systems

MDPI

Article

When Easy Becomes Boring and Difficult Becomes Frustrating: Disentangling the Effects of Item Difficulty Level and Person Proficiency on Learning and Motivation

Mariola Moeyaert [1,*,†], Kelly Wauters [2,†], Piet Desmet [3] and Wim Van den Noortgate [4]

1 Department of Educational Psychology and Methodology, State University of New York, 1400 Washington Ave, Albany, NY 12222, USA
2 Federale Overheidsdienst Financien, 50 Kruidtuinlaan, Brussels 1000, Belgium; kelly@allnice.be
3 Faculty of Arts and iMinds-ITEC, KU Leuven 3000, Belgium; piet.desmet@kuleuven-kulak.be
4 Faculty of Psychology and Educational Sciences and iMinds-ITEC, KU Leuven 3000, Belgium; wim.vandennoortgate@kuleuven-kulak.be
* Correspondence: mmoeyaert@albany.edu; Tel.:+1-518-618-6056
† These authors contributed equally to this work.

Academic Editors: Nguyen-Thinh Le, Jon Mason, Christian Gütl and Kiyoshi Nakabayashi
Received: 24 September 2015; Accepted: 22 February 2016; Published: 3 March 2016

Abstract: The research on electronic learning environments has evolved towards creating adaptive learning environments. In this study, the focus is on adaptive curriculum sequencing, in particular, the efficacy of an adaptive curriculum sequencing algorithm based on matching the item difficulty level to the learner's proficiency level. We therefore explored the effect of the relative difficulty level on learning outcome and motivation. Results indicate that, for learning environments consisting of questions focusing on just one dimension and with knowledge of correct response, it does not matter whether we present easy, moderate or difficult items or whether we present the items with a random mix of difficulty levels, regarding both learning and motivation.

Keywords: adaptive item sequencing; item-based learning; computer-assisted learning; difficulty; proficiency

1. Introduction

The appearance and functionality of electronic learning environments have changed tremendously as a result of both technological advances and the increased attention of researchers and companies [1,2]. Research is today mainly focused on creating an adaptive learning environment in which one or more characteristics of the learning environment (e.g., difficulty of the items and type of feedback) are adapted to one or more features of the learner. Different classifications of learner features are suggested in the previous research. For instance [3] differentiate between proficiency and learning styles; [4]) go a step further and divide learner's features into static (e.g., age and tongue language) and dynamic categories (e.g., proficiency and learning styles); [5] differentiates between domain-specific (e.g., proficiency and skills) and domain-independent information (*i.e.*, learning preferences and demographics). In this study, the learner's feature of interest is the learner's proficiency, and the learning environment characteristic of interest is the item difficulty level. Such a personalized/individualized learning environment can, for instance, incorporate an adaptive item curriculum sequencing algorithm that provides a sequence of items that is contingent on the performance of the learner on previous items and on the difficulty of the remaining unsolved items [6,7].

Adaptive curriculum sequencing requires two main processes: The first process implies estimating the learner's proficiency level and the difficulty level of the course material (*i.e.*, the difficulty of the

items presented to the learners). The second process makes use of the obtained estimates to optimize the interaction between the learner and the learning material given to him [8]. The focus of this article is on the second process and investigates the relationship between item difficulty level and person proficiency level, and their impact on learning and motivation using Item Response Theory (IRT) in an item-based learning environment.

Before this can be investigated, we further clarify the kind of e-learning environment that is the focus of this study, because the characteristics of the e-learning environment affects the applied estimation method for determining the learner's prior proficiency and the difficulty level of the learning material, as well as the choice of sequencing algorithm. We focus on item-based e-learning environments that consist of small, independent tasks (which we call items), in which users learn from making items and getting immediate feedback on their answers. Items differ in the amount of ability, but not the kind of ability that is needed to solve them correctly. Most previous research attention is drawn to the adaptive sequencing algorithm underlying learning environments with tasks, items, or learning materials that are not independent, but rather linked by some kind of relationship. For instance, if a learner answers an item related to the dimension comprehensive reading correctly, then the probability is high that the learner will also correctly answer items related to the dimension technical reading because technical reading is a prerequisite for reading comprehension. The adaptive sequencing algorithm most frequently implemented in such learning environments is the classic rule-based curriculum sequencing techniques [9,10] and the probabilistic graphical models, such as Bayesian networks [11]. In rule-based curriculum sequencing, the learning path is based on the relationship between multiple dimensions defined by experts. Each dimension represents items related to a specific latent proficiency [12,13]. For instance, reading fluency and verb conjugation are two different latent proficiencies (each with their own series of items), and both proficiencies are expected to be correlated. The relationship in simple rule-based curriculum sequencing techniques is the prerequisite relationship, and curriculum sequencing is based on a single rule: Learn prerequisite knowledge first. Bayesian networks differ from rule-based models because they incorporate uncertainty into the different relationships by modeling the strength of a relationship as a probability. Furthermore, Bayesian networks are used to update the probabilities when information comes in.

One disadvantage of these curriculum sequencing techniques is that they are only applicable to learning environments that include learning material consisting of dimensions that are linked by some kind of relationship (e.g., prerequisite, analogy, *etc.*). However, some learning environments consist of unrelated dimensions. While there is a fairly large body of research related to adaptive curriculum sequencing in learning environments with linked dimensions (*i.e.*, multidimensional learning environments), less research has been conducted on sequencing algorithm techniques in learning environments where this relationship is absent, *i.e.*, unidimensional learning environments [14]. Even though [12] states that such a simple curriculum can only be offered by random question sequencing, the aim of this study is to explore whether this statement can be underpinned by empirical research or whether a specific sequencing algorithm can be applied to item-based adaptive learning environments in order to improve learning efficiency, taking motivation into account. To reach our goal, a brief elaboration on the estimation of the learner's proficiency level and the item difficulty level is offered, followed by the implementation of this estimation method for adaptive item sequencing in testing environments and in item-based adaptive learning environments. Finally, we will argue that the optimal relative difficulty (depending on the learner's proficiency and item difficulty) might not be fixed for a given learner, but could increase or decrease according to his or her proficiency level [12].

2. Adaptive Item Sequencing in Item-Based Learning Environments

Simultaneous estimation of the item difficulty parameters and the learner's proficiency level and its application. It is important to estimate both the item difficulty level and the learner's proficiency level because the relative difficulty might be more influential than the absolute item difficulty. In item-based learning environments, the probability of success can be estimated by means

of item response theory (IRT) [15]. IRT is a psychometric approach that emphasizes the fact that the probability of a discrete outcome, such as the correctness of a response to an item, is influenced by qualities of the item and by qualities of the person. Various IRT models exist, differing in degree of complexity, with the simplest IRT model stating that a person's response to an item depends on the person's proficiency level and the item's difficulty level [16]. The item difficulty parameter (*i.e.*, β_i) and the person proficiency parameter (*i.e.*, θ_s) can be found in the following Equation:

$$\pi_{pi} = P(X_{si} = 1) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)} \quad (1)$$

As a consequence, IRT makes it possible to estimate the probability of success (*i.e.*, $\pi_{pi} = P(X_{si} = 1)$ for each combination between an item difficulty and a person's proficiency. The person and item parameter can be placed on the same continuous scale, making it possible to match the difficulty of the item to the proficiency level of the learner. More specifically, the difficulty level of an item can be interpreted as the proficiency needed to have a 0.5 probability (*i.e.*, $\pi_{pi} = 0.5$) of giving a correct answer. The higher the person's proficiency compared to the item difficulty (*i.e.*, $\theta_s - \beta_i$ is high and positive), the greater the probability of giving a correct answer (*i.e.*, the higher π_{pi}). The reverse results in a smaller probability. As a consequence, the Rasch model presented in Equation (1) takes into account that the difficulty of an item is relative to the learner's ability (and this is reflected in the probability to give a correct answer, the relative item difficulty).

A major implementation of IRT is situated in computerized adaptive testing (CAT) [17,18] where adaptive item sequencing is used to get an estimate of the true underlying students' proficiency, based upon the item difficulty and the student proficiency. By varying the difficulty level of the item, one can evaluate the change in probability of giving a correct answer. More specifically, the sequencing algorithm in CAT, guided by the objective of precise measurement, targets the item that provides the most information on the person's proficiency level. For the Rasch model, this means that items are administered for which the person is expected to have about a 50% probability of answering the item correctly (*i.e.*, items with a difficulty level close to the proficiency level). The sequencing algorithm in CAT can be described using following steps: (1) A prior calibration study to create calibrated items is conducted (*i.e.*, the items difficulty parameter value were estimated); (2) items to the participants with an optimal difficulty level are administered (more specifically, the optimal difficulty level of an item can be interpreted as the proficiency needed to have a 0.50 probability (*i.e.*, $\pi_{pi} = 0.50$) of giving a correct answer); (3) the participants proficiency level is estimated; (4) items with adjusted difficulty level are administered (based upon the estimated proficiency level in step 3); (5) the estimated proficiency level is adjusted; (6) step 1 is repeated, or, if the proficiency level is accurately estimated (or if the test length has attained its maximum size), the sequence is stopped. The purpose of this manuscript is step (4), estimating the adjusted difficulty level.

An important difference between testing and learning environments makes us question whether this CAT item selection algorithm is also suitable for learning purposes. While in testing environments, the objective is to select the item that would be most informative for refining the person's proficiency estimate; in learning environments, the objective is to select the item that optimizes the probability of progressing to a higher proficiency level [19].

Adaptive item sequencing. Researchers have recognized the importance of developing a productive adaptive curriculum sequencing strategy as a strategy that leads to effective and efficient learning. Whether this strategy alternates between difficult and easy items, aims at resolving misconceptions, or makes the decision based on the ideas of CAT, the overall objective is to enhance learning and increase or maintain motivation. Some theories, such as the flow theory [20] and the self-determination theory [21], state that learners are more motivated by challenging tasks. According to the flow theory, the learner's perceived challenge and his or her proficiency should be balanced. If there is not a good balance, feelings of anxiety (with high challenge and low ability) and boredom or uninvolvement (with low challenge and high proficiency) will be the result [22]. Moderately

challenging tasks, *i.e.*, tasks that are somewhat beyond the learner's current proficiency [20,21,23], make learners, on the one hand, aware that they lack some proficiency but on the other hand keep them involved [24,25]. Hence, those intermediate difficult problems engage learners and lead to a greater enjoyment of the task. Furthermore, learners conducting moderately challenging tasks feel more successful, efficacious, and in control of their own learning [25]. On the other hand, overly challenging tasks can have an adverse effect on motivation and persistence. More specifically, tasks that are too difficult relative to the learner's actual proficiency or his/her perceived proficiency have a negative impact on the feeling of competence, expectations of success, and enjoyment of the activity, and increase anxiety [21,22,26,27] . The underlying idea, supported by the flow theory, is that those feelings of anxiety may inhibit the learner's involvement and task engagement [20,28]. Furthermore, overly challenging or difficult tasks can be perceived as a threat to the learner's sense of competence, resulting in lower self-efficacy [26]. The learner's interest in the learning material may buffer the negative effects of overly challenging tasks on motivation [29]. Interest is composed of both intrinsic motivation [30] and task value/task motivation [31]. Learners who are interested in the task are more likely to enjoy challenging tasks, while learners who are not interested in the task are more likely to avoid challenge [32,33]. The underlying process is possibly mediated by arousal and attention [34].

In sum, for maximizing the motivation of learners, tasks should still provide an intermediate probability of success, rather than offering an almost certain probability of success (*i.e.*, the probability of correctly answering an item is always close to 1) or failure (*i.e.*, the probability of correctly answering an item is always close to 0 [29,35,36]. In other words, research is needed to find the optimal probability to answer an item correct (*i.e.*, π_{pi} in Equation (1)) in order to keep the learner engaged. This probability depends on both the learner's proficiency and the item difficulty as reflected in the Rasch algorithm presented in Equation (1). In instructional game research, it is indeed found that too easy or too difficult games can lead to a reduction in motivation and, in time, on task [37]. [38] further argue that this effect may result in less positive learning outcomes [39]. However, with regard to learning outcome, results are found to be inconsistent. [40] studied the effect of feedback and adaptive sequencing of tasks on learning outcome and learning efficiency. Results indicated that adaptive task sequencing does not lead up to more effective learning. On the other hand, some studies did find a significant effect of adapting the difficulty and support of learning tasks to the learner's competences and perceived cognitive load [41–43]. Other researchers found a positive effect of IRT-based adaptive item sequencing on learning. More precisely, adaptive environments in which items were selected because the learner had a 50% probability of answering them correctly yielded faster learning than a non-adaptive learning environment [44,45]. Furthermore, research on CAT suggests that administering easier items would foster motivation and lead to a higher performance score, especially for persons with a low proficiency level [46]. Hence, the selection of challenging tasks is not only supposed to enhance motivation, but could also have an effect on learning. In addition to that, prior research found that learner's characteristics and, in particular, the learner's proficiency can influence learning outcomes [38,47–50] and the need for any enhancement to the basic learning material, such as adaptive task sequencing [51–55]. The overall finding is that students with low proficiency benefit more from adaptive learning environments than do students with high proficiency [38,53,55]. However, one study found that adapting the difficulty is more beneficial for advanced learners than it is for the novice or intermediate learners [52]. Based on this prior research, the present study will also examine the influence of the learner's prior proficiency level on the relationship between the adaptive item sequencing algorithm and motivation and learning outcomes in item-based learning environments.

Previous studies did not differentiate between different levels of difficulty and different levels of proficiency, which is needed for an accurate estimate of the relative item difficulty. In this study, we sought to provide initial evidence as to whether a particular relative difficulty level is more effective than others in a specific item sequencing algorithm in terms of learning and motivation and hence aim to answer the following question: What is the optimal relative difficulty to use in an item-based learning environment where the item difficulty level and the learner's proficiency level is estimated by

means of IRT? The relationship between item difficulty level and person proficiency level, and their impact on learning and motivation, will be disentangled.

3. Experiment

To date, no previous research has systematically compared item selection algorithms in item-based adaptive learning environments by considering different levels of item relative difficulty. This research sought to provide initial evidence as to whether a particular relative item difficulty level (*i.e.*, the probability of answering an item correctly) is more effective than others in terms of maintaining learner's motivation and, in turn, enhancing learning outcomes. We chose to examine six different item selection algorithms in the learning environment: items for which the learner has a probability between 0.40 and 0.50 of answering the item correctly (*i.e.*, π_{pi} in Equation (1) ranges from 0.40 to 0.50), a probability between 0.50 and 0.60, 0.60 and 0.70, 0.70 and 0.80, 0.80 and 0.90, and a selection algorithm that randomly selects items for which the learner has a probability between 0.40 and 0.90 of answering the item correctly. The model used to estimate the relative item difficulty is presented in Equation (1). The outcome score (*i.e.*, π_{pi}, the relative difficulty or probability of correctly responding the item) is a function of the item difficulty and learner's proficiency parameters, β_i and θ_s, respectively

Following previous studies focusing on adaptive technologies [40,56], we predict that adaptive item sequencing will result in higher learning outcomes and a higher level of motivation.

This results in following research hypotheses:

(1) Items with a moderate relative difficulty (π_{pi} = 0.60–0.70) will result in higher task involvement, higher interest, and higher perceived competence than when presenting more difficult items (π_{pi} = 0.40–0.60).

(2) Relatively easy items (π_{pi} = 0.70–0.90) will result in lower task involvement (effort). The learner's interest (intrinsic motivation and task motivation) is presumed to buffer the negative effect that difficult items have on motivation.

(3) Proficiency has a moderating effect on the relationship between the relative item difficulty level and learning outcome. In other words, the relation between relative item difficulty and learning outcome depends on proficiency.

4. Method

Participants. Students from ten educational programs in the Flemish part of Belgium (1st and 2nd year of the Bachelor Linguistics and Literature—KU Leuven; 1st, 2nd and 3rd year of the Bachelor Teacher-Training for primary education—Katho Tielt; 1st and 2nd year of the Bachelor Teacher-Training for secondary education—Katho Reno; 1st and 2nd year of the Bachelor of Applied Linguistics—HUB and Lessius; and 1st year of the Bachelor Educational Science—KU Leuven) were contacted to participate in the experiment. Two hundred twenty participants completed the entire study (*i.e.*, pre-test, learning phase and post-test). Descriptive statistics of the participants are presented in Table 1.

Design. In a pre-test, proficiency and motivation were measured. A covariate adaptive randomization design was used with proficiency (6 levels) as covariate. Participants within each covariate level were randomly assigned to one of the six between-subject conditions (*i.e.*, relative difficulty level) that were part of the learning phase: (1) very difficult (VD), in which participants had a probability between 0.40 and 0.50 to answer an item correctly; (2) difficult (D), in which participants had a probability between 0.50 and 0.60 to answer an item correctly; (3) moderate (M), in which participants had a probability between 0.60 and 0.70 to answer an item correctly; (4) easy (E), in which participants had a probability between 0.70 and 0.80 to answer an item correctly; (5) very easy (VE), in which participants had a probability between 0.80 and 0.90 to answer an item correctly; and (6) random (R), in which participants were presented a random set of items for which they had a probability between 0.40 and 0.90 of answering those items correctly. Every difficulty condition included a similar number of participants: $n(VD) = 36$; $n(D) = 34$; $n(M) = 38$; $n(E) = 37$; $n(VE) = 39$; $n(R) = 31$). After the

learning phase, a post-test was administered, consisting of a proficiency test and post-experimental motivation measurement.

Table 1. Descriptive Statistics of Study Participants.

Variables	Min	Max	M	SD	Frequency	Percentage
Age	18	48	18.630	2.164		
Weekly hours of French in the last year of secondary education	0	6	3.580	0.750		
Course credits in French from current academic year in current education[b]	0	56	3.520	8.128		
Gender						
Missing					5	2.27
Male					21	9.55
Female					194	88.18
Current Education						
Missing					9	4.09
1st Year Bachelor Linguistics & Literature					20	9.09
2nd Year Bachelor Linguistics & Literature					4	1.82
1st Year Bachelor Teacher-Training primary education					1	0.45
2nd Year Bachelor Teacher-Training primary education					5	2.27
3rd Year Bachelor Teacher-Training primary education					0	0
1st Year Bachelor Teacher-Training secondary education					10	4.54
2nd Year Bachelor Teacher-Training secondary education					4	1.82
1st Year Bachelor of Applied Linguistics					0	0
2nd Year Bachelor of Applied Linguistics					0	0
1st Year Bachelor Educational Sciences					167	75.91
Secondary Education						
Missing					14	6.36
GSO					201	91.36
TSO					5	2.27

Note: GSO = General Secondary Education. TSO = Technical Secondary Education. The total number of participants is 215.

Material. *The web-based learning environment*. In this study, the open source software Moodle 2.0 ® (http://www.moodle.org) was used to create and administer: (1) the pre-test, (2) the course of the learning phase, and (3) the post-test. The testing and learning material (*i.e.*, items) consisted of fill-in exercises on French verb conjugation. Every item contained one example of the required verbal form, followed by the actual verb that the learner needed to conjugate. After completing an item, participants received explanatory feedback on the correct response. Each item had an associated item difficulty parameter value. The items were calibrated (*i.e.*, the items difficulty parameter value were estimated) by means of a conducted by SELOR (Selectie en Orientatie, is the official assessment center of the federal Belgian government that selects and tests candidate civil servants in Belgium). Items were calibrated using the Rasch model, based on the data from 2961 examinees. The examinees of SELOR completed the calibration study because the administered items are used to test the examinees proficiency of French verb conjugation. The examinees that successfully completed the test got promoted at the government. The examinees are not part of the current study.

Introduction and pre-test. All participants completed a proficiency test consisting of 25 fill-in items testing French verb conjugation. The test was not time-limited and the average time to complete the test was close to 20 min. The pre-test total scores ranged from 4 to 25 with a mean of 15.81 and standard deviation of 4.53.

To measure motivation, we adapted the Motivated Strategies for Learning Questionnaire (*i.e.*, MSLQ) developed by [57] so that this questionnaire would be applicable to French language

learning. Sample items include the following: (1) "For learning French connector words, I prefer tasks that really challenge me so I can learn new things"; (2) Understanding the use of French connector words is very important to me". The questionnaire consisted of 18 6-point Likert type items (1 = strongly disagree, 6 = strongly agree), divided over three scales: (1) self-efficacy and performance, (2) motivation, and (3) task value.

Based on the responses of the 215 study participants who filled out the questionnaire, we found that these scales are internally consistent (by calculating Cronbach's α, [58]): *intrinsic motivation*, consisting of four items ($\alpha = 0.732$), asking students why they are engaging in the learning task; *task value*, consisting of six items ($\alpha = 0.836$), asking students how interesting, important, and useful they find the task; and *self-efficacy and performance*, consisting of eight items ($\alpha = 0.937$), asking students for their expectancy for success and self-efficacy. The motivation questionnaire (as measured by the three subscales) was found to be reliable ($\alpha = 0.803$), with all subscales showing a positive correlation ($p < 0.01$). Both *intrinsic motivation* and *task value* are regarded as pre-experimental motivation/interest, while *self-efficacy and performance* was considered a separate scale.

Learning phase. For each combination of prior proficiency ($n = 6$) and difficulty condition ($n = 6$), random sets of 80 items were compiled. All items were on French verb conjugation and were scored binarily (1 for a correct response and 0 for an incorrect response). Learning phase total scores ranged from 21 to 76 with a mean equal to 55.54 and a standard deviation of 11.11.

Post-experimental phase. After the learning phase, all participants received 25 fill-in items on French verb conjugation, with equal content for all conditions. The post-test scores ranged from 5 to 25, with a mean equal to 16.59 and a standard deviation of 4.21. To measure post-experimental motivation, a translated version of the Intrinsic Motivation Inventory (IMI) [59] was used. We selected four relevant subscales. The questionnaire consisted of 25 6-point Likert type items (1 = strongly disagree, 6 = strongly agree) divided into four subscales that were found to be reliable in the present study (n = 215, Cronbach's α): *interest/enjoyment*, consisting of seven items ($\alpha = 0.924$), *perceived competence*, consisting of six items ($\alpha = 0.918$), *value/usefulness*, consisting of seven items ($\alpha = 0.923$), and *effort/importance*, consisting of five items ($\alpha = 0.853$). The motivation questionnaire (as measured by the four subscales) was found to be reliable ($\alpha = 0.797$), with all subscales showing a positive correlation ($p < 0.01$).

Procedure. *Introduction and pre-test.* During the pre-experimental phase, the participants first received a short introduction on the experiment. Subsequently they signed the informed consent, provided some background information and filled in the motivated strategies for the learning questionnaire. After completing the MSLQ, the participants completed the proficiency test consisting of 25 fill-in items.

Intermediate analysis. The proficiency of participants was assessed by applying the Rasch model (Equation (1)) on the participants' scores on the 25 fill-in items with the known difficulty of the proficiency test. Based on the resulting proficiency estimates, participants were grouped into six proficiency levels: $]-\infty;-1[$, $[-1;0[$, $[0;1[$, $[1;2[$, $[2;3[$, and $[3;\infty[$. Within each proficiency level, participants were randomly assigned to one of the six experimental conditions.

Learning phase. One week after the pre-test, participants completed 80 items during a learning phase. After each response, they received feedback on the correctness of their answer; at the same time, the correct response was provided [60].

Post-experimental phase. During the post-experimental phase the participants completed the post-test consisting of 25 items. Subsequently, they filled in the IMI. The total duration of the learning and post-experimental phase was approximately one hour and a half.

Data Analysis. The total number of students who completed the pre-test, learning phase and post-test in the experiment was 220. Participants with a score on the pre-test, learning set, or post-test of 3 SDs below or above the average score were also excluded from the analysis ($n = 4$). We choose 3 SDs as criterion for identifying outliers because scores deviating more than 3SDs from the mean are unlikely (*i.e.*, 0.3% of the scores if we assume a normal distribution [61].

All excluded participants had a score of more than 3 SDs below the average score (*i.e.*, $\overline{X}_{pre}$ = 15.77, $\overline{X}_{leer}$ = 54.95, $\overline{X}_{post}$ = 16.48), possibly due to a lack of effort those participants had put into the experiment. 45 out of the 215 (20.83%) study participants had missing values on either the MSLQ or IMI scale (completely at random), which was used for the post-experimental motivation analysis (*i.e.*, the effect of proficiency, prior motivation, and difficulty on post-experimental motivation). Instead of deleting the participants with missing values on these scales, we applied the regression-based multiple imputation technique (after investigating the percentage of missing data per variable and per case and investigating the pattern of missing values [62]). Values were imputed borrowing strength of the known values for the different variables in the dataset. A sensitivity analysis was conducted to investigate whether the imputation method had an effect on the results by comparing the results of using the regression-based imputation method with those using maximum likelihood estimation. Only small differences were found, and conclusions remained the same. Therefore, we report here the results of the regression-based imputation method.

Participants with a score on the different scales of MSLQ and IMI of more than 3 SDs below or above the average score were also excluded from this specific analysis (n = 1). The excluded participant had a score of more than 3 SDs below the average score on the subscale value/usefulness of the Intrinsic Motivation Inventory survey. In sum, a total number of 215 study participants were included in the analysis with imputed values on the MSLQ and IMI variables. Every difficulty condition included a similar number of participants: $n(VD)$ = 36; $n(D)$ = 34; $n(M)$ = 38; $n(E)$ = 37; $n(VE)$ = 39; $n(R)$ = 31).

The influence of the difficulty condition (*i.e.*, independent variable, grouping variable) on the learning outcome (measured as the difference between the post-test and pre-test score) controlling for self-efficacy, prior motivation, and proficiency (*i.e.*, the covariates) was investigated. Analysis of covariance (*i.e.*, ANCOVA) is the most recommended analysis method. Prior to the analysis, we tested the homogeneity assumption using Levene's test. Based on this, we concluded that the homogeneity assumption was not violated, $F(5,209) = 1.69$, $p = 0.14$. In addition, we evaluated the normality assumption by applying the Kolmogorov-Smirnov and Shapiro-Wilk test, and no significant deviations from normality were identified (Kurtosis statistic varies between -0.64 and 1.14 and Skewness between -0.230 and 0.230). In addition, ANCOVA was robust, as our group sizes are very similar, there are at least 20 degrees of freedom, and the smallest response category contained at least 20% of all responses [62].

A multivariate analysis of covariance (MANCOVA) was applied to investigate the influence of the grouping variable (difficulty level) and covariates (*i.e.*, prior motivation and self-efficacy) on multiple dependent variables (*i.e.*, the four subscales of post-intervention motivation). The following assumptions were evaluated: independence of observations, multivariate normality assumption using Kolmogorov-Smirnov and ShapiroWilk test, homogeneity of covariance matrix using Box's test, homogeneity of error variance using Levene's test, and the assumption of no multicollinearity. Ideally, the dependent variables are moderately correlated with each other. If correlations are low, it is better to run separate one-way ANOVAs; if the correlations are larger than 0.9, than there is such a strong multicollinearity that an analysis of one of the dependent variables is sufficient. Box's test of equality of covariance matrices indicates that there is no significant difference between the covariance matrix of the four dependent variables [Box's $M = 50.007$, $F(50, 44676.961) = 0.937$, $p = 0.601$]. In addition, Levene's test of equality of error variances indicates that the error variances are equal across the groups. The Kolmogorov-Smirnov test and Shapiro-Wilk test both indicate that the data are multivariate normally distributed. The correlation between the 4 dependent variables was found moderate in size (ranging from 0.305 to 661), supporting the choice for a MANCOVA.

5. Results

All analyses reported in the present study used a significance level of 0.05. The equality of conditions (VD, D, M, E, VE, and R) was ascertained for proficiency, as measured by the total score on the pre-test, $F(5,209) = 0.320$, $p = 0.904$, and prior motivation, $F(5,200) = 1.44$, $p = 0.211$.

This means that there was no systematic difference between the six conditions in terms of proficiency and prior motivation.

Manipulation check. A logistic regression analysis was conducted with the binary response on the learning phase as the dependent variable and five of the six difficulty groups (VD, D, M, E and VE) as the independent variable. The difficulty groups had a statistically significant effect on the outcome score [$t(5) = 558.35$, $p < 0.001$]. The random difficulty condition was not included in this analysis because, in this condition, a random mix of difficulty levels was presented. The proportion correct score for the learning phase by the difficulty condition and the confidence intervals of the mean proportion correct score for the learning phase by the difficulty condition can be found in Figure 1.

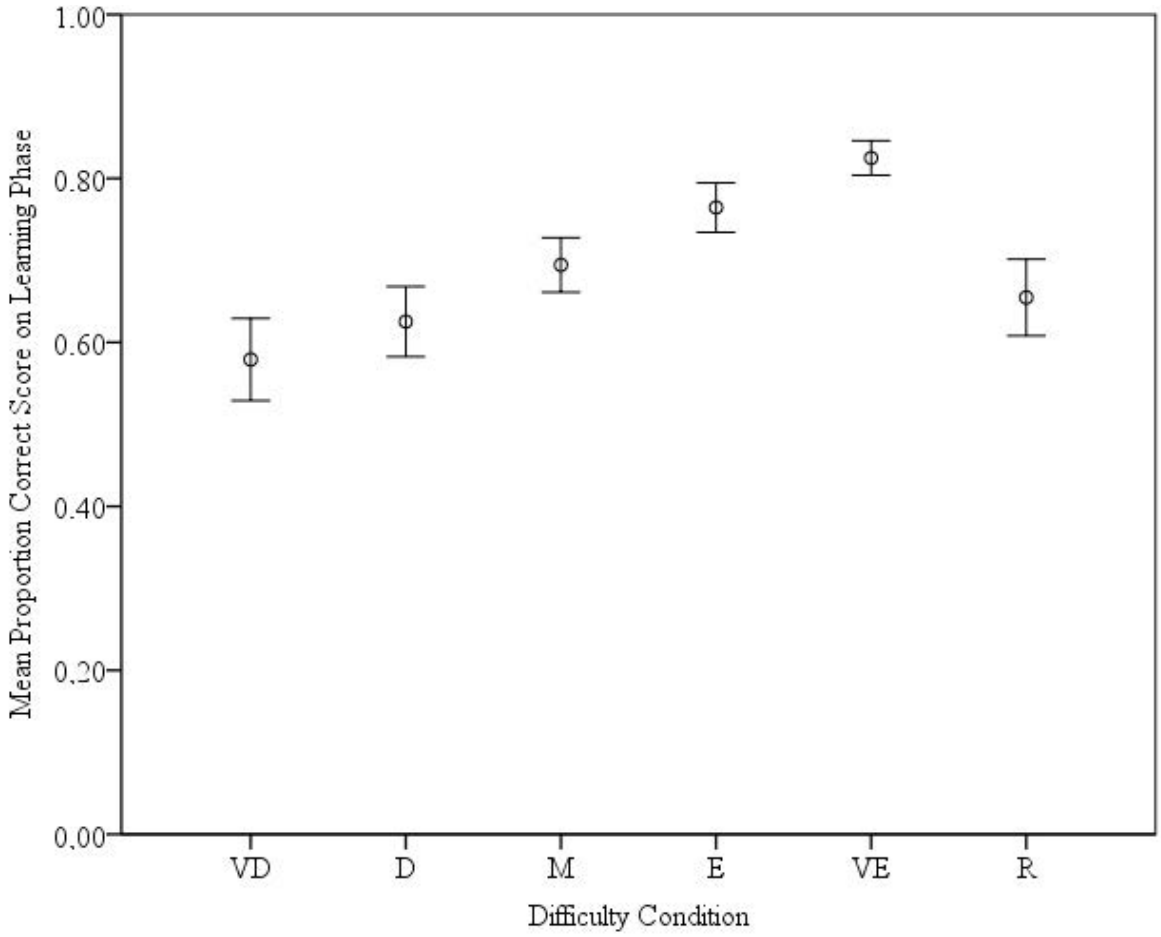

Figure 1. Mean proportion correct score for each difficulty condition. Error bars represent 95% confidence intervals. VD = Very Difficult; D = Difficult; M = Moderate; E = Easy; VE = Very Easy; R = Random.

Learning outcome. The mean of the pre-test ($\overline{X}_{pre} = 15.85$) was significantly lower than the mean of the post-test ($\overline{X}_{post} = 16.63$), $t(215) = -3.37$, $p = 0.001$. Since both tests were equally difficult (the true score at $\theta = 0.5$ is 12.504 for the pre-test and 12.505 for the post-test), the results suggest that learning occurred.

An ANCOVA with self-efficacy (measured by the MSLQ), prior motivation (*i.e.*, intrinsic motivation and task value measured by the MSLQ), and proficiency as covariates, and the difficulty condition as the independent variable (VD, D, M, E, VE, and R), was tested to explain the variances in learning outcome (*i.e.*, the difference between the score on the post-test and the score on the pre-test). Learner's self-efficacy score did not affect learning outcome $F(1,205) = 1.09$, $p = 0.297$, $\eta_p^2 = 0.003$. Prior motivation had a positive but small effect on learning outcome $F(1,205) = 0.07$, $p = 0.787$ $\eta_p^2 = 0.0002$.

Proficiency had a statistically significant affect on learning outcomes $F(1,205) = 83.43$, $p < 0.001$, $\eta_p^2 = 0.0282$. The difficulty condition had a non-significant effect on learning outcome, $F(5,205) = 1.34$, $p = 0.248$, $\eta_p^2 = 0.023$.

Post-experimental motivation. A MANCOVA was conducted with the four subscales of the post-experimental motivation questionnaire as dependent variables and self-efficacy, prior motivation, and proficiency as covariates, and the difficulty condition (VD, D, M, E, VE, and R) as the independent variable. Detailed statistics are provided in Table 2.

Table 2. Multivariate analyses of covariance (MANCOVA) assessing post-experimental motivation.

Independent	Dependent	Wilk's λ	*F*	df_1	df_2	*p*	η_p^2	*b*	*SE*
Self-efficacy	Post-experimental motivation	0.901	5.54	4	203	0.0003	0.0097		
	Interest/Enjoyment		24.26	1	213	<0.001	0.010	0.347	0.09
	Effort/Importance		7.03	1	213	0.001	0.032	0.174	0.07
	Perceived Competence		61.29	1	213	<0.001	0.223	0.489	0.06
	Value/Usefulness		9.90	1	213	0.002	0.044	0.191	0.06
Prior motivation	Post-experimental motivation	0.943	3.05	4	203	0.018	0.036		
	Interest/Enjoyment		23.50	1	213	<0.001	0.099	0.391	0.08
	Effort/Importance		16.18	1	213	<0.001	0.071	0.296	0.07
	Perceived Competence		28.63	1	213	<0.001	0.119	0.406	0.08
	Value/Usefulness		19.54	1	213	<0.001	0.084	0.300	0.008
Prior knowledge	Post-experimental motivation	0.901	5.57	4	203	<0.001	0.013		
	Interest/Enjoyment		7.37	1	213	0.007	0.033	0.051	0.019
	Effort/Importance		0.00	1	213	0.951	0.000	0.001	0.017
	Perceived Competence		46.79	1	213	<0.001	0.180	0.112	0.016
	Value/Usefulness		4.25	1	213	<0.001	0.0200	0.032	0.016
Difficulty	Post-experimental motivation	0.915	0.910	20	674.22	0.575	0.017		

Post-experimental motivation was significantly affected by prior motivation (Wilks' λ, $F(4,204) = 6.86$, $p < 0.001$, $\eta_p^2 = 0.063$). Higher prior motivation was associated with higher post-experimental Interest/Enjoyment ($b = 0.358$), higher Value/Usefulness ($b = 0.285$), and higher Effort/Importance ($b = 0.332$). Self-efficacy for learning and performance significantly affected post-experimental motivation (Wilks' λ, $F(4,203) = 5.54$, $p = 0.0003$, $\eta_p^2 = 0.009$). The higher the self-efficacy rating, the higher the perceived competence ($b = 0.332$). Proficiency significantly affected post-experimental motivation [Wilks' λ, $F(4,207) = 5.98$, $p < 0.001$, $\eta_p^2 = 0.013$]. The higher the participants' proficiency level, the higher the perceived competence ($b = 0.071$). The difficulty condition had no significant effect on post-experimental motivation [Wilks' λ, $F(20,674.22.) = 0.91$, $p = 0.575$].

Moderator effect of proficiency. A hierarchical multiple linear regression was conducted to determine whether the difficulty condition and the proficiency has a significant interaction effect on learning outcome. We wanted to investigate whether the effect of prior proficiency on learning outcome is dependent on different level of difficulty. The difficulty condition and proficiency were entered in Step 1, explaining 24.01% of the variance in the learning outcome scores. The predictive model for Step 1 was statistically significant, $F(6,208) = 10.96$, $p < 0.001$. After entering the interaction term at Step 2, the total variance explained by the model as a whole was 25.63%, $F(11, 203) = 6.36$, $p < 0.001$. The interaction term, therefore, hardly explains any additional variance in the learning outcome; the proportion only explained variance that changed with 0.016 [F (5, 203) = 0.995, $p = 0.422$]. Examination of the beta values highlighted the significant contribution of proficiency (b = −0.413, $p < 0.001$). This demonstrated that learning outcomes decrease as proficiency level increases. The interaction effect of the difficulty condition and proficiency ($b = 0.014$, $p = 0.601$) and the main effect of the difficulty condition ($b = -0.384$, $p = 0.394$) were non-significant. Detailed statistics of this hierarchical multiple regression are provided in Table 3.

Table 3. Testing Moderator Effect Using Hierarchical Multiple Regression.

Step/Predictor	Step 1			Step 2		
	b	***t***	***p***	***b***	***t***	***p***
1. Difficulty 1 [a]	1.252	1.69	0.093	0.389	0.14	0.888
Difficulty 2 [a]	0.831	1.11	0.270	0.389	0.14	0.523
Difficulty 3 [a]	1.461	2.00	0.047	−0.435	−0.15	0.879
Difficulty 4 [a]	1.280	0.736	0.083	1.739	0.62	0.537
Difficulty 5 [a]	1.012	1.39	0.165	−2.513	−0.89	0.378
Proficiency	−0.358	−7.77	<0.001	−0.413	−3.11	0.002
2. Difficulty 1 [a] × Proficiency				0.054	0.32	0.749
Difficulty 2 [a] × Proficiency				−0.069	−0.38	0.705
Difficulty 3 [a] × Proficiency				0.120	0.69	0.493
Difficulty 4 [a] × Proficiency				−0.027	−0.16	0.875
Difficulty 5 [a] × Proficiency				0.219	1.28	0.204
R^2		0.240			0.256	
ΔR^2		0.240			0.016	
ΔF		10.96			0.995	
Df		6.209			5.204	
p		<0.001			0.422	

[a] Difficulty 1, 2, 3, 4, and 5 are k-1 dummy variables for k different difficulty conditions in the study. Difficulty 1 = VD; Difficulty 2 = D; Difficulty 3 = M; Difficulty 4 = E; and Difficulty 5 = VE.

6. Discussion

In this study, we aimed at identifying the optimal item sequencing algorithm in item-based adaptive learning environments by disentangling the relationship between the item difficulty level and the learning outcome and motivation. As little experimental research has been conducted on evaluating the efficacy of item sequencing algorithms with varying item difficulty levels, this study tried to bridge the gap by evaluating six difficulty conditions in which participants had a varying

probability of answering an item correctly: (1) between 0.40 and 0.50; (2) between 0.50 and 0.60; (3) between 0.60 and 0.70; (4) between 0.70 and 0.80; (5) between 0.80 and 0.90; and (6) between 0.40 and 0.90. The six difficulty conditions were evaluated on learning outcome and motivation.

Results showed that the difficulty condition had no significant effect on either learning outcome or on motivation. Because the number of participants is relatively large, and the lack of significance, therefore, does not seem to be a consequence of lack of power, this finding suggests that, in item-based adaptive learning environments covering only one latent proficiency (in this study French verb conjugation), it makes no important difference whether you present items that are adapted to the learner's proficiency level or whether you select items randomly. As a consequence, Hypothesis (1), stating that items with a moderate difficulty will result in higher task involvement, higher interest, and higher perceived competence than when presenting more difficult items, and Hypothesis (2), stating that relatively easy items will result in lower task involvement (effort), could not be confirmed. Higher proficiency appeared to be predictive of lower learning outcomes, and this was independent of the difficulty condition. Therefore, Hypothesis (3), assuming that proficiency has a moderating effect on the relationship between the relative item difficulty level and learning outcome, could not be confirmed either.

Furthermore, no single difficulty condition maximized the learning outcome relative to others. Hence, the results provide empirical evidence for Brusilovky's statement [11] that simple curriculum learning does not benefit from adaptive sequencing compared to random question sequencing. Besides, the results are in line with the findings of [40], who found that adaptive task sequencing did not yield more efficient learning.

In this study, the learning outcome is measured by means of a post-test that was of approximately equal difficulty as the pre-test. Because both pre-test and post-test consisted of items from a calibrated item bank (*i.e.*, the item difficulty parameters are known and located on one continuous scale), the score on the pre-test and the post-test could be compared and could function as a measure of learning outcome. Other possible measurement methods, such as retention and time it takes to learn, have not been taken in to consideration. Because we randomly assigned the participants to experimental conditions and conditions only differed in the difficulty of the items in the learning phase, we can exclude the influence of confounding factors. However, we must be more prudent in interpreting the improvement of the average score from pre-test to post-test: Whereas this improvement suggests a positive effect of the learning phase, it is not excluded that the improvement is (partly) due to, for instance, study activities in the days between pre- and post-test.

Furthermore, it needs to be considered that the proportion correct score on the learning phase for each difficulty condition, and particularly for the more difficult conditions, were substantially higher than expected on the basis of the item difficulty parameter. This minor shortcoming in the difficulty manipulation might result in too small a distinction between the different difficulty conditions, leading to non-significant effects of the difficulty condition. Furthermore, this could explain why the difficulty condition was found to have no negative effect on post-experimental motivation. The difficult items might not have been difficult enough to have an adverse effect on motivation. Furthermore, a possible explanation for the high values of proportion correct score in the learning phase might be attributed to learning taking place in the learning phase.

The MSLQ [57] was used to measure prior motivation (interest—*i.e.*, intrinsic motivation and task value—and self-efficacy). The IMI [59] was used to measure post-experimental motivation (interest/enjoyment, perceived competence, effort/importance and value/usefulness). The two questionnaires contain distinct subscales; consequently, it would have been better to choose one questionnaire to present to the learners before and after the learning phase. Furthermore, asking the learner to rate their agreement with specific attitudes, beliefs, and activities is only one method to measure motivation. Future research could also focus on behavior in the learning environment as an indicator of motivation.

Because the literature reports inconsistent results with regard to the presence or absence of the relationship between difficulty level and learning outcome, future research should focus on inferring the specific characteristics of the learning environments in which this relationship does or does not hold. Besides, the grammar items in our study are non-authentic, while some researchers suggest that authentic tasks can speed up the learning of grammar [63]. Furthermore, simple knowledge of correct response feedback might not be enough to effectively promote learning. According to [40], elaborated feedback would ensure that the assessment itself is a valid learning experience. Besides examining the specific characteristics of the learning environments in which the studied relationship does or does not hold, future research may also consider different item selection algorithms. In this article, we explored an item selection algorithm that is comparable to the item selection algorithm in CAT. Other procedures of sequencing the items in an item-based learning environment are available, but require further investigation, such as alternating between relatively difficult and relatively easy items or incorporating a moving window as in the moving test approach [64,65].

In sum, this study provides initial evidence as to whether a particular relative item difficulty level is more effective than others in terms of maintaining learner motivation and, in turn, enhancing learning outcome in an item-based learning environment. Findings indicate that, for learning environments consisting of simple questions (*i.e.*, questions dealing with one proficiency) provided with knowledge of the correct response, it does not matter whether we present easy, moderate, or difficult items or whether we present the items with a random mix of difficulty levels. This research may instigate further examination, which could take other characteristics of the learner and the learning environment into consideration.

Author Contributions: Mariola Moeyaert and Kelly Wauters analyzed the data and wrote the paper. Kelly Wauters, Wim Van den Noortgate, and Piet Desmet designed and performed the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Motivated Strategies for Learning Questionnaire (*i.e.*, MSLQ), by [57]

Motivated Strategies for Learning Questionnaire	Totally Disagree					Totally Agree
1. I prefer challenging tasks so I can learn new things.	1	2	3	4	5	6
2. What I learned from this task will be useful for other courses.	1	2	3	4	5	6
3. I believe I understand the basics of this topic.	1	2	3	4	5	6
4. I expect to perform well on this task.	1	2	3	4	5	6
5. I prefer harder tasks that make me curious even though this may mean that those tasks are harder to study.	1	2	3	4	5	6
6. I believe that I can perform well on a test about French conjugation.	1	2	3	4	5	6
7. It is important to me to study the content of this topic.	1	2	3	4	5	6
8. I am convinced that I will be able to understand the most difficult content of this topic.	1	2	3	4	5	6
9. Mastering this topic as thoroughly as possible makes me most satisfied.	1	2	3	4	5	6
10. I believe that I will be able to understand the most complex study materials concerning this topic.	1	2	3	4	5	6
11. I am very interested in the content of this course.	1	2	3	4	5	6
12. I believe that I will obtain an excellent grade after the learning phase.	1	2	3	4	5	6
13. Learning this topic of the course is important.	1	2	3	4	5	6
14. When given the opportunity, I will choose the tasks that enable me to learn the most.	1	2	3	4	5	6
15. I like the topic of the course.	1	2	3	4	5	6
16. I am convinced that I will become proficient in this topic.	1	2	3	4	5	6
17. Understanding this topic is very important for me.	1	2	3	4	5	6
18. If I consider the difficulty level of this topic, the type of instruction and my own proficiency, then I will perform well for this course.	1	2	3	4	5	6

Appendix B

Intrinsic Motivation Inventory (*i.e.*, IMI) [59]

Intrinsic Motivation Inventory	Totally Disagree					Totally Agree
1. Overall, I liked the task.	1	2	3	4	5	6
2. I think my performance on this task is better compared to peers.	1	2	3	4	5	6
3. This task was useful for learning French verb conjugation.	1	2	3	4	5	6
4. By working on this task, I felt more competent in French verb conjugation.	1	2	3	4	5	6
5. I put a lot of effort into this task.	1	2	3	4	5	6
6. While conducting the task, I realized that I really liked the task.	1	2	3	4	5	6
7. The task helped me to master French verb conjugation.	1	2	3	4	5	6
8. I did not put a lot of effort in completing the task.	1	2	3	4	5	6
9. The task was fun.	1	2	3	4	5	6
10. The task was difficult.	1	2	3	4	5	6
11. This task can be useful for me in the future.	1	2	3	4	5	6
12. I could not concentrate while conducting the task.	1	2	3	4	5	6
13. This task is important.	1	2	3	4	5	6
14. I think I did well on the task.	1	2	3	4	5	6
15. I am happy with the results of the task.	1	2	3	4	5	6
16. I did not need to put a lot of effort into this task to perform well.	1	2	3	4	5	6
17. I would be willing to redo the task because of its usefulness.	1	2	3	4	5	6
18. This task was boring.	1	2	3	4	5	6
19. It was important for me to perform well on this task.	1	2	3	4	5	6
20. I would describe this task as very interesting.	1	2	3	4	5	6
21. I believe that by completing this task, my performance will be enhanced.	1	2	3	4	5	6
23. I did the best I could for this task.	1	2	3	4	5	6
24. This task is very important because it can enhance my French conjugation proficiency.	1	2	3	4	5	6
25. I found this task fairly fun.	1	2	3	4	5	6

References

1. Brusilovsky, P. Adaptive hypermedia: From intelligent tutoring systems to Web-based education (Invited talk). In *Intelligent Tutoring Systems. Lecture Notes in Computer Science*; Gauthier, G., Frasson, C., VanLehn, K., Eds.; Springer-Verlag: Berlin, Germany, 2000; pp. 1–7.
2. Perry, E.H.; Pilati, M.L. Online learning. *New Direct. Teach. Learn.* **2011**, *128*, 95–104. [CrossRef]
3. Brusilovsky, P.; Millán, E. User models for adaptive hypermedia and adaptive educational systems. In *the Adaptive Web*; Brusilovsky, P., Kobsa, A., Nejdl, W., Eds.; Springer-Verlag: Berlin, Germany, 2007; pp. 3–53.
4. Jeremić, Z.; Jovanović, J.; Gašević, D. Student modeling and assessment in intelligent tutoring of software patterns. *Exp. Syst. Appl.* **2012**, *39*, 210–222. [CrossRef]
5. Holden, H.K. Understanding current learner modeling approaches. In *Design Recommendations for Adaptive Intelligent Tutoring Systems*; Sottilare, R.A., Graesser, A.C., Hu, X., Holden, H., Eds.; U.S. Army Research Laboratory: Adelphi, MD, USA, 2013.
6. Brusilovsky, P. Adaptive and intelligent technologies for Web-based education. *Künstliche Intell.* **1999**, *13*, 19–25.
7. Wauters, K.; Desmet, P.; van den Noortgate, W. Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *J. Comput. Assist. Learn.* **2010**, *26*, 549–562. [CrossRef]
8. Mödritscher, F. Adaptive e-learning. In *Adaptive E-Learning Environments: Theory, Practice, and Experience*; Mödritscher, F., Ed.; Verlag Dr. Müller: Saarbrucken, Germany, 2008; pp. 45–60.

9. Peachey, D.; McCalla, G. Using planning techniques in intelligent tutoring systems. *Int. J. Man-Mach. Stud.* **1986**, *24*, 77–98. [CrossRef]
10. De Bra, P.; Smits, D.; Stash, N. Creating and Delivering Adaptive Courses with AHA! In *Lecture Notes in Computer Science: Vol. 4227. Innovative Approaches for Learning and Knowledge Sharing*; Nejdl, W., Tochtermann, K., Eds.; Springer-Verlag: Berlin, Germany, 2006; pp. 21–33.
11. Conati, C.; Gertner, A.; VanLehn, K. Using Bayesian networks to manage uncertainty in student modeling. *J. User Model. User-Adapt. Interact.* **2002**, *12*, 371–417. [CrossRef]
12. Brusilovsky, P. A framework for intelligent knowledge sequencing and task sequencing. In *Lecture Notes in Computer Science: Vol. 608. Intelligent Tutoring Systems*; Frasson, C., Gauthier, G., McCalla, G., Eds.; Springer-Verlag: Berlin, Germany, 1992; pp. 499–506.
13. Chang, X.; Zheng, Q. Knowledge element extraction for knowledge-based learning resources organization. In *Lecture Notes in Computer Science: Vol. 4823. Advances in Web Based Learning—ICWL 2007*; Leung, H., Li, F., Lau, R., Li, Q., Eds.; Springer-Verlag: Berlin, Germany, 2008; pp. 102–113.
14. Qinghua, Z.; Li, L.; Tian, F.; Ding, J. Adaptive english reading articles recommendation system using IRT. In Proceedings of the International Conference on Web-Based Learning, Edinburgh, UK, 15–17 August 2007.
15. Van der Linden, W.J.; Hambleton, R.K. *Handbook of Modern Item Response Theory*; Springer: New York, NY, USA, 1997.
16. Rasch, G. *Probabilistic Models for some Intelligence and Attainment Tests*; University of Chicago Press: Chicago, IL, USA, 1960.
17. Liu, H.; You, X.; Wang, W.; Ding, S.; Chang, H.-H. The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *J. Classif.* **2013**, *30*, 152–172. [CrossRef]
18. Van der Linden, W.J.; Glas, C.A.W. *Computerized Adaptive Testing: Theory and Practice*; Kluwer: Norwell, MA, USA, 2000.
19. Wang, F.H. A fuzzy neural network for item sequencing in personalized cognitive scaffolding with adaptive formative assessment. *Expert Syst. Appl.* **2004**, *27*, 11–25. [CrossRef]
20. Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*; Harper Perennial: New York, NY, USA, 1991.
21. Deci, E.L.; Ryan, R.M. *Intrinsic Motivation and Self-Determination in Human Behavior*; Plenum Press: New York, NY, USA, 1985.
22. Schweinle, A.; Turner, J.C.; Meyer, D.K. Striking the right balance: Students' motivation and affect in elementary mathematics. *J. Educ. Res.* **2006**, *99*, 271–293. [CrossRef]
23. Van Velsor, E.; McCauley, C.D. Our view of leadership development. In *The Center for Creative Leadership: Handbook of Leadership Development*; McCauley, C.D., van Velsor, E., Eds.; Jossey-Bass: San Francisco, CA, USA, 2004; pp. 1–22.
24. Stacey, K.; Sonenberg, E.; Nicholson, A.; Boneh, T.; Steinle, V. A teaching model exploiting cognitive conflict driven by a Bayesian Network. In Proceedings of the Ninth International Conference on User Modeling (UM2003), Johnstown, PA, USA, 18 June 2003.
25. Pintrich, P.R.; Schunk, D.H. *Motivation in Education: Theory, Research, and Applications*; Merrill Prentice Hall: Upper Saddle River, NJ, USA, 2002.
26. Eccles, J.S.; Wigfield, A. In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personal. Soc. Psychol. Bull.* **1995**, *21*, 215–225. [CrossRef]
27. Shernoff, D.J.; Csikszentmihalyi, M.; Schneider, B.; Shernoff, E.S. Student engagement in high school classrooms from the perspective of flow theory. *Sch. Psychol. Q.* **2003**, *18*, 158–176. [CrossRef]
28. Csikszentmihalyi, M.; Nakamura, J. The dynamics of intrinsic motivation: A study of adolescents. In *Research on Motivation in Education, Vol. 3: Goals and Cognitions*; Ames, C., Ames, R., Eds.; Academic Press: New York, NY, USA, 1989.
29. Fulmer, S.M.; Frijters, J.C. Motivation during an excessively challenging reading task: The buffering role of relative topic interest. *J. Exp. Educ.* **2011**, *79*, 185–208. [CrossRef]
30. Deci, E.L.; Ryan, R.M. A motivational approach to self: Integration in personality. In *Perpectives on Motivation. Nebraska Symposium on Motivation, 1990*; Dienstbier, R.A., Ed.; University of Nebraska Press: Lincoln, NE, USA, 1991; pp. 237–288.
31. Schiefele, U. Interest, learning, and motivation. *Educ. Psychol.* **1991**, *26*, 299–324. [CrossRef]
32. Hidi, S.; Renninger, K.A. The four-phase model of interest development. *Educ. Psychol.* **2006**, *41*, 111–127. [CrossRef]

33. Inoue, N. Why face a challenge?: The reason behind intrinsically motivated students' spontaneous choice of challenging tasks. *Learn. Individ. Differ.* **2007**, *17*, 251–259. [CrossRef]
34. Ainley, M.; Hidi, S.; Berndorff, D. Interest, learning, and the psychological processes that mediate their relationship. *J. Educ. Psychol.* **2002**, *94*, 545–561. [CrossRef]
35. Belanich, J.; Sibley, D.; Orvis, K.L. *Instructional Characteristics and Motivational Features of a PC-Based Game*; Research Report No. 1822; U.S. Army Research Institute for the Behavioral and Social Sciences: Arlington, VA, USA, 2004.
36. Malone, T.W.; Lepper, M.R. Making learning fun: A taxonomy of intrinsic motivations for learning. In *Aptitude, Learning and Instruction*; Snow, R.E., Farr, M.J., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1987; pp. 223–253.
37. Paas, F.; Tuovinen, J.E.; van Merriënboer, J.J.G.; Darabi, A.A. A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educ. Technol. Res. Dev.* **2005**, *53*, 25–34. [CrossRef]
38. Orvis, K.A.; Horn, D.B.; Belanich, J. The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Comput. Hum. Behav.* **2008**, *24*, 2415–2433. [CrossRef]
39. Colquitt, J.A.; LePine, J.A.; Noe, R.A. Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *J. Appl. Psychol.* **2000**, *85*, 678–707. [CrossRef] [PubMed]
40. Shute, V.J.; Hansen, E.G.; Almond, R.G. Evaluating ACED: The impact of feedback and adaptivity on learning. In *Artificial Intelligence in Education—Building Technology Rich Learning Contexts that Work*; Luckin, R., Koedinger, K., Greer, J., Eds.; IOS Press: Amsterdam, The Netherlands, 2007; pp. 230–237.
41. Camp, G.; Paas, F.; Rikers, R.; van Merriënboer, J. Dynamic problem selection in air traffic control training: A comparison between performance, mental effort and mental efficiency. *Comput. Hum. Behav.* **2001**, *17*, 575–595. [CrossRef]
42. Corbalan, G.; Kester, L.; van Merriënboer, J.J.G. Selecting learning tasks: Effects of adaptation and shared control on efficiency and task involvement. *Contemp. Educ. Psychol.* **2008**, *33*, 733–756. [CrossRef]
43. Salden, R.J.C.M.; Paas, F.; Broers, N.J.; van Merriënboer, J.J.G. Mental effort and performance as determinants for dynamic selection of learning tasks in air traffic control training. *Instr. Sci.* **2004**, *32*, 153–172. [CrossRef]
44. Chen, C.M.; Duh, L.J. Personalized Web-based tutoring system based on fuzzy item response theory. *Expert Syst. Appl.* **2008**, *34*, 2298–2315. [CrossRef]
45. Chen, C.M.; Lee, H.M.; Chen, Y.H. Personalized e-learning system using item response theory. *Comput. Educ.* **2005**, *44*, 237–255. [CrossRef]
46. Betz, N.E.; Weiss, D.J. Validity. In *Handbook of Measurement and Evaluation in Rehabilitation*; Bolton, B., Ed.; University Park Press: Baltimore, MD, USA, 1976; pp. 39–60.
47. Clarke, T.; Ayres, P.; Sweller, J. The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educ. Technol. Res. Dev.* **2005**, *53*, 15–24. [CrossRef]
48. Kalyuga, S.; Sweller, J. Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educ. Technol. Res. Dev.* **2005**, *53*, 83–93. [CrossRef]
49. Schnotz, W.; Rasch, T. Enabling, facilitating, and inhibiting effects of animations in multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educ. Technol. Res. Dev.* **2005**, *53*, 47–58. [CrossRef]
50. Tobias, S. Interest, prior knowledge, and learning. *Rev. Educ. Res.* **1994**, *64*, 37–54. [CrossRef]
51. Barla, M.; Bieliková, M.; Ezzeddinne, A.B.; Kramár, T.; Šimko, M.; Vozár, O. On the impact of adaptive test question selection for learning efficiency. *Comput. Educ.* **2010**, *55*, 846–857. [CrossRef]
52. Mitrovic, A.; Martin, B. Evaluating adaptive problem selection. In *Lecture Notes in Computer Science: Vol. 3137. Adaptive Hypermedia and Adaptive Web-Based Systems*; Nejdl, W., De Bra, P., Eds.; Springer-Verlag: Berlin, Germany, 2004; pp. 185–194.
53. Razzaq, L.; Heffernan, N.T. Scaffolding *vs.* hints in the ASSISTment System. In *Lecture Notes in Computer Science: Vol. 4053. Intelligent Tutoring Systems*; Ikeda, M., Ashley, K.D., Chan, T.W., Eds.; Springer-Verlag: Berlin, Germany, 2006; pp. 635–644.
54. Razzaq, L.; Heffernan, N.T.; Lindeman, R.W. What level of tutor interaction is best? In *Artificial Intelligence in Education—Building Technology Rich learning Contexts that Work*; Luckin, R., Koedinger, K., Greer, J., Eds.; IOS Press: Amsterdam, The Netherlands, 2007; pp. 222–229.

55. VanLehn, K.; Graesser, A.C.; Jackson, G.T.; Jordan, P.; Olney, A.; Rose, C.P. When is reading just as effective as one-on-one interactive human tutoring? In Proceedings of the 27th Annual Meeting of the Cognitive Science Society, Stresa, Italy, 21–23 July 2005; pp. 2259–2264.
56. Jameson, A. Adaptive interfaces and agents. In *Human-Computer Interaction Handbook*, 2nd ed.; Jacko, J.A., Sears, A., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2006; pp. 305–330.
57. Pintrich, P.R.; Smith, D.A.F.; Garcia, T.; McKeachie, W.J. *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*; National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan: Ann Arbor, MI, USA, 1991.
58. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. [CrossRef]
59. McAuley, E.; Duncan, T.; Tammen, V.V. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Res. Q. Exerc. Sport* **1987**, *60*, 48–58. [CrossRef] [PubMed]
60. Narciss, S. The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Exp. Psychol.* **2004**, *51*, 214–228. [CrossRef] [PubMed]
61. Howell, D.C. *Statistical Methods in Human Sciences*; New York: Wadsworth, OH, USA, 1998.
62. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*; The National Academies Press: Washington, DC, USA, 2010.
63. Embong, A.M.; Abdullah, M.R.T.L.; Yaacob, R.A.I.R.; Noor, A.M.; Abdullah, A. The sustainability of utilizing the authentic materials in English classroom: From the perspective of the theory of learning. *Int. J. Basic Appl. Sci.* **2011**, *11*, 132–134.
64. Leutner, D. Das testlängendilemma in der lernproweßbegleitenden wissensdiagnostik (The test-length dilemma in assessing knowledge during learning). *Z. Pädagogische Psychol.* **1992**, *6*, 233–238.
65. Leutner, D. Instructional design principles for adaptivity in open learning environments. In *Curriculum, Plans, and Processes in Instructional Design: International*; Seels, N.M., Dijkstra, S., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2004; pp. 289–308.