

Does difficulty even matter? Investigating difficulty adjustment and practice behavior in an open-ended learning task

**Anan Schütt, Tobias Huber, Jauwairia Nasir, Cristina Conati,
Elisabeth André**

Angaben zur Veröffentlichung / Publication details:

Schütt, Anan, Tobias Huber, Jauwairia Nasir, Cristina Conati, and Elisabeth André. 2024. "Does difficulty even matter? Investigating difficulty adjustment and practice behavior in an open-ended learning task." In LAK '24: proceedings of the 14th Learning Analytics and Knowledge Conference, Kyoto, Japan, 18-22 March 2024, 253-62. New York, NY: Association for Computing Machinery (ACM).
<https://doi.org/10.1145/3636555.3636876>.

Nutzungsbedingungen / Terms of use:

licsonst

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under the conditions:

Sonstige Open-Access-Lizenz

Weitere Informationen finden Sie unter: / For more information see:
https://www.bibliothek.uni-augsburg.de/opus/lic_sonst.html





 Latest updates: <https://dl.acm.org/doi/10.1145/3636555.3636876>

RESEARCH-ARTICLE

Does Difficulty even Matter? Investigating Difficulty Adjustment and Practice Behavior in an Open-Ended Learning Task

ANAN SCHÜTT, University of Augsburg, Augsburg, Bayern, Germany

TOBIAS HUBER, University of Augsburg, Augsburg, Bayern, Germany

JAUWAIRIA NASIR, University of Augsburg, Augsburg, Bayern, Germany

C. CONATI, The University of British Columbia, Vancouver, BC, Canada

ELISABETH ANDRÈ, University of Augsburg, Augsburg, Bayern, Germany

Open Access Support provided by:

The University of British Columbia

University of Augsburg



PDF Download
3636555.3636876.pdf
18 December 2025
Total Citations: 2
Total Downloads: 747

Published: 18 March 2024

[Citation in BibTeX format](#)

LAK '24: The 14th Learning Analytics and Knowledge Conference

March 18 - 22, 2024

Kyoto, Japan

Does Difficulty even Matter? Investigating Difficulty Adjustment and Practice Behavior in an Open-Ended Learning Task

Anan Schütt

anan.schuett@uni-a.de
University of Augsburg
Augsburg, Germany

Tobias Huber

tobias.huber@informatik.uni-augsburg.de
University of Augsburg
Augsburg, Germany

Jauwairia Nasir

jauwairia.nasir@uni-a.de
University of Augsburg
Augsburg, Germany

Cristina Conati

conati@cs.ubc.ca
University of British Columbia
Vancouver, British Columbia, Canada

Elisabeth André

andre@informatik.uni-augsburg.de
University of Augsburg
Augsburg, Germany

ABSTRACT

Difficulty adjustment in practice exercises has been shown to be beneficial for learning. However, previous research has mostly investigated close-ended tasks, which do not offer the students multiple ways to reach a valid solution. Contrary to this, in order to learn in an open-ended learning task, students need to effectively explore the solution space as there are multiple ways to reach a solution. For this reason, the effects of difficulty adjustment could be different for open-ended tasks. To investigate this, as our first contribution, we compare different methods of difficulty adjustment in a user study conducted with 86 participants. Furthermore, as the practice behavior of the students is expected to influence how well the students learn, we additionally look at their practice behavior as a post-hoc analysis. Therefore, as a second contribution, we identify different types of practice behavior and how they link to students' learning outcomes and subjective evaluation measures as well as explore the influence the difficulty adjustment methods have on the practice behaviors. Our results suggest the usefulness of taking into account the practice behavior in addition to only using the practice performance to inform adaptive intervention and difficulty adjustment methods.

CCS CONCEPTS

- Human-centered computing → User studies; User models;
- Applied computing → Computer-assisted instruction; Interactive learning environments.

KEYWORDS

Difficulty Adjustment, Adaptive Practice, Clustering, Educational Data Mining

ACM Reference Format:

Anan Schütt, Tobias Huber, Jauwairia Nasir, Cristina Conati, and Elisabeth André. 2024. Does Difficulty even Matter? Investigating Difficulty

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

LAK '24, March 18–22, 2024, Kyoto, Japan

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1618-8/24/03.

<https://doi.org/10.1145/3636555.3636876>

Adjustment and Practice Behavior in an Open-Ended Learning Task. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3636555.3636876>

1 INTRODUCTION

Practice exercises are an important part of learning [14]. They allow students to apply the information they learned to problems, and thus better retain what they learned [35]. Since students who work through practice exercises have their own prior knowledge and learn at different rates, it makes sense to adapt the practice exercises to suit them. In this regard, previous research has shown that choosing the appropriate difficulty level of practice exercises has positive effects on learning gains [38] and learning experience [5, 23]. Additionally, when looking through the lens of the flow theory [6], exercises that are too difficult for the student could cause anxiety, while exercises that are too easy could cause boredom. Therefore, avoiding both of these is important for reaching the state of flow, which has been shown to improve learning outcomes in different learning scenarios [10, 34].

To this end, there are many approaches to adjust the difficulty level of practice exercises. One is to allow the students to choose what difficulty they will get next, giving them autonomy. This is supported by research that shows that including the learner's choice in a learning activity improves learning gain [37], mainly in sports [3, 45]. Another approach is to estimate the student's ability level from their previous performance, and then automatically serve them with a corresponding exercise. In the context of computer-based education, this process of automatically adjusting the difficulty level of exercises in this way, has been called by different keywords, such as computer adaptive practice [22, 33], adaptive curriculum [2], personalized task difficulty [46], or Dynamic Difficulty Adjustment (DDA) [17]. In this paper, we refer to the automated and adaptive difficulty adjustment as Dynamic Difficulty Adjustment (DDA) due to it being self-explanatory.

Although both self-determined difficulty and DDA have been applied in different domains, to the best of our knowledge, studies directly comparing these two approaches in the same application are rare, with [37] being the only example we found. Moreover, we are interested in difficulty adjustment in an open-ended task, which is defined as a task with multiple ways to complete, and potentially

multiple solutions. In this work, we evaluate the two approaches of difficulty adjustment in an open-ended reasoning task, more specifically, a graph theory task. The choice of the task is inspired by the fact that there is an increasing emphasis on integrating more constructivist and open-ended learning activities into the curriculum to allow for a possibility for the students to become aware of the knowledge construction process through exploration and exploitation of the environment complemented by reflection [44]. The exploratory nature of such tasks makes it non-trivial to gauge students' ability; hence, making it more challenging to adapt the difficulty. On the other hand, an open-ended task requires multiple actions from the student to solve, and so these actions can be traced with other techniques to more deeply assess the student. To our knowledge, adapting difficulty in such open-ended reasoning tasks is not a widely explored topic in the literature yet, the closest research that comes to our work is [15, 37], however, there are important differences between the proposed work and the previous works that we point out in Section 2.1. Thus, as a first contribution in this paper: *we extend the research in this direction by evaluating the two aforementioned difficulty adjustment approaches in an open-ended reasoning task to provide new insights into this problem area.*

In an open-ended task, students perform multiple fine-grained actions to complete a single exercise. While evaluating the submission of the exercise indicates how well a student does, the fine-grained actions a student performs, such as the string of click actions in a computer application, is also a valuable piece of information. By grouping students by how they interact with the learning system, one can gain an insight into which behavior is helpful for learning [18, 41]. In the same sense, to better understand the students' behavior that links to learning and to investigate the relationship between the method of difficulty adjustment and the resulting behavior type, we further analyze the fine-grained data as a post-hoc analysis. We discover the behavior types from clickstreams using clustering, following [4, 24]. Ultimately, as a second contribution in this paper, *we identify different student practice behaviors that link to learning differently in an open-ended reasoning task in an effort to tie those behavioral differences back to difficulty adjustment.*

Overall, we are interested in three research questions.

- **RQ1:** How do the different difficulty adjustment methods affect the learning outcomes and the subjective experience of the students in an open-ended learning task?
- **RQ2:** What types of students' practice behaviors exist in such a task and how do they relate to the learning outcomes?
- **RQ3:** How does the method of difficulty adjustment influence the students' practice behaviors?

2 RELATED WORKS

2.1 Difficulty Adjustment in Learning

Previous works have used difficulty adjustment mechanisms for educational practice exercises. We divide the existing methods of difficulty adjustment into two main categories: self-determined and Dynamic Difficulty Adjustment (DDA). The self-determined approach refers to allowing the student to select the difficulty level of his exercises. The reasoning comes from the Self-determination theory [8], which lists autonomy as a basic requirement of human

psychology, and also an important component in learning [30]. Hughes et al. [16] studied how a student chooses difficulty levels in a practice session of a first-person shooter and how the difficulty level influences the post-test performance, but didn't directly address the learning gains. Chiviacowsky et al. [3] showed the effectiveness of the self-determined approach in a motor-learning task, comparing between a self-determined condition and a yoked condition. A participant in the yoked condition doesn't get to choose the difficulty level, but instead gets the choices that a paired participant from the self-determined condition. Still, the work addressed a close-ended task, where there is no flexibility for exploratively using different methods to complete the task.

The other category is DDA. Here, the difficulty level of a practice exercise depends on the performance on previous exercises. Students who previously performed better get more challenging exercises, while students who performed worse get easier exercises. Romero et al. [36] implemented a DDA for a cardiac life support simulator, showing that adaptive training helps students learn the same content in less time, improving efficiency. Sampayo-Vargas et al. [38] showed improved learning gains through DDA in a Spanish vocabulary practice application. Other works have shown increased engagement with the practice exercises, which then lead to improved learning gains. Klinkenberg et al. [22] showed increased engagement and learning gains in basic arithmetic exercises, and Pelanek et al. [33] in geography facts practice. However, these tasks are also close-ended in nature, making the way a student interacts with the learning application relatively fixed. There are also works that explored open-ended tasks. Hooshyar et al. [15] applied DDA in a block-based programming learning scenario. Still, they compared an adaptive gamified learning session against a lecture of the same content, and did not compare it against a non-adaptive version of the same game.

The work that was closest to our scenario was from Salden et al. [37]. They studied two difficulty adjustment methods in an open learning scenario, namely an air traffic control task. They had a condition with self-determination difficulty level and a DDA condition. They found an increased learning gain between each condition and their corresponding yoked conditions, but no significant differences between the two difficulty adjustment conditions. We extend this work by comparing against a more neutral baseline, and additionally studying the behavior of the students during practice.

2.2 Effects of Practice Behavior on Learning

Previous works have shown that how a student interacts with the learning application (which we henceforth call the practice behavior) affects how well a student will learn. Käser and Schwartz [21] explored students' practice behavior in an open-ended educational game, where the student has to find out the algebraic rules governing the results of a tug-of-war game. Students can input different tug-of-war configurations, which the game will simulate and output the winner. There are differences in which configurations students try out, as some more logically reveal more information about the rules, and some seem more like unprincipled trial-and-error. What configuration a student chooses to simulate predicts how well they will learn. Other works in different domains have also shown that practice behavior predicts the learning outcome,

including in programming [9], video learning [31, 41], and English [7]. As a further step in application, learning about the efficacy of students' behavior types can help derive a pedagogical policy to help students learn better, by nudging the student's behavior to be more similar to the high-learning counterpart [20, 24]. We extend the literature addressing direct behavior change by studying the changes in practice behavior through another mechanism, which is difficulty adjustment in our case.

To extract the behavior types, previous works have proposed clustering methodologies. These types can then be analyzed and linked to the learning outcomes. In an effort to gauge productive engagement, Nasir et al. [28] proposed a forward-backward clustering approach that reveals the link between higher learning gains and multiple behavior types in an open-ended learning environment underlying a reasoning task. Fratamico et al. [11] also used clustering to find behavior types in an electronic simulator, with different types indicating differences in learning outcome. Additionally, they used association rule mining [1] to find out the defining features of the behavior types, creating a set of explanations that can inform instructors. Taking inspiration from these works, we use clustering on our dataset to discover types of practice behavior.

3 STUDY DESIGN

3.1 Research Question

We originally designed the user study to answer our first research question about the effect of different methods for selecting exercise difficulty on the learning gain and on the affective state of the participants. We preregistered this study online¹. Additionally, we use the data from this study to investigate our second and third research questions about what practice behaviors are indicative of learning and how this ties with the difficulty selection method. Due to this, we made some minor modifications in some of the evaluation details from the initial preregistration, in particular how we filter the participants.

3.2 Task

The task in our study is the maximum independent set (MIS) learning task, which we previously proposed in Schütt et al. [40]. The rules of this task are easy to explain, but can lead to a complex exercise to solve, given a more complicated graph. The maximum independent set of a graph G is the largest set of vertices V_{MIS} , such that no two vertices $u, v \in V_{MIS}$ are adjacent. In the graphical interface of the study (Fig. 1), a graph is shown with all the vertices in black, denoting that they are not selected. The participants' task is to find and select a set of vertices that form a valid MIS for the graph, and then submit their selection. When the participant clicks on an unselected vertex, we consider this action as adding that vertex to the selection set. The participant can also unselect a vertex from the selection set by clicking on a selected vertex. The last action is reset (*clear* in the figure), which unselects every vertex, emptying the selection set. It must be noted that there are several

¹<https://aspredicted.org/cg92w.pdf> for comparing the *pref* condition against the *self-det* condition.

<https://aspredicted.org/i6zm7.pdf> for comparing *pref* and *self-det* conditions against the *DDA* condition.

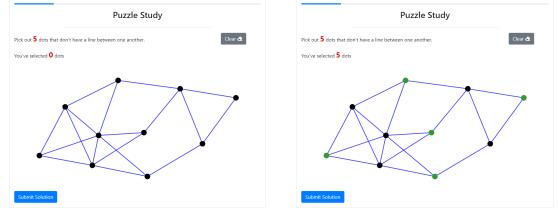


Figure 1: Main interface of the user study. The screen shows the graph, the buttons for submission and reset, the number of vertices to choose, and the number of vertices the participant already chose. The left figure shows the screen upon arriving on the page, with no vertex selected. The right figure shows a valid solution to the graph.



Figure 2: Stages of the user study.

sequences of actions that reach the same MIS, as well as there can be multiple MIS solutions for the same problem.

The interface also displays the number of currently selected vertices, the size of the correct MIS, and a button for submitting the current selection of vertices. Note here that the button can be clicked at any time, so the participant could submit an incorrect answer. The participant thus has to manually check if he has a correct solution before submitting. The participant is told that there is a time limit, but they do not know how long it is (90 seconds). This is to reduce the sensation of time, to not be in conflict with flow. To indicate the time, a red flag is shown 5 seconds before the time runs out. If the time runs out without a submission, it is counted as an incorrect solution. After the submission, there is a pop-up saying whether the solution was correct or incorrect.

Each practice exercise and each exercise in the pre- and post-test is associated with a difficulty level, represented by a real number. This value is important, as it is the main point of adjustment. The difficulty level of an exercise is calculated from the attributes of the graph, such as the number of edges and the size of the MIS. The exact method of calculation is described in the appendix of [40]. This difficulty value is used to determine the test exercises, and used for adjustment in the *Predef* and *Self-det* conditions, later described in 3.4. The *DDA* condition relies on a knowledge tracing model for adjustment, which includes its own method for determining difficulty value.

3.3 Study Procedure

We conducted an online user study consisting of five stages, as shown in Figure 2. The first stage is the consent form and demographics questionnaire. Here, the participants state their age, gender, field and degree of education, their experience with graph theory, puzzle and strategy games, and programming, as well as their affinity for mathematics and puzzle games. We provide the

details about the questionnaire items in the appendix A. Then, the participants are provided with a tutorial about the task that resembles the part of a graph theory lecture that introduces the Maximum Independent Set (MIS). To show that they understood the task, the participants had to solve one tutorial exercise before moving on. After that is the main phase of the study, containing a pre-test, a practice stage, and a post-test. The pre-test and post-test each contain one easy, one medium, and one hard exercise. These test exercises are fixed for all the participants and all the conditions. The tests provide a bonus payment to the participants, which is the extrinsic motivation for the participants to practice. The practice stage consists of 12 exercises that are chosen differently depending on the participants' condition, and do not provide bonus payment. After the post-test comes the post-questionnaire, where the participants have to complete the Flow Short Scale questionnaire [10] and the NASA TLX questionnaire [13]. They are also prompted to give some free text comments on their feelings about the practice stage which is not reported in this paper due to space constraints. We provide further details about the pre- and post-questionnaire, listing all the questions in Appendix A.

3.4 Conditions

The difficulty levels of the practice exercises are selected differently, depending on the condition the participant is in. The first condition is the predefined difficulty level (*Predef*). The first practice exercise is an easy one, then the second is slightly more difficult, and so on until the hardest on the twelfth. The difficulty levels will be in this sequence, no matter how well the participant does. The second condition is the self-determined difficulty level (*Self-det*). The first practice exercise is at a medium difficulty level. After each exercise, the participants choose after each exercise whether they want the next exercise to be easier, harder, or of the same difficulty. The third condition is the dynamic difficulty adjustment (*DDA*). In this condition, the difficulty level of each exercise is determined by a *DDA* algorithm proposed in our previous work [40]. The algorithm is based on knowledge tracing that tries to find exercises whose difficulty level matches the participants' ability level by looking at the participant's previous performance. Prior work has demonstrated that the algorithm is successful in achieving the desired success rate, appropriately adapting the difficulty to the participants.

3.5 Dependent Metrics

We consider three main metrics in our study. First is the normalized learning gain (NLG), which captures the difference in score between the pre- and post-test. Each test has a full score of 3 points, and the normalized learning gain is defined in Equation 1 [27].

$$NLG = \begin{cases} \frac{post - pre}{3 - pre} & \text{if } post > pre \\ \frac{post - pre}{pre} & \text{if } post \leq pre \end{cases} \quad (1)$$

The two other metrics come from the post-questionnaire the participants have to fill out at the end. One of them is flow; the state of being focused on the task and being less aware of extraneous factors, including the passage of time. For this, we use the Flow Short

Scale to measure this [10]. The scale provides ten 7-point Likert scale items, measuring different components of flow. We calculate the flow score by taking the average score of all the scale items. Secondly, we also measure the perceived difficulty (PDif) of the practice exercises, which is an item of the NASA TLX questionnaire [13]. The participant chooses the perceived difficulty on a slider, with values going from 0 to 20. The numbers are not shown on the slider to the participant.

3.6 Participants and Compensation

We recruit 30 participants for each condition, totalling 90 participants. The participants are recruited on Prolific, and are required to be fluent in English. To filter inattentive online participants, we exclude participants who did not pass an attention check or had a contiguous gap of inactivity of more than 3 minutes during the puzzle phases. To account for outliers, we additionally remove participants whose NLG differs more than two standard deviations from the mean. In total, this leaves us with 86 participants with 28 in the *Predef* condition, 30 in the *Self-det* condition, and 28 in the *DDA* condition. The participants include 46 males and 40 females, with a mean age of 28.55. There were no meaningful differences between the three conditions for age, programming experience, puzzle and strategy gaming experience, and graph theory experience, as well as affinity for mathematics or puzzle games. The *DDA* and *Predef* conditions had 43% and 39% percent of female participants while the *Self-det* condition had 57%. The median time the experiment took was 23 minutes, with the pre-test, practice, and post-test together taking a median of 14 minutes. A majority of the participants have a Bachelor's degree as their highest education level.

Each participant was paid £3.9 for a successful participation. Additionally, for each correctly solved pre- and post-test exercise, they were paid £0.1 - 0.2, depending on the time they needed. They got a higher bonus if they completed the tests more quickly. This totals up to a potential bonus of £1.2, and was communicated to the participants before the pre-test.

4 ANALYSIS

4.1 Difficulty Adjustment Condition Analysis

We compare the different experimental conditions to see the effect of the difficulty adjustment methods. More precisely, we look at the three main dependent variables, namely the normalized learning gain, flow, and the perceived difficulty. We use the ANOVA test if normality and equal variance assumptions are fulfilled, and Kruskal-Wallis H-test otherwise.

4.2 Practice Behavior Clustering Analysis

For the second part of our methodology that allows us to investigate RQ2 and RQ3, we analyze the recorded learning trajectories from the user study participants, finding groups of behavior. The data from each participant consists of two parts: behaviors from the practice phase, and the evaluation measures. Our methodology can be summarized as follows: we first preprocess the data to prepare it for clustering, including filtering incongruent participants and outliers. We then cluster the participants by their behavior during the practice phase. To ensure meaningful differences between the

clusters, we statistically test whether the clusters differ in terms of the evaluation measures. We then use association rule mining to find the defining sets of behaviors pertaining to each cluster. Figure 3 illustrates the clustering method.

4.2.1 Feature Extraction & Preprocessing. To identify the different types of participants, we extract behavior features from the log data during the practice phase. These features needed to be formulated as a fixed-length vector. To this end, we use the number of clicks of each possible action in the exercises and the timing as elaborated below.

There are three possible actions in the exercises: set, unset, and reset. Set refers to adding an unset vertex into the selection. Unset does the opposite, removing a set vertex from the selection. Reset refers to clearing the entire selection. The counts of these three actions, summing over all the practice exercises, make up three features in our dataset. The fourth feature representing the participant is the average time between two consecutive actions. This feature cannot be calculated if a participant performs zero actions in one of the exercises, so such participants are removed from the dataset. These four behavior features together make up the feature vector representing one participant.

After extracting the behavior features, we now remove outliers from the dataset. The outliers are defined as those who have extreme values in the four features that we use for clustering. We filter out outliers using the Mahalanobis distance, used for multivariate data, with the cutoff threshold $\alpha = 0.01$, following [25]. Finally, to prepare the data for clustering, we normalize all the features with min-max normalization, such that all features lie inside the range $[0, 1]$.

4.2.2 Clustering. We use k-means to cluster the participants, using the scikit-learn library [32]. For this, we need to decide on the number of clusters, which we do by using the kneedle algorithm [39], the mathematical formulation of the commonly used elbow method. The kneedle algorithm finds the point of maximum curvature on the curve of the inertia against the number of clusters. The inertia is the sum of squared distances to the closest cluster center of each point.

4.2.3 Between-Cluster Comparisons. Here we apply statistical tests to check for differences between the clusters in terms of learning gain and subjective experience. First, we perform an omnibus test on three variables: learning gain, flow, and perceived difficulty. We use the ANOVA test or the Kruskal-Wallis H-test, as described in Section 4.1. Since we are particularly interested in the learning gain, we also perform pairwise comparisons, using either the two-sample t-test or the Mann-Whitney U-test, depending on the normality. We correct the pairwise comparisons by using the Benjamini-Hochberg procedure.

4.2.4 Association Rule Mining. After obtaining the clusters, we examine them more closely by finding out how the behavior features cause the different participants to land in different clusters. We use the Apriori algorithm [1] for association rule mining (ARM). ARM takes sets of items as input, called transactions, and outputs the extracted association rules. An association rule has a left-hand side (LHS) and a right-hand side (RHS), each of which is a set of items. A rule indicates that if a transaction contains the LHS, then it is

likely to also contain the RHS. This approach was previously used to point out differences between behavior clusters in [19, 24].

To apply association rule mining, we need to transform a feature vector \vec{x} representing a participant into a set of items, called transaction T . The behavior feature vector is written below in Equation 2:

$$\vec{x} = (x_{\text{Set}} \quad x_{\text{Unset}} \quad x_{\text{Reset}} \quad x_{\text{TimeBtwClicks}}) \quad (2)$$

where x_{type} is the different features. To convert this feature vector into a transaction, each feature x_{type} is transformed into an item type_j , where $j \in \text{hi, med, low}$, depending on the value x_{type} . We use a discretizer available from [32] based on 1-dimensional k-means of each feature to decide on the item. x_{type} gets translated to type_{hi} if x_{type} is closest to the highest centroid from k-means, and similar for the medium and lower centroid. T contains the 4 items from the behavior features, plus one more item, c_i , indicating that the participant is clustered into cluster i . We then feed the transactions into the Apriori algorithm, obtaining the association rules. We consider only rules with at most four behavior feature items on the left-hand side, and exactly one cluster item on the right-hand side. To filter rules that explain the clusters sufficiently well, we require the mined rules to have confidence of at least 0.6 and per-cluster support of 0.5, as defined in Equations 3 and 4 [19].

$$\text{Confidence} = p(RHS \subset T | LHS \subset T) \quad (3)$$

$$\text{Per-cluster support} = \frac{|\{T : (LHS \cup RHS) \in T\}|}{|\{T : RHS \in T\}|} \quad (4)$$

Confidence is the probability that a participant is clustered into the cluster on the right-hand side of the rule, given that the participant fulfills the left-hand side. Per-cluster support indicates the ratio of participants that the rule holds for that cluster, which is the right-hand side.

4.3 Cluster and Condition Comparison

To check whether the conditions have an effect on the participants' behavior during the practice phase, and by extension on the evaluation measures, we look at the clusters and the conditions together. More precisely, we look at how many participants in each cluster come from each condition, forming a contingency table. We apply the Chi-Square test on the contingency table to find if the difficulty adjustment methods influence the practice behavior of the participants. This analysis is designed to address RQ3.

5 RESULTS

5.1 Difficulty Adjustment Condition Analysis

We first present the comparison between the different difficulty adjustment methods, *Predef*, *Self-det*, and *DDA*. Because of non-normality as described in 4.1, we apply the ANOVA test for flow and Kruskal-Wallis (KW) H-test on learning gains and self-perceived difficulty. We also report η^2 as the effect size (*estimated* η^2 in case of KW H-test), as described in [43]². The plots are shown in Figure 4. Here we observe that there is no significant difference for learning gain (KW H-test, $H(2) = 1.31$, $p = 0.52$, $\eta^2 = -0.008$), flow (ANOVA,

²The classic effect size labels small, medium, and large are associated with threshold values for η^2 of 0.010, 0.059, 0.14 or for Cohen's d of 0.2, 0.5, 0.8 [12].

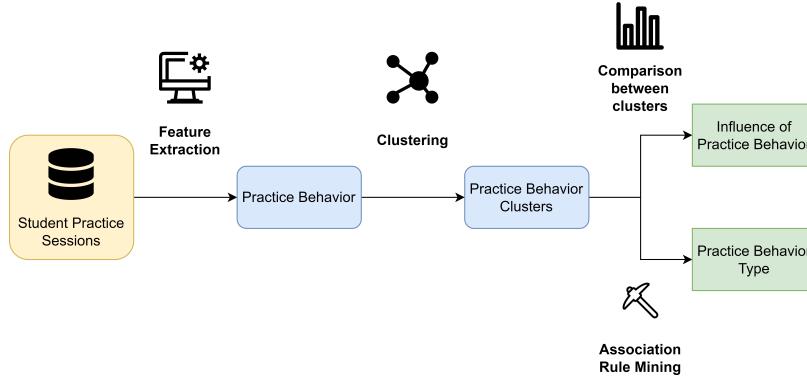


Figure 3: The clustering method. The practice behavior is extracted from recorded practice sessions, and then used for clustering. We compare the evaluation measures on the clusters to validate the differences between clusters and mine rules to see the defining characteristics of each cluster.

$F = 0.479, p = 0.94, \eta^2 = 0.001$), or perceived difficulty level (KW H-test, $H(2) = 1.47, p = 0.48, \eta^2 = -0.006$).

5.2 Practice Behavior Clustering Analysis

Now we turn to the clustering results. There are 6 students who didn't click in at least one of the practice exercises, and therefore cannot be used for clustering based on the average time between clicks. Removing the outliers by the Mahalanobis distance removes 5 more. After this, 75 participants remain for clustering. The kneedle algorithm reports that the optimal number of clusters is 5. K-means results in five clusters with sizes 24, 14, 9, 8, 20, respectively.

Figure 5 shows the learning gain, the self-reported flow, and the perceived difficulty of the practice phase for each cluster. To determine the differences in each measure, we apply a statistical test, either ANOVA or KW H-test, along with the effect size, as described in 5.1. For NLG, there is a significant difference between the clusters (KW H-test, $H(4) = 18.43, p = 0.0010, \eta^2 = 0.206$). For flow, there is no significant difference (ANOVA, $F = 1.0378, p = 0.39, \eta^2 = 0.056$). Finally, there is also a significant difference in perceived difficulty (ANOVA, $F = 5.21, p = 0.0010, \eta^2 = 0.229$).

To get more comprehensive differences among the clusters in terms of NLG, we apply statistical tests between each pair of clusters. Because of non-normality, we use Mann-Whitney (MW) U-tests for this. We report the test results and the probability of superiority (PS)³ as effect sizes [12]. To account for the large number of tests, we use Benjamini-Hochberg correction to correct the significance of the tests. Through this, we obtain three significant pairs: Clusters 0-4 (MW, $U = 103.5, p = 0.0007, PS = 0.784$), 1-4 (MW, $U = 62.0, p = 0.005, PS = 0.779$), and 3-4 (MW, $U = 27.5, p = 0.006, PS = 0.828$). From this, we interpret cluster 4 to have high NLG, and clusters 0, 1, and 3 to have low NLG. To verify grouping clusters 0, 1, and 3 together, we look at the differences between them. The PSs between these three clusters are classified as small or negligible (0-1: PS = 0.600, 0-3: PS = 0.518, 1-3: PS = 0.580) [12] so we consider them all as having low learning gains. As for cluster 2, we can't make

a statement about their NLG compared to other groups from the statistical tests.

Table 1 shows the association rules of the practice behavior clustering extracted by the Apriori algorithm. We only present rules whose left-hand side is not totally contained within another rule, which is more specific and describes the cluster with higher confidence. These rules point out the defining behavior of each cluster. In addition to the rules, we report the confidence and the per-cluster support, as defined in Equations 3 and 4.

Through clustering and the association rules, from Figure 5 and Table 1, we identify five behavior types during the practice phase, each of which leads to different learning gains. Using the left-hand side of the rules addressing the clusters as well as looking at the normalized learning gain (NLG) of each cluster, we give the clusters representative names to refer to them further on in the paper. From the low number of actions exhibited by the participants in cluster 0, the short time between the actions, as well as low NLG, we call cluster 0 the *few-shot rushers*. Participants in cluster 1, also exhibited low NLG, performed few actions, and took a long time between the actions, can be termed as the *strugglers*. Contrarily, participants in cluster 2 did a lot of sets and unsets, took a short time between the actions, and yielded a relatively high NLG, so we call them the *fast explorers*. Participants in cluster 3, which also had low NLG like clusters 0 and 1, use a low amount of unsets. This is also the only cluster in which the participants made moderate use of resets. They chose to clear their whole selection instead of backtracking more carefully with unsets. We thus call cluster 3 the *board-clearers*. Finally, participants in cluster 4 had medium sets, unsets, and time between clicks, as well as high NLG, which is why we call them the *thoughtful searchers*.

5.3 Cluster-Condition Relationship

Table 2 shows the number of participants in each condition and each cluster. Because of the small sample size and the large contingency table, we cannot apply the Chi-squared test and the Fisher's exact test to check for significant differences [42]. However, we can observe that other than cluster 3, clusters have comparable participants from each condition.

³Probability of superiority is associated with the effect size labels small, medium, and large by the threshold values of 0.56, 0.64, and 0.71.

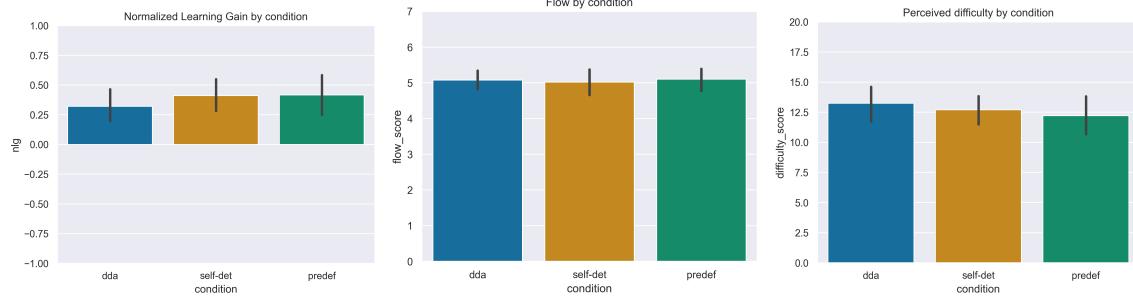


Figure 4: Comparison of different learning outcomes between the three conditions. The plots show from left to right, learning gain, flow, and perceived difficulty of each cluster of students. The error bars show 95% confidence interval. The results are provided in numbers in Table 3.

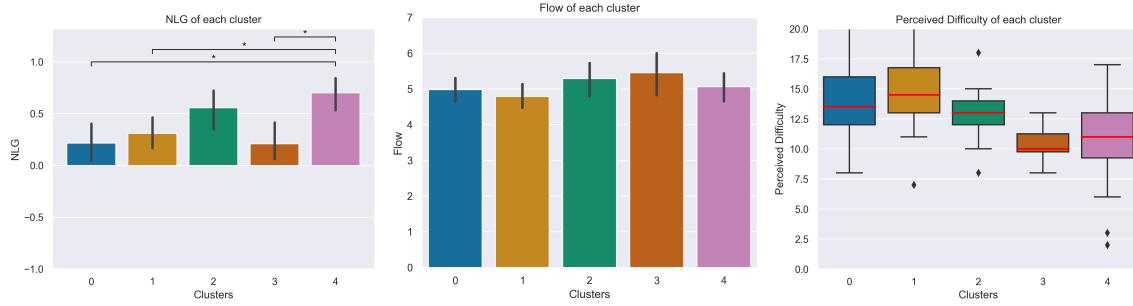


Figure 5: Results from clustering by practice behavior. The plots show from left to right, the learning gain, flow, and perceived difficulty of each cluster of students. The error bars show 95% confidence intervals for bar plots. The whiskers of the box plots span 1.5 times the interquartile range. The learning gain plot also shows significant pairs of differences after the Benjamini-Hochberg correction. The results are provided in numbers in Table 3. The sizes of the clusters are 24, 14, 9, 8, and 20, respectively.

Table 1: Rules extracted from the Apriori algorithm.

Left hand side	Right hand side	Confidence	Per-cluster support
{Set _{low} , Unset _{low} , Reset _{low} }	c ₀	0.577	0.625
{TimeBtwClicks _{low} , Reset _{low} }	c ₀	0.800	0.571
{TimeBtwClicks _{hi} , Unset _{low} , Reset _{low} }	c ₁	1.000	0.500
{TimeBtwClicks _{hi} , Set _{low} , Unset _{low} }	c ₁	1.000	0.500
{Reset _{low} , Set _{hi} , Unset _{hi} }	c ₂	1.000	0.667
{TimeBtwClicks _{low} , Set _{hi} , Reset _{low} }	c ₂	1.000	0.667
{TimeBtwClicks _{med} , Set _{med} , Unset _{low} }	c ₃	0.500	0.800
{TimeBtwClicks _{med} , Set _{med} , Reset _{med} }	c ₃	0.500	1.000
{Set _{med} , Unset _{low} , Reset _{med} }	c ₃	0.625	1.000
{TimeBtwClicks _{med} , Set _{med} , Unset _{med} , Reset _{low} }	c ₄	0.700	1.000

Table 2: Number of participants in each cluster and each condition.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Predef	7	3	4	1	8
Self-det	8	5	2	6	6
DDA	9	6	3	1	6

6 DISCUSSION

To answer RQ1, we compared the results of the difficulty adjustment methods on the practice exercises in terms of learning gains and subjective measures. Here, as mentioned in section 5.1, we found no significant differences between the three conditions. This suggests that in our open-ended reasoning context, the different methods of adjusting difficulty don't seem to play a role in influencing the

evaluation measures. This is not in line with our initial assumptions and motivations in the paper, and also not in line with other previous literature [22, 36, 37]. We hypothesize that the lack of a difference could be due to the simplicity of the task, in that there aren't many rules to learn. Our practice task stands in contrast to tasks in previous works, where there are many rules or mechanisms that the student needs to understand to successfully complete the task. If there were many rules to learn, easier levels could act as a scaffold to support students to learn a few things at a time, as was suggested in [38]. One more potential difference is the recruitment of the participants. Previous studies set their experiments inside an actual learning environment, whereas we recruited participants online, who might not be as intrinsically motivated to practice. Despite these potential explanations of our outcomes, we were still interested in investigating the influence of the practice phase more closely by analyzing students' practice behavior via clustering.

To investigate RQ2, we see if the data reveals associations between certain student practice behaviors and learning. From our results, clustering showed groups of behavior types, which are linked to different learning gains, as listed in Section 5.2. These results agree with previous works [4, 29]. Looking at the different characteristics, we found one cluster that learned significantly well: the *thoughtful searchers*. The defining feature of this cluster is using many sets and unsets and taking some time between the actions. This suggests going through different options to explore the environment and taking time to think about what to do next or what has been done, i.e., exhibiting reflection. This goes in line with what previous literature also highlights regarding the connection between exploratory and reflective behavior in open-ended tasks with learning [29]. Next up is the cluster that couldn't be defined as low- or high-learning: the *fast explorers*. They also used a lot of sets and unsets, but did so quickly. This seems to indicate that while they explored a lot, they gave less thought to each action that they did. Looking at their learning gain, they seem to have done better than low-learners, but not as well as the *thoughtful searchers*. This indicates that their behavior is in the right direction, but they could still improve more, if they were nudged to spend more time to reflect on their actions.

There are also three clusters that didn't learn well. Two low-learning clusters, the *few-shot rushers* and the *strugglers*, both used few sets, unsets, and resets. In line with the aforementioned line of thought, a lack of exploration, i.e., not using enough actions during practice sensibly may lead to poor learning. Interestingly, the third low-learning cluster, the *board-clearers*, used a moderate number of sets, but a low number of unsets, choosing to use resets instead. This seems to indicate that using resets is counterproductive for learning. Aside from that, we also exploratively found that the *board-clearers* have higher pre-test scores than other clusters. Potentially, this could have contributed to the lower learning gain of this cluster as there is less margin to improve.

Having seen the differences between the behavior types, they suggest that student's behaviors can also inform about the student's learning in addition to just the practice exercise outcomes.

In order to answer the third RQ, we looked at how the students from the three conditions are distributed across the behavioral clusters as shown in Table 2. As mentioned in the previous section, since it is not possible to apply the Chi-squared or the Fisher's exact

test to the data because of the small sample size in relation to the contingency table, we look at the trends. Judging from the numbers, the overall distribution seems to be uniform across the clusters with the exception of *board-clearers*, cluster 3, which seems to have most students from the *Self-det* condition. It is interesting to note that this is the only cluster with higher pre-test scores and a relatively high number of resets suggesting a trend that when difficulty is adjusted based on students' perceived notion of difficulty (*Self-det* condition), a behavior of excessively resetting the environment might suggest the students come with a higher prior knowledge. This may inform the intervention scheme as the action of resetting the environment can mean different things for learning.

Our work has some limitations. Firstly, we have only used data from one user study on one task. It would be valuable to see results from other open reasoning tasks, and with more participants. In particular, we weren't able to use statistical tests to address RQ3 because of the sparsity of participants. Secondly, the user study we used recruited participants through crowdsourcing, and was not embedded as a part of a curriculum that students have to go through. This might limit the generalizability of the results, particularly in comparing different difficulty adjustment methods. Thirdly, the features we have used in this work are summative features, aggregating over all the exercises. Although the features were able to explain differences between participants, they don't account for the differences in exercises' difficulty levels and the sequence of exercises.

7 CONCLUSION

In this work, we have compared three different methods of difficulty adjustment, applying them in an open reasoning task. Secondly, we have identified different types of practice behaviors in the task using clustering. Thirdly, we have examined the link between behavior types and difficulty adjustment methods. Our results suggest some interesting insights and takeaways for the community: (i) Differences in the way exercises are presented to students based on difficulty might not always lead to differences in learning gains, particularly in open-ended reasoning tasks, contrasting some previous works. (ii) As we observe differences in the behaviors that link to high and low learning, this suggests the usefulness of complementing in-task performance measures with students' behavior types to inform the difficulty adjustment interventions in open-ended learning environments.

As future works, one straightforward continuation is to design an intervention to lead students to practice behaviors indicative of improved learning. Further evaluation of intervention methods could yield guidelines for adaptive application designs. Another interesting line is to more deeply understand the impact of difficulty on the practice behavior at the per-exercise level, rather than at the per-condition level as we did. Results from there could potentially allow us to understand the impact of difficulty levels more deeply.

ACKNOWLEDGMENTS

This paper was partially funded by the DFG through the Leibniz award of Elisabeth André (AN 559/10-1).

REFERENCES

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Santiago, Chile, 487–499.

[2] Robert Belfer, Ekaterina Kochmar, and Julian Vlad Serban. 2022. Raising Student Completion Rates with Adaptive Curriculum and Contextual Bandits. In *International Conference on Artificial Intelligence in Education*. Springer, 724–730.

[3] Suzete Chiviacowsky, Gabriele Wulf, Rebecca Lewthwaite, and Tiago Campos. 2012. Motor learning benefits of self-controlled practice in persons with Parkinson's disease. *Gait & Posture* 35, 4 (2012), 601–605.

[4] Cristina Conati and Samad Kardan. 2013. Student modeling: Supporting personalized instruction from problem solving to exploratory open ended activities. *Ai Magazine* 34, 3 (2013), 13–26.

[5] Gemma Corbalan, Liesbeth Kester, and Jeroen JG Van Merriënboer. 2008. Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology* 33, 4 (2008), 733–756.

[6] Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[7] Alana M De Morais, Joseana MFR Araujo, and Evandro B Costa. 2014. Monitoring student performance using data clustering and predictive modelling. In *2014 IEEE frontiers in education conference (FIE) proceedings*. IEEE, 1–8.

[8] Edward L Deci, Robert J Vallerand, Luc G Pelletier, and Richard M Ryan. 1991. Motivation and education: The self-determination perspective. *Educational psychologist* 26, 3-4 (1991), 325–346.

[9] Andrew Emerson, Andy Smith, Fernando J Rodriguez, Eric N Wiebe, Bradford W Mott, Kristy Elizabeth Boyer, and James C Lester. 2020. Cluster-based analysis of novice coding misconceptions in block-based programming. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 825–831.

[10] Stefan Engeser and Falko Rheinberg. 2008. Flow, performance and moderators of challenge-skill balance. *Motivation and emotion* 32 (2008), 158–172.

[11] Lauren Fratamico, Cristina Conati, Samad Kardan, and Ido Roll. 2017. Applying a framework for student modeling in exploratory learning environments: Comparing data representation granularity to handle environment complexity. *International Journal of Artificial Intelligence in Education* 27 (2017), 320–352.

[12] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* 141, 1 (2012), 2.

[13] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[14] John Hattie. 2012. *Visible learning for teachers: Maximizing impact on learning*. Routledge.

[15] Danial Hooshyar, Margus Pedaste, Yeongwook Yang, Liina Malva, Gwo-Jen Hwang, Minhong Wang, Heuiseok Lim, and Dejan Delev. 2021. From gaming to computational thinking: An adaptive educational computer game-based learning approach. *Journal of Educational Computing Research* 59, 3 (2021), 383–409.

[16] Michael G Hughes, Eric Anthony Day, Xiaolian Wang, Matthew J Schuelke, Matthew L Arsenault, Lauren N Harkrider, and Olivia D Cooper. 2013. Learner-controlled practice difficulty in the training of a complex task: Cognitive and motivational mechanisms. *Journal of Applied Psychology* 98, 1 (2013), 80.

[17] Robin Hunicke. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. 429–433.

[18] Emily Jensen, Tetsumichi Umada, Nicholas C Hunkins, Stephen Hutt, A Corinne Huggins-Manley, and Sidney K D'Mello. 2021. What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 121–131.

[19] Samad Kardan. 2017. *A data mining approach for adding adaptive interventions to exploratory learning environments*. Ph.D. Dissertation. University of British Columbia. <https://doi.org/10.14288/1.0348694>

[20] Samad Kardan and Cristina Conati. 2015. Providing adaptive support in an interactive simulation for learning: An experimental evaluation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3671–3680.

[21] Tanja Käser and Daniel L Schwartz. 2020. Modeling and analyzing inquiry strategies in open-ended learning environments. *International Journal of Artificial Intelligence in Education* 30, 3 (2020), 504–535.

[22] Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. 2011. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education* 57, 2 (2011), 1813–1824.

[23] Danny Kostons, Tamara van Gog, and Fred Paas. 2010. Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education* 54, 4 (2010), 932–940.

[24] Sébastien Lallé and Cristina Conati. 2020. A data-driven student model to provide adaptive support during video watching across MOOCs. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I* 21. Springer, 282–295.

[25] Christophe Leys, Olivier Klein, Yves Dominicy, and Christophe Ley. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of experimental social psychology* 74 (2018), 150–156.

[26] Gabriel Lins de Holanda Coelho, Paul HP Hanef, and Lukas J. Wolf. 2020. The very efficient assessment of need for cognition: Developing a six-item version. *Assessment* 27, 8 (2020), 1870–1885.

[27] Jeffrey D Marx and Karen Cummings. 2007. Normalized change. *American Journal of Physics* 75, 1 (2007), 87–91.

[28] Jauwairia Nasir, Pierre Dillenbourg, Utku Norman, and Barbara Bruno. 2020. When Positive Perception of the Robot Has No Effect on Learning. In *29th IEEE International Conference on Robot and Human Interactive Communication, ROMAN 2020, Naples, Italy, August 31 - September 4, 2020*. IEEE, 313–320. <https://doi.org/10.1109/RO-MAN47096.2020.9223343>

[29] Jauwairia Nasir, Aditi Kothiyal, Barbara Bruno, and Pierre Dillenbourg. 2021. Many are the ways to learn: identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning* 16, 4 (2021), 485–523.

[30] Christopher P Niemiec and Richard M Ryan. 2009. Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and research in Education* 7, 2 (2009), 133–144.

[31] Kamalesh Palani, Paul Stynes, and Pramod Pathak. 2021. Clustering Techniques to Identify Low-engagement Student Levels.. In *CSEDU* (2). 248–257.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[33] Radek Pelánek, Jan Papoušek, Jiří Rihák, Vít Stanislav, and Juraj Nižnan. 2017. Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction* 27, 1 (2017), 89–118.

[34] Young K Ro, Yi Maggie Guo, and Barbara D Klein. 2018. The case of flow and learning revisited. *Journal of education for business* 93, 3 (2018), 128–141.

[35] Henry L Roediger III and Jeffrey D Karpicke. 2006. The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science* 1, 3 (2006), 181–210.

[36] Cristóbal Romero, Sebastián Ventura, Eva L Gibaja, Cesar Hervás, and Francisco Romero. 2006. Web-based adaptive training simulator system for cardiac life support. *Artificial Intelligence in Medicine* 38, 1 (2006), 67–78.

[37] Ron JCM Salden, Fred Paas, and Jeroen JG Van Merriënboer. 2006. Personalised adaptive task selection in air traffic control: Effects on training efficiency and transfer. *Learning and Instruction* 16, 4 (2006), 350–362.

[38] Sandra Sampayo-Vargas, Chris J Cope, Zhen He, and Graeme J Byrne. 2013. The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education* 69 (2013), 452–462.

[39] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.

[40] Anan Schütt, Tobias Huber, Ilhan Aslan, and Elisabeth André. 2023. Fast Dynamic Difficulty Adjustment for Intelligent Tutoring Systems with Small Datasets. In *Proceedings of the 16th International Conference on Educational Data Mining*. Mingyu Feng, Tanja Käser, and Partha Talukdar (Eds.). International Educational Data Mining Society, Bengaluru, India, 482–489. <https://doi.org/10.5281/zenodo.8115740>

[41] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131* (2014).

[42] Merle W Tate and Leon A Hyer. 1973. Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small. *J. Amer. Statist. Assoc.* 68, 344 (1973), 836–841.

[43] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. (2014).

[44] Ernst Von Glaserfeld. 1998. Cognition, construction of knowledge, and teaching. In *Constructivism in Science Education: A Philosophical Examination*, Michael R. Matthews (Ed.). Springer, 11–30.

[45] Gabriele Wulf, Heidi E Freitas, and Richard D Tandy. 2014. Choosing to exercise more: Small choices increase exercise engagement. *Psychology of Sport and Exercise* 15, 3 (2014), 268–271.

[46] Yaqian Zhang and Wooi-Boon Goh. 2021. Personalized task difficulty adaptation based on reinforcement learning. *User Modeling and User-Adapted Interaction* 31, 4 (2021), 753–784.

Table 3: Comparison of different learning outcomes of the three conditions and five clusters in numbers, written as (Mean \pm SD).

	Predef	Self-det	DDA	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
<i>NLG</i>	0.41 ± 0.45	0.41 ± 0.38	0.32 ± 0.37	0.22 ± 0.43	0.31 ± 0.29	0.56 ± 0.29	0.21 ± 0.27	0.70 ± 0.37
<i>Flow</i>	5.10 ± 0.85	5.02 ± 0.99	5.07 ± 0.69	4.98 ± 0.83	4.79 ± 0.64	5.29 ± 0.71	5.45 ± 0.84	5.06 ± 0.88
<i>Pcv. diff.</i>	12.21 ± 4.28	12.70 ± 3.32	13.25 ± 3.77	14.21 ± 3.43	14.36 ± 3.20	13.00 ± 2.71	10.38 ± 1.50	10.50 ± 3.72

A QUESTION ITEMS IN THE USER STUDY

We list the items in the pre- and post-questionnaire of the user study here.

A.1 Pre-Questionnaire

The pre-questionnaire is the first thing the participant does in the study. The questionnaire starts out by asking about the demographics: age, gender, and degree and field of study. The answers for gender, degree, and field of study are chosen from a drop-down list. Then it asks about previous experience with mathematics and computer science, asking the participant to choose how well each statement describes him on a 7-point Likert scale. The items are:

- I have heard of graph theory before.
- I like mathematics.
- I like puzzle games.
- I know how to procedurally solve puzzles (Sudoku, Rubik's cube, etc.).
- I have experience in programming.
- I have played strategy games before (Chess, Go, Red Alert, etc.).

The Likert scale also contains an attention check item, where the participant is told to pick the fourth option. If a participant doesn't pick the right item, their data is removed from the analysis. The participant must answer every item to continue to the next page.

A.2 Post-Questionnaire

The post-questionnaire is the very last page of the study. It asks about the experience during the practice phase. The first set of questions comes from the Flow Short Scale [10], prompting the participant to state how much they agree with each statement about their feelings while doing the exercises on a 7-point Likert scale:

- I feel just the right amount of challenge.
- My thoughts/activities run fluidly and smoothly.
- I don't notice time passing.
- I have no difficulty concentrating.
- My mind is completely clear.
- I am totally absorbed in what I am doing.
- The right thoughts/movements occur of their own accord.
- I know what I have to do each step of the way.
- I feel that I have everything under control.
- I am completely lost in thought.

Again, the Likert scale contains an attention check item. The questionnaire proceeds by asking about the need for cognition, with items taken from [26]. The participant should answer how characteristic or uncharacteristic each item is for them on a 5-point Likert scale:

- I would prefer complex to simple problems.
- I like to have the responsibility of handling a situation that requires a lot of thinking.
- Thinking is not my idea of fun.
- I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
- I really enjoy a task that involves coming up with new solutions to problems.
- I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

Then comes the NASA TLX questionnaire [13], where participants describe their feelings on a continuous slider. The questions are:

- How mentally demanding was the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?
- How difficult was the task overall?

Finally, there are two free-text questions: "Please give comments about the difficulty level of the puzzles and their impact on your training" and "Please give an impression of how you felt during the puzzles.". As in the pre-questionnaire, the participant must answer every item to continue to end the study.

B RESULTS IN TABLES

We provide the results from Figures 4 and 5 as numbers in Tables 3.