out the pre- and post-test assessments which serve to model learning gains in students (and form a component of our algorithm for determining which objects to present to each student).

We hypothesised that a group of students using our peer-based technique for selecting learning objects (Algorithm 1) would show greater learning gains than a control group that had learning objects randomly assigned to them.

Each session lasted approximately one hour. 48 participants were involved in our experiment, including graduate students, undergraduate students and staff members at the university. All were at least 18 years old, fluent in English and not an expert in autism spectrum disorders.

4.2 Wizard of Oz

For our human study, we used what is known as a ‘Wizard of Oz’ approach. From Human-Computer Interaction (Kelley, 1983, 1984; Akers, 2006; Höysniemi et al., 2004). The idea behind this approach is that participants interact with a system that is at least partially controlled by a human.

In a typical ‘Wizard of Oz’ study the participants are led to believe that the human controlled elements are actually provided by software. In our study we were upfront about what was being done by the computer and what was being done by the experimenter. Deception was unnecessary, as we were investigating the efficacy of the educational material presented rather than the participants’ reaction to a fully-functional system. Instead of creating a complete computer-based educational environment, we wrote a program that would use the student assessments and our curriculum sequencing approach to recommend a learning object to the student. The recommended learning objects were then given (some as paper articles, others as videos played on a netbook) to the participants, while assessments were given as paper-and-pencil quizzes that were evaluated and entered into the system by the experimenter.
4.3 Procedure

The focus of our study was validating our proposed curriculum sequencing algorithm. We determined our repository by taking an extensive, sound set of articles and videos provided by our psychologist and distilling the information into what we felt would be five-minute lessons. These five-minute streamlined versions were shown again to the psychologist. She confirmed that they still presented sound information as well as sufficient information to answer the assessment questions developed.

In recruiting participants for this study we focused on obtaining participants who were students or staff at the University of Waterloo. The recruitment letter clearly indicated that we were excluding anyone with significant expertise in autism spectrum disorders. In fact, we eliminated two volunteers upon learning that they had received training to work with children with autism.

We decided to begin with just two participants in order to troubleshoot our experimental approach, proposed set of learning objects and assessment questions. We had in mind that these students would receive a random set of learning objects with each subsequent learning object not being influenced by their experience with the previous object. We had at that time a proposed assessment quiz that was administered before and after each learning object (the same assessment quiz each time). This quiz was a set of very general questions about autism that was not tied to the learning objects specifically. We observed that the two participants improved on their assessments, even though subsequent learning objects did not offer information that served to answer the questions on which they had improved. We learned that the participants were in fact improving because of the assessment quiz itself (using information from one question to help answer another question). From here we interacted with two psychologists experienced in research methods and concluded that it was important for us to ensure that our assessment quiz was much more directly tied to the learning objects that would be experienced. We then revised our assessment quiz and assembled a group of participants for our study, using this revised quiz for assessing these participants.

Each participant experienced five learning objects and was assessed before and after each for a total of six assessments. The assessments were the same ten multiple choice questions each time. This was done in part to ensure that we were modelling comparable learning experiences from the participants. Note that we considered, and rejected, the idea of using different assessments, counter-balanced using a Latin square, for two reasons. First, since our primary evaluation of each participant’s session would be their learning gain, it would be problematic if pre-test and post scores were based on different assessments. Secondly, using the same assessments allowed us to ensure that a larger amount of data points would be available for each learning object, of value for our peer-based technique.

In the end, our quiz was designed so that each question was covered well by different learning objects in the repository (and more than one learning object served to help a student to respond to that question).

We decided at the outset that we would use a control group, who experienced learning objects not using our techniques, and a treatment group, whose learning object sequence was determined by our techniques, in order to contrast the two groups. We hypothesised that the curriculum sequencing approach would lead to greater learning gains in the treatment group. The first 12 participants were randomly assigned learning objects during their session (control 1). They were used both as a control group and to
provide training data for our technique. The next 13 participants experienced a curriculum sequence provided by our approach, using only the data from the control group (treatment 1). Going beyond this initial pilot study, we then had 15 participants experience a curriculum sequence provided by our approach using all data that had been gathered up to that point in the experiment when they participated (treatment 2). Finally we had six participants experience a curriculum sequence suggested by a trained applied behaviour analysis worker who had ten years experience doing respite care with a young girl with autism (control 2). Since this expert did not have information about the specific students who would be using the system (beyond knowing that they were staff or students at The University of Waterloo), the curriculum targeted a more general audience, and was not personalised to specific students.

Participants read hard copy articles or watched videos on a provided netbook as a ‘Wizard of Oz’ style study was performed (see Section 4.2). For our technique, a program was written using the CLA (Algorithm 1) and the answers provided by participants in their pre-test assessments served as the current student assessment; a new recommendation for a learning object was then determined. This sequence continued until the student had experienced five different learning objects. In essence, the first 12 participants served to prime the system for the remaining participants. After this phase, each learning object in the repository had three experiences recorded: while the initial control group of students were shown a random set of objects, which objects would be presented to each was determined off-line in a way that ensured that each object would be shown to three different participants. The net-benefit obtained by each subject in the control group (number of questions correct between pre and post-test) became part of that object’s interaction history. As a result, each learning object in the repository acquired an interaction history with exactly three entries. This then ensured that the students in the treatment group were experiencing a sufficiently rich collaborative learning algorithm.

For the participants in our experimental group, determining the similarity between the current student and previous peers was measured by comparing the number of questions on the assessment that were answered identically. Only the data collected from the training group was used to make recommendations to the ‘treatment 1’ group. Following our proposed approach and continually adding data for the program to make recommendations, which we did for ‘treatment 2’, caused the participants to be assigned learning objects based on an evolving repository of data. However, this also caused this experimental group to fail to be provided with a consistent treatment. This can be thought of as many separate experimental conditions, each with a single data point, which makes statistical analysis difficult. Our ‘control 2’ group used an expertly divided curriculum instead of random assignment (as was done for ‘control 1’).

To act as ‘wizard’, we took the first assessment quiz completed by the subject and entered their answers to each question into a program on the netbook. This produced a list of 5 learning objects with the highest predicted benefit (according to the CLA) in order. The most preferred learning object that the student had not previous seen was either handed to the student as a print out, or the video was loaded on the netbook and shown to the participant. Note that we made a decision not to show the same learning object twice; this was reasonable, given the very brief length of the learning objects and of the total instructional experience. This same procedure was repeated after each student had completed an assessment.

For our human study there was a challenge in applying the multi-dimensional model of knowledge. Unlike the case of our simulations, where we could directly model and
measure the simulated students’ understanding of various knowledges, with human students we only had the answers to quiz questions to use for our calculation of student similarity. The domain of instruction, care of children on the autism spectrum, was inherently complex and determining what abilities make up the overall knowledge of the course of study and categorising assessment questions according to this ontology would have been challenging.

Our simplifying assumption was that each question on the quiz represented a dimension in the model of knowledge for the students. This gave us a ten-dimensional model (one for each question), and the similarity of students could be matched based on how they answered each question, regardless of whether that answer was correct or not. For example two students who both incorrectly answered ‘speech therapy’ to question 9 would be more similar than a student who answered physical therapy (which was also incorrect).

It was possible for us to do this since there was only a single quiz for assessment for the entire experiment. Had we used multiple quizzes, it would not have been possible to compare the results in this manner (they would be in different vector spaces). In this case there would have been a need to map the assessments to a global knowledge, which could then have been compared.

4.3.1 Results

We first considered the learning gains of our 11 experimental group participants, namely the post-test (their final assessment) minus the pre-test (their first assessment).

These results can be interpreted that, on average, participants in the control 1 group got 1.83 more questions correct (out of ten) after completing the five learning objects. Participants in the ‘treatment 1’ group got an average of 3.08 more questions correct (see Table 3).

The results were statistically reliable at $p < 0.05$ (one-sided, two samples, unequal variance t-test) when contrasting either of the treatments to ‘control 1’. ‘Treatment 2’ was significantly different from ‘control 2’, however ‘treatment 1’ was not.15

<table>
<thead>
<tr>
<th>Mean</th>
<th>S.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control 1</td>
<td>1.83</td>
</tr>
<tr>
<td>Treatment 1</td>
<td>3.08</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>3.33</td>
</tr>
<tr>
<td>Control 2</td>
<td>1.50</td>
</tr>
</tbody>
</table>

Next, we compared the proportional learning gains of participants (see Table 4). This was to take into consideration the suggestion of (Jackson and Graesser, 2007) that simple learning gains are ‘biased towards students with low pretest scores because they have more room for improvement’. This is measured using \( \frac{\text{posttest} - \text{pretest}}{\text{maximum} - \text{pretest}} \). While this is a less intuitive measure than simple learning gains, it can be thought of as the initially unknown portion of the testable material that the student learned, expressed as a real number, with a maximum of 1.0 (indicating mastery of the material). A mean value of .610 for ‘treatment 2’ means that students learned 61% of material that had been initially unknown, whereas for ‘control 1’ only 31.6% of that material was learned.
Table 4  Comparison of proportional overall learning gains of users in control and experimental groups

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>S.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control 1</td>
<td>0.316</td>
<td>0.214</td>
</tr>
<tr>
<td>Treatment 1</td>
<td>0.498</td>
<td>0.281</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>0.610</td>
<td>0.307</td>
</tr>
<tr>
<td>Control 2</td>
<td>0.195</td>
<td>0.355</td>
</tr>
</tbody>
</table>

The results were statistically reliable at $p < 0.05$ (one-sided, two samples, unequal variance t-test) between ‘control 1’ and both the treatment groups, and between ‘treatment 2’ and ‘control 2’. The difference between ‘treatment 1’ and ‘control 2’ was not statistically significant.16

In Figure 6 we plot the normalised learning gains of each participant against the amount of data from past participants that this experimental condition had access to. That is, the ‘control 1’ group had randomly assigned learning objects, which is equivalent to not having access to any data, which gave all 12 participants 0 data from past participants. The ‘Treatment 1’ group had recommendations based exclusively on the data from the ‘control 1’ group, so each of these 13 participants received recommendations based on data from 12 past participants. The ‘Treatment 2’ group had full access to all past data, including past participants in the ‘Treatment 2’ group, and are represented by the amount of data on the x-axis which begins at 25 and goes up to 39. The ‘control 2’ group are not represented, because an alternative approach (namely an expertly crafted curriculum), was used to make these recommendations.

Figure 6  Linear regression of normalised learning gains (see online version for colours)

The linear regression of this data allows us to predict the normalised learning gains from amount of data from past students by the following formula: 

$$NLG = 0.0084 \times Data_{fromPastParticipants} + 0.35; R^2 = 0.1559.$$ The standard error of the
slope was 0.003. Given that a slope of 0 is not within the standard error of the slope of the linear regression, this is a statistically significant result.

Taken together, our results indicate that students presented with learning objects determined by our algorithm achieved greater learning gains than those who were randomly assigned objects. Additionally, in the situation where our approach has access to larger amounts of data, our methods were shown to achieve greater learning gains than a curriculum determined by a subject expert.

5 Conclusions and discussion

This paper presents an approach for designing intelligent peer-based tutoring systems that reflect an ecological approach, in contexts where there is a repository of learning objects. The proposed algorithm and its accompanying modelling of student learning is described in detail and is validated as effective, through a combination of simulation and human studies. Our techniques provide a straightforward, inexpensive approach to making any computer-based educational system personalisable and adaptive. The model is applicable for a wide range of possible applications; an illustration for the domain of health has been provided, offering an opportunity for patients and caregivers to learn how to manage their illnesses through a personalised presentation of the most effective learning materials. Throughout, our focus is on curriculum sequencing and content selection and how best to resolve that for students. Below we highlight some of the novel features of our approach and emphasise its inherent value, in contrast with the work of other researchers.

5.1 Our use of simulations

The model presented in this paper offers a methodology for designing intelligent tutoring systems that propose learning objects for students to experience, based on benefits in learning achieved by similar peers, in the past. The first approach for validating our central curriculum sequencing uses a novel concept of simulated student learning which then enables a quantification of knowledge gains achieved by students, using our methods, against the backdrop of a considerable, varied complement of peers, over an extended period of time, with latitude for some error in the assessments used to drive the labelling of student achievement levels (and thus the representation of their knowledge).

The results that emerge from our simulations lead to certain important conclusions, as follows. From Figure 2 we see that, compared to random selection, the knowledge gains achieved by the three variations of our central algorithm are dramatically better, even over a short number of trials. This serves to validate that it is indeed possible to effectively operationalise McCalla’s ecological approach (as enshrined in our Collaborative Learning Algorithm). In addition, McCalla’s (2004) prediction that a greater number of learners will lead to greater learning gains is upheld and definitively quantified. In summary, one central contribution of our efforts is to confirm the inherent value of ecological content sequencing, as the basis for intelligent tutoring of students.

Since our solution rests on the use of assessments of students in order to determine peer similarity and benefits delivered by learning objects, we included experiments that allowed for some error. The results of the successful outcomes in Figure 3 are, in fact, a reflection of the robustness of our proposed method and our strategy for simulating the learning.
If an error in assessment leads to an inappropriate learning object being proposed for a new student, this should result in a poor increase in knowledge. When the new student’s assessment is modelled, it will then be attached as one of the experiences with that learning object. The average of these assessments will approach the true value, even if some of those assessments are distorted by noise. As this happens, the system is less likely to recommend the bad object, and will increasingly direct students to a better choice. As a result, this learning object would now be less likely to be assigned to students in the future and the system self-corrects.

The other element integrated into our simulations is time. As with other decisions in life, time a student devotes to one activity is time that can not be devoted to other activities (opportunity cost). Making the most efficient use of the available time is clearly a worthwhile goal, and we expanded the sophistication of the simulations of learning objects to allow a variable time of instruction. This allowed interesting tradeoffs, such as whether to interact with a learning object that would give a greater learning gain or another object that could be completed sooner. This can be thought of as the proportionate benefit or the learning per minute. Simulations with the error and variable time incorporated in the simulation can be seen in Figure 4 and the fact that our algorithm is robust to variations in time of lesson is encouraging.

Other researchers, such as VanLehn et al. (1994) and Matsuda et al. (2007), have investigated the use of simulated students. They were primarily interested in using simulated student to model human students or to allow a teacher to practice. Our system is different from these, in that our simulated students are used entirely to evaluate the efficacy of our techniques, and are not used as peers for humans or to predict their actions.

Perhaps most importantly, using a simulation to validate an ITS fills a much needed gap that is left when the value of these systems is confirmed through studies with human users alone. This is because:

- It is possible to simulate the learning achieved by a large number of students (whereas human studies are challenged to incorporate a very large sample size of participants).
- It is possible to reflect the learning that would be achieved when a very large repository of objects has been experienced by previous peers (which is also a challenge when relying solely on the use of a human study).

In particular, our large corpus experiment (Figure 5) would have been challenging to assemble and provide to human students in a way that enables a wide range of the repository to be experienced. With a simulation, large corpus sizes can be created by changing a parameter. We do multiple runs with the 50 students that we inject into our simulation (for all three variants) which serves to be modelling an extensive learning experience, for each student. These experiments would be comparable to running an experiment for 50 students, in 20 iterations, for each of the five conditions: the equivalent of running a study with 5,000 participants. With a simulated annealing approach, in particular, we are able to ensure that a wide range of objects are introduced to the simulation, exploring the full repository.

Looking at systems presented at ITS2010 for instance, we note however that the number of students used in the ten human studies presented was approximately 50
participants on average, which is considerable lower than we are able to achieve with our simulations.

We also note, if our algorithm for intelligent tutoring were to be validated on the basis of a human study alone, we would be challenged in bootstrapping the system (assembling a record of peer experiences which form the basis of the future learning of the new students). We did, in fact, conduct a user study as an additional source of validation for the CLA (in Section 4), however, and the results were encouraging. We ended up training the system with an initial set of 12 students, using only this data to make recommendations for an additional 13 participants and found the improvements in learning for the students noticeably better than competing algorithms for content selection. In order to assess our approach with evolving student data, we then introduced a second experimental condition, involving 15 participants where each participant’s recommendations were based on the data from all previous students, including previous students in the second experimental condition. Again, the results from this group demonstrated important improvements in learning gains.

In sum, our validation efforts can be viewed in the following light. Simulations and models are necessarily an approximation of reality. Despite any enticing benefits they offer, their fidelity to real world events must be supported. Typically this is done by using the simulation or model to make predictions, then measuring the difference between that prediction and the actual events. To this end, we examined different conditions for real world students and contrast the efficacy of our approach during these experiments to those predicted by our simulations. While our techniques are intended for large repositories with massive numbers of students, a clear effect was shown, even with small scale human trials.

5.2 Contrast with other intelligent tutoring systems

In contrast to efforts for designing peer-based intelligent tutoring systems such as Cheng and Vassileva (2006) focused on enabling real-time engagement of peers, in our approach each student’s learning is directed by considering all experiences of previous students, thus allowing for a continuous redirection of possible content. Personalisation is maintained throughout, as well. This is achieved by modelling the knowledge levels of each student and an assessment of their current overall understanding in order to perform matching to like-minded peers, for the selection of learning objects.

Previous work on collaborative learning (Harrer et al., 2006; Hausmann et al., 2008) has attempted to use interactions between students and the system to provide a better experience for subsequent students (as we do). Harrer et al. (2006) created a program that would capture user problem solving behaviours in the system and use this data to begin the development of a tutor. In contrast, our approach does not try to model specific user actions. Instead it pragmatically considers the sequence that learning material is experienced and how successful the students were. Hausmann et al. (2008) examined dialogs between two students and their joint understanding of a dialog with the ITS (hence the ‘Trialog’). This was shown to both enhance student understanding and reduce off-task behaviour. Such a technique would be compatible with our approach and two students using a system together may have a more nuanced, productive interaction with the system: two students working together on a task could simply be considered a learning object that is recommended like any other.
5.3 Specific features of our model and its validation

Our system, which assigns learning objects based on assessments, does not give recommendations based on the validity of educational content of learning objects and, in fact, does not perform any reasoning about this. Instead, recommendations are pragmatically made based on what has helped a student learn. Whatever pedagogical approach or specific learning objects facilitate this for a student will be more likely to be shown to similar students in the future. Therefore, this deflects the need to explicitly screen every learning object in the repository in order to remove the ones with weaker educational value. The system itself will identify the stronger and weaker learning objects.

Our approach for simulated student learning can also be viewed as a specific proposal for representing student knowledge within user models that are employed within intelligent tutoring systems. Here, we offer a multi-dimensional distinction for each student, tracking knowledge levels in a set of different required major concepts. This aligns well with current user modelling efforts that promote multi-dimensionality in user modelling (Vassileva et al., 2003).

Legaspi et al. (2004) has provided approaches to allow self-improving instructional planning. The authors created an agent that uses reinforcement learning to automatically derive categories and allows self-improvement of the system by using these categories to revise existing (or create new) instructional plans. Their approach requires extensive, explicit information about the learning objects (such as lesson goals and topic of instruction), students (such as cognitive ability, learning style, knowledge scope and lists of errors committed) and instructional plans (and, if they were ineffective, changes that could make them effective). Our approach, using data that will already be gathered in instructional contexts and not requiring creation of extensive metadata, can be viewed as faster, easier and cheaper to add to an existing system.

We note as well that while our work may seem to over-focus on performance, in actuality this is not the case. Our system does evaluate the learning gains from a student’s interaction with a learning object, but the specifics of this interaction are completely abstract. A learning object could be a 35-minute interaction with an ITS, and part of the learning gains made could be from the student experiencing confusion. If this is the case, our system will record the resulting improved learning and recommend that learning object in preference to objects without such effective pedagogical approaches. Our approach is agnostic towards the specific pedagogy and pragmatically focuses on the long-term learning of students and their similarities to other students.

5.4 Future work

One direction for future work is to explore different variations for the simulations and the human study. For the simulations, we are interested in conducting additional experimentation to determine the ideal choice between ecological, ecological with pilot and simulated annealing approaches (which we suspect will be domain specific); exploring metrics other than mean average knowledge, for example clustering students into subpopulations and tracking separately the knowledge gains of each; experimenting with greater or lesser penalties for being assigned a learning object with an inappropriate target level of instruction during the simulation. For the human study, the obvious next
step is to enlist actual caregivers of autistic children as our participants for greater insight into the potential for patient-led healthcare.

Other future work involves adjusting the model at the centre of our proposed solution. One direction is to explore a different metric for learning gains than is currently used in the Collaborative Learning Algorithm. Historically, learning gain has been the accepted metric for measuring a learning event for ITS researchers. One alternative which is being considered is to use the proportional learning gains (Jackson and Graesser, 2007)\(^7\) which is defined as: \[
\frac{\text{post-test} - \text{pretest}}{\text{pretest}}
\] This would be a useful alternative for avoiding a bias towards interactions where a student has a low initial score, if this formula were used instead on the right hand side of the equation in line 8 of Algorithm 1. For example, with this metric, advancing from A to A+ is a greater gain than advancing from B to B+. We have begun an initial exploration of the use of this metric in our human study, discussed in Section 4.

A series of additional extensions to the model can also be explored. It may be worthwhile to consider recommending learning objects as a pair in addition to individually, for situations where two learning objects complement one another and have pedagogical synergies. It may also be useful to reason about sequences of learning objects that may reinforce one another and give a better experience as a set, rather than selecting individual learning objects. In addition, our work has used approaches that precisely measure the similarity between two students when making recommendations, and in the future we are interested in examining clustering techniques (e.g., K-means, Y-means, COP-networks) as well as other measures of similarity beyond academic performance. A final extension would be to introduce a garbage collection algorithm that could be used to consider the set of all learning objects in the corpus, reasoning about which learning objects should be retained and which add little instructional value to any of the students who may use the system.

Final efforts for future work would involve integrating entirely different components proposed by other researchers. Cheng and Vassileva (2006) investigated recommendation of academic papers and motivating the participation of users in a small-scale, online community. Their approach has been useful to inform, develop and evaluate our system which has a broader focus. Their work on incentivising participation will be useful during larger scale human studies. In addition, working with a more elaborate system for student modelling is a more forward looking plan for future work. Kay (2008) considers how to approach building a user model to support universal, personalised lifelong learning. This works well with our techniques and a combined system would be a large contribution.

References


Notes

1. Adapted from Breese et al.’s (1998) collaborative filtering paper.
2. The negative impact was introduced to simulate the possibility of misinformation from a poor quality learning object or a learning object that does a good job teaching one concept, while undermining the understanding of another concept.
3. This can be thought of as the ‘Goldilocks’ zone because it is ‘just right’ for the student.
4. We assume that a learning object can be experienced more than once by the same student.
5. Because there was an even mixture of learning objects which improve or degrade the students’ knowledge, when learning objects are randomly assigned the overall average knowledge tends to stay the same, leading to the flat curve seen in the graphs.
6. Even with an ideal assessment tool, there will still be situations where students mistakenly give incorrect information that they understand (known as a slip) or accidentally give the right answer to something they do not understand (known as a guess).
7. It is important to note that there are different approaches to model a ‘bad assessment’. By randomly adding noise, we are modelling an assessment that has variability in every assessment. This does not model an assessment with a systemic bias, for example, one that always evaluates C+ students as D students.
8. The idea is that if a student could be modelled with an erroneous assessment level (e.g., B vs. A) then with greater standard deviation, the likelihood of an erroneous label increases. Note that values closer to the true value will still be the most likely to be assigned.
9. Although we use the units minutes, these are best thought of as unspecified time units, since no effort was spent trying to calibrate whether the educational gains were appropriate for the length of instruction.
10. Participants were not provided with the results of their assessment.
11. After creating a rough cut of the lesson, the first author timed himself reading through them and looked at the timestamps on the video to verify that they took approximately five minutes each.
12. The first assessment, before the participants have experienced any learning objects, measures their initial knowledge about the subject at the beginning of the session.
13. For the 12 participants, three random lists of all integers between 1 and 20 were created. The first participant saw the first 5 entries on the first list, the second participants saw entries 6–10 on the first, etc., with the 12th participant seeing the last five entries on the third list. This ensured that every student was assigned five distinct random learning objects and that each learning object had data from three separate students.
14. For our purposes, we considered this to be a single group. We provide a linear regression (see Section 4.3.1) which shows a statistically significant improvement in learning gains as data accumulates on learning objects. Furthermore, our ‘treatment 1’ condition also shows a statistically significant improvement, while maintaining a consistent experimental condition.
15. This may be due in part to the small number of participants in ‘control 2’. We note that the difference between ‘control 1’ and ‘control 2’ was also not statistically significant.
16. As with our learning gains experiment, this may well have been influenced by the small number of participants for ‘control 2’. It was also the case that the ‘control 1’ and ‘control 2’ difference was not statistically significant.
17. This is distinct from the proportionate benefit detailed above.