applied sciences

MDPI

Article

Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks

Younyoung Choi [1,*] **and Cayce McClenen** [2]

[1] Department of Adolescent Coaching Counseling, Hanyang Cyber University, Seoul 04763, Korea
[2] Department of Computer Science, McGill University, Montreal, QC H3A 1G1, Canada; cayce.mcclenen@mail.mcgill.ca
* Correspondence: 1120008@hycu.ac.kr

Received: 8 October 2020; Accepted: 16 November 2020; Published: 19 November 2020

Abstract: Online formative assessments in e-learning systems are increasingly of interest in the field of education. While substantial research into the model and item design aspects of formative assessment has been conducted, few software systems embodied with a psychometric model have been proposed to allow us to adaptively implement formative assessments. This study aimed to develop an adaptive formative assessment system, called computerized formative adaptive testing (CAFT) by using artificial intelligence methods based on computerized adaptive testing (CAT) and Bayesian networks as learning analytics. CAFT can adaptively administer personalized formative assessment to a learner by dynamically selecting appropriate items and tests aligned with the learner's ability. Forty items in an item bank were evaluated by 410 learners, moreover, 1000 learners were recruited for a simulation study and 120 learners were enrolled to evaluate the efficiency, validity, and reliability of CAFT in an application study. The results showed that, through CAFT, learners can adaptively take item s and tests in order to receive personalized diagnostic feedback about their learning progression. Consequently, this study highlights that a learning management system which integrates CAT as an artificially intelligent component is an efficient educational evaluation tool for a remote personalized learning service.

Keywords: computerized adaptive testing; formative assessment; learning management systems; artificial intelligence (AI) in education; e-learning technologies; learning analytics

1. Introduction

The integration of artificial intelligence into an online formative assessment system using modern measurement techniques directly benefits the adaptive collection of individual personalized information. Specifically, instead of giving all students the same test in online formative assessment, tests can be administered adaptively to each learner in terms of the learner's characteristics. This adaptive assessment system makes it possible to efficiently collect personalized diagnostic information as well as to tailor a test with respect to a student's ability, eventually offering a meaningful e-learning system [1,2].

In recent decades, many online adaptive formative assessments have been developed [3]. For example, SIETTE is one of the web-based learning systems that can provide an adaptive testing based on a learner model grounded on the leaner's response to previous questions. Additionally, COMPASS is a popular web-based learning system that can offer adaptive formative testing, tutoring, and feedback. Even though these systems adaptively select exams in terms of a learner's characteristics and ability, there is little research confirming that online formative assessments function based on accurately estimated evidence from statistical methods. Moreover, the assessments have been more focused on the evaluation of learning instead of “for learning”. Recently, there has been a big movement in the evaluation to include process-, formative-, and diagnostic- based assessment. However, most web-based formative assessment systems offer very little insight regarding the process and diagnostic purposes for learning based on what individual students should be provided and what actually does improve their academic ability for learning [4].

The concept of formative assessment is that a student’s learning status and progression are assessed during instruction as well as at the end of the course [5–7]. Therefore, the focus of assessment shifts toward evaluating students’ learning progress rather than just their final achievements. Formative assessment systems provide information about the knowledge, skills, and abilities (KSAs) that a student has reached and the learning trajectory of a student over time [8,9]. Recently, the use of formative assessment has been expanded to provide diagnostic information for reducing student weaknesses by identifying a gap between actual student levels and desired levels of performance in school [10,11]. Additionally, it can provide evidence about the change in a student’s ability level beyond that of other methods which only monitor the general proficiency of a student [8]. Due to these educational benefits, formative assessment has been popularly used in educational fields to help instructors monitor their students’ learning progression, select instructional strategies, and utilize alternative instructional approaches.

However, due to the COVID-19 pandemic, education institutions around the world have been closed and exams are temporarily suspended. Since teachers cannot interact with students in the classroom, they are not able to evaluate students face to face. This global crisis makes many conventional educational instruction and evaluation methods useless. Specifically, educators are not able to promptly evaluate students and provide feedback about their learning status and progression [9]. Consequently, online formative assessment systems using e-mail, instant messaging platforms, and online computer-based educational tools are increasingly of interest in the field of education. One of the benefits of online formative assessment tools is that they provide real-time updates on a student’s learning progression throughout a course [12]. Such information can offer quick remedial actions and feedback to students, instructors, and curriculum developers remotely [13,14].

Online formative assessment systems, web-based adaptive learning systems such as intelligent tutoring systems, and e-learning management systems are increasingly of interest in the field of education since the COVID-19 outbreak. The success of an adaptive learning system is fundamentally grounded on accurate information about what a student has learned and knows about specific concepts [15,16]. As such, a valid and meaningful adaptive learning system necessitates accurate assessment of a student’s ability and diagnosing what concepts a student knows and has learned. To achieve this, the adaptive assessment system must be able to interact with the learning system.

Computerized adaptive testing (CAT) provides a customized item set by dynamically selecting appropriate items aligned with the learner’s ability. CAT has been used for one-time testing such as in selection assessments or certificate exams [17]. With CAT, each item of a test is exposed to a learner depending on their response to the previous item, so the items are adaptively administered to learners in terms of their abilities. Therefore, the items administered to different learners are unique and depend on the learner’s ability [18,19]. Figure 1 shows the procedure of CAT. In order to implement CAT, all items should first be calibrated by item response theory (IRT) [20]. Various IRT models can be used for estimating item characteristics such as item difficulty and item discrimination in order to build an item bank. Once the item characteristics including item features, difficulty, and discrimination, have been estimated using various statistical estimation methods, the items are stored in the item bank along with their information. When administering a test using CAT, the process starts with selecting the first item from the item bank. There are several statistical methods proposed for selecting the first item from an item bank such as a random or Bayesian method. In the next step, an examinee’s ability and the standard error of measurement of the examinee’s ability are estimated based on their response to the first item using IRT. After this, the next item is adaptively selected from the item bank based on the previously estimated examinee's ability and the computed standard error of measurement. In more detail, CAT selects the item that maximizes the discriminatory and informative values for the given examinee's ability. For example, if an examinee answers correctly, the next item will be slightly more difficult than the previous one, however, if an examinee answers incorrectly, the next item will be slightly less difficult than the previous one. This procedure is repeated until the stopping criterion has been met. The stopping criterion of this iterative process can be (1) a certain level of standard error of measurement for a learner's ability, which is about measurement accuracy or precision or (2) a pre-set maximum number of items [21].

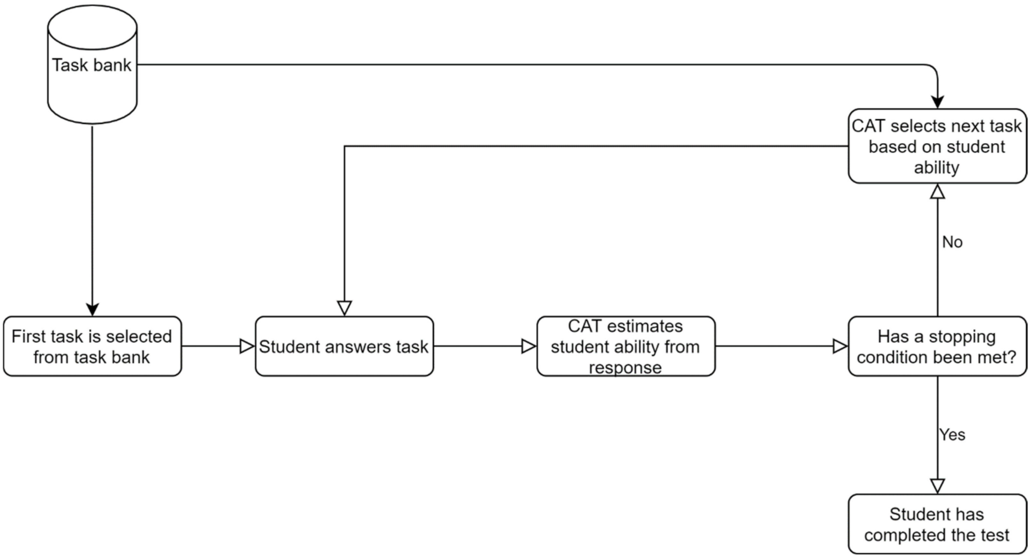

Task bank
CAT selects next task based on student ability
No
First task is selected from task bank
Student answers task
CAT estimates student ability from response
Has a stopping condition been met?
Yes
Student has completed the test

Figure 1. The procedure of computerized adaptive testing (CAT).

Consequently, by integrating CAT technology into a formative assessment system, the formative assessment can be adaptively implemented in terms of a learner's characteristics. By combining the statistical and mathematical properties of CAT with a formative assessment system, a new assessment system is created that can adaptively measure not only what a student knows, but also what a student's learning progression is over multiple measurement time points. From a diagnostic perspective, it offers information about which parts of a curriculum are difficult for a given student to learn as well as how well the student is doing throughout a course. Furthermore, since all students do not need to answer the same items- rather, an adaptive number of items are administered in terms of their abilities based on the CAT algorithm-, testing can be more efficient, accurate, and informative [22–24].

Historically, most CAT applications have been used for one-time testing instead of for multiple measurements over time such as in formative assessments. As such, CAT has not been applied for estimating the learning progression over multiple measurement time points from a longitudinal perspective. In order to develop an adaptive formative assessment system, a psychometric method that can analyze learning progression over multiple time points should be employed along with the CAT system. Many psychometric models have been proposed for measuring proficiency change over time. Proficiency change as a continuous variable is often expressed as quantitative growth modeled by means of latent growth curve approaches. In this case, quantitative growth can be defined in terms of an increase or decrease in the amount of knowledge or ability. In contrast, movement between stages or stage sequential changes are often described by qualitative growth. A typical example of qualitative growth is Piaget's model based on the cognitive development of children. Qualitative growth is measured by the critical pinpoints that represent a qualitatively different way of thinking and doing. With formative systems, the curriculum is modeled through several discrete learning stages, and a change in the student's ability through the stages is considered as qualitative growth. Furthermore, the notion of formative assessment was initially based on the concept of "mastery learning," in which students do not progress to the next learning objective until they have mastered the current one; hence, learning progression based on formative assessment consists of several learning stages and measures sequential stage change linked to the curriculum and instructions provided. Therefore, analytic models require the estimation of qualitative level changes over time, where item design and content domain theory provide a theoretical framework for creating and modeling observable evidence. For this purpose, dynamic Bayesian networks (DBNs) offer a promising approach [25], as they have been used in intelligent tutoring systems, game-based learning systems, and simulation-based learning [26,27]. A DBN is a probability based statistical modeling framework, which can make inferences about the previous, current, and future states of a student's learning progression over a specific period of time. Therefore, computerized formative adaptive testing (CAFT) utilizes the adaptive system of CAT and the probabilistic qualitative growth modeling from DBNs.

2. Research Objectives

The purpose of this study was to introduce a framework for an adaptive formative assessment system using computerized adaptive testing (CAT) and dynamic Bayesian networks (DBNs) that can adaptively assess a learner's ability over multiple measurement time points. Since previously established online formative assessment systems administer fixed, identical tests to learners, the adaptive function has not yet been implemented. The adaptation is achieved with several statistical algorithms from CAT. The fundamental concepts of the adaptation in CAT are (1) A learner's ability and the standard error of measurement about the learner's ability are estimated based on IRT, and (2) CAT selects the item from the item bank that maximizes the discriminatory and informative values for the given examinee's ability from the item bank. This procedure is repeated until a certain level of measurement precision regarding a learner's ability is satisfied. Since CAT chooses the next item based on the learner's answer to the previous item, all learners with different abilities take different items adaptively. The length and path of the test vary in terms of a learner.

CAFT is an assessment system created by combining CAT and DBNs. The adaptive system of CAT is applied at the item and test levels over multiple time points by adaptively selecting the next item of a test and the next test in sequential testing, respectively. In addition, DBNs offer real-time updates on the estimation of a learner's ability across multiple tests. The feedback system between CAT as an adaptive selection method and DBNs as a real-time estimation method is the core goal of CAFT as an integration of artificial intelligence methods into an e-learning system. Therefore, this study developed CAFT with several layers. The first layer is a test generation system, which contains (1) an instructor that generates items and publishes tests from an item bank and (2) calibration of item characteristics including item difficulty and discrimination. This layer provides basic storage for an adaptive formative assessment. The second layer is a CAT engine, the parameters of which a user can customize through a graphical interface in order to choose different item/test selection statistical algorithms. The third layer consists of DBNs, which estimate a learner's past, current, and future abilities across different measurement time points. This offers probabilistic real-time updates of a learner's ability through interaction with the CAT engine.

Therefore, in this study, first, we developed an adaptive formative assessment system using CAT and DBNs, called CAFT. After this, the validity, reliability, and efficiency of CAFT were evaluated by simulation study and application studies. In the simulation study, the learners' estimated abilities and the simulated learner's true abilities were compared for evaluation of the validity and reliability. The number of items used in CAFT was also measured in order to compare how many items are required to reach the same level of measurement precision under CAFT as when the full item set was used. Additionally, an application study was conducted for examining the reliability and efficiency of the system using data collected from the introduction to statistics course of an online university.

3. Materials and Methods

3.1. Materials

Figure 2 presents a high-level abstraction of the adaptive formative assessment system using the CAT procedure. Basically, there are five layers including (1) the item generation system: educators generate items and build several tests in any discipline. Item banks may only contain items from the same discipline, and educators can authorize tests for a particular discipline by assembling items from the respective item bank. The sequential collection of such tests serves as formative assessment across measurement time points. In addition, this layer conducts calibration of the item bank. (2) The CAT system: this system adaptively selects an item from the item bank based on the examinee's response to the previous item through several different statistical algorithms. (3) The formative assessment system: an instructor publishes a set of tests for formative assessment during their instruction of a course. This system adaptively selects a test from the set for examinees based on their current ability level across measurement time points aligned with the instruction or curriculum. (4) The real-time estimation system: this system estimates an individual learner's past, current, and future abilities and learning paths in real-time using DBNs. (5) The delivery system: this system reports diagnostic feedback about a student's learning progression including the student's current level and how their level has changed during the course. This information can be linked to the adaptive learning system for adaptively choosing an instructional strategy and learning materials.

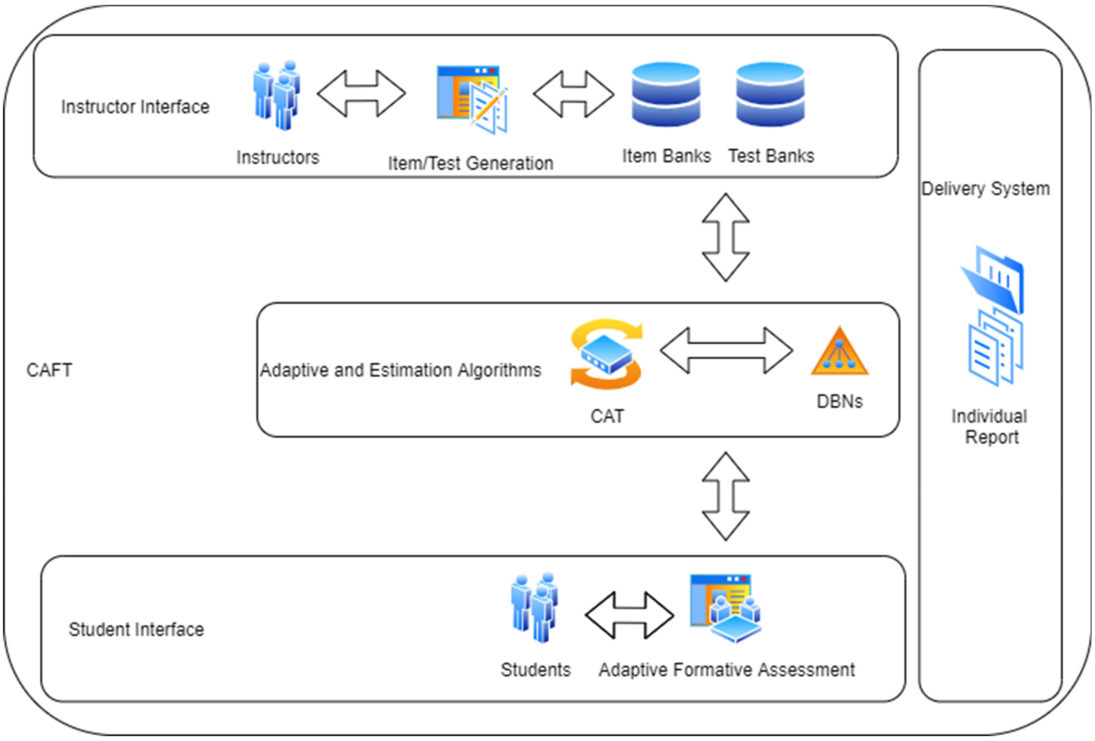


Figure 2. A system of computerized formative adaptive testing (CAFT) architecture.

3.2. Calibartion and Construction of the Item Bank for a Simulation Study and an Application Study

Forty items were generated for assessing two sample t-tests in an introduction to statistics course. The content objectives of the items were (1) knowing how to make a statistical research question related to t-tests, (2) understanding the basic statistical properties and assumptions about t-tests, (3) knowing the procedure of conducting a hypothesis testing of t-tests, and (4) applying and understanding the interpretation of t-test results. The quality of the item bank was evaluated before conducting a simulation and application study. Then, 410 samples were recruited from an introduction statistics course at an online university, 69% of which were females and 31% were males. The average age was 42.1 with a standard deviation of 6.9 and a range of 20–58.

The psychometric properties of all items in the item bank were assessed by (1) unidimensionality, (2) local independence, and (3) item fit. The unidimensionality as one of the assumptions in the IRT was evaluated by exploratory and confirmatory factor analysis. The Q3 statistic [28] was used for the test of local independence. Lastly, the item fit was evaluated using S-X^2 [29]. After assessing the quality of the items, item parameters such as item difficulty and discrimination were used for generating simulation data. All items and learners were estimated using the graded response model (GRM) [30] as one of the polytomous item response theory models because all items were multiple choice items in this study.

3.3. A Simulation Study

A simulation study was conducted for evaluating the performance of CAFT using the Monte-Carlo (MC) simulation method. The data of 1000 examinees were simulated using ability parameters randomly drawn from the normal distribution (Mean = 0, Standard Deviations = 1), and the item parameters were estimated by real data. The validity and efficiency, conditional measurement bias, mean squared error, root mean squared error and test overlap rate (T) were computed under different stopping rules [31,32].

$$Bias = \frac{\sum_1^N \left(\hat{\theta}_n - \theta_n\right)}{N} \tag{1}$$

$$MSE = \frac{\sum_1^N \left(\hat{\theta}_n - \theta_n\right)^2}{N} \tag{2}$$

$$RMSE = \sqrt{MSE} \tag{3}$$

$$T = \frac{I}{B} \tag{4}$$

where N is number of examinees, θ is the examinee ability, $\hat{\theta}$ is estimated the examinee ability, I is the number of items in the item bank B, and S^2 is the variance in the exposure rates of all items in the item bank.

3.4. An Application Study

The real application study was conducted using 120 students for implementing CAFT in the statistics discipline. One hundred and twenty students took the adaptive formative assessment during an introductory statistics class, and their learning progression was evaluated to determine whether they possessed the required knowledge. Table 1 shows the descriptive statistics of the subjects.

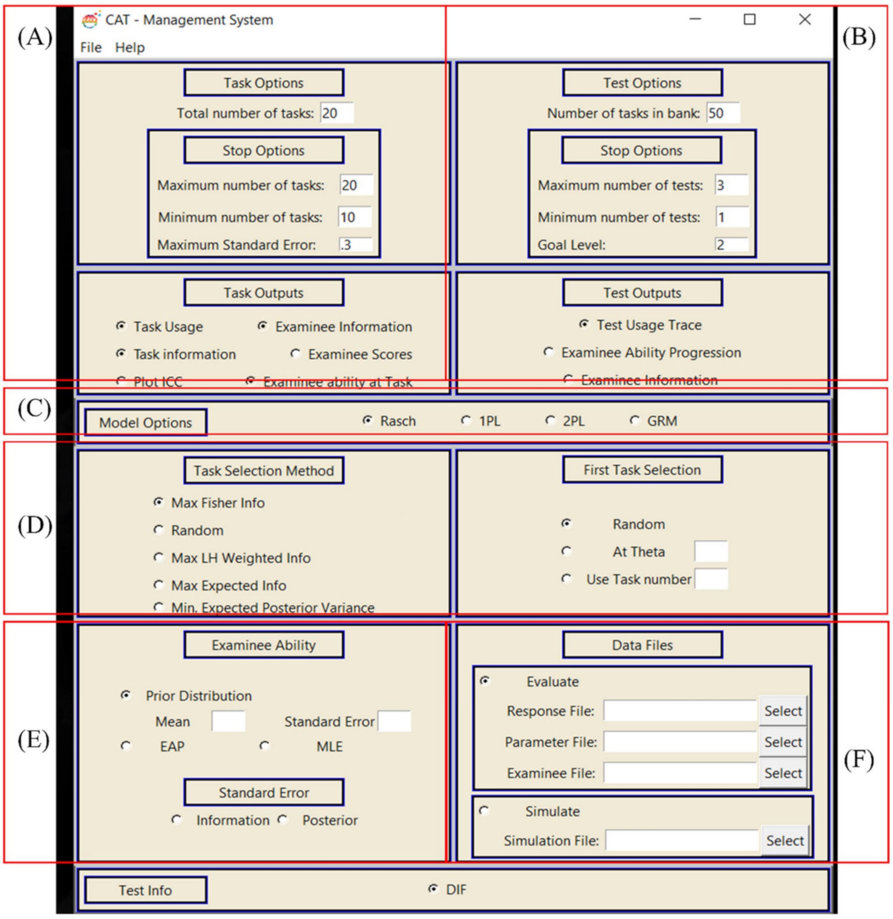
Table 1. Descriptive statistics of the subjects.

		Percent (%)	**Count**
Educational Level	1 year	66.7	80
	2 years	12.5	15
	3 years	12.5%	15
	4 years	8.3	10
Age	20	25.0	30
	30	37.5	45
	40	16.7	20
	50	20.8	25
Gender	Male	40.0	48
	Female	60.0	72
Job Status	Full time	36.7	44
	Part time	23.3	28
	No	40.0	48

3.5. Methodology

3.5.1. Statistical Functions of the CAFT Software

We developed the CAFT engine in such a way that it can administer adaptive formative assessments. Figure 3 shows a screenshot of the graphical user interface in the CAFT engine. CAFT consists of seven sections containing functions including (a) test options, (b) item options, (c) IRT model selection, (d) adaptive options: adaptive selection algorithms, (e) learner estimation method, and (f) output files and simulation.



Task Options
Total number of tasks: 20
Stop Options
Maximum number of tasks: 20
Minimum number of tasks: 10
Maximum Standard Error: .3

Test Options
Number of tasks in bank: 50
Stop Options
Maximum number of tests: 3
Minimum number of tests: 1
Goal Level: 2

Task Outputs
● Task Usage ● Examinee Information
● Task information ○ Examinee Scores
○ Plot ICC ● Examinee ability at Task

Test Outputs
● Test Usage Trace
○ Examinee Ability Progression
○ Examinee Information

Model Options ● Rasch ○ 1PL ○ 2PL ○ GRM

Task Selection Method
● Max Fisher Info
○ Random
○ Max LH Weighted Info
○ Max Expected Info
○ Min. Expected Posterior Variance

First Task Selection
● Random
○ At Theta
○ Use Task number

Examinee Ability
● Prior Distribution
Mean Standard Error
○ EAP ○ MLE
Standard Error
○ Information ○ Posterior

Data Files
● Evaluate
Response File: Select
Parameter File: Select
Examinee File: Select
○ Simulate
Simulation File: Select

Test Info ● DIF

Figure 3. An example of CAFT using the CAT system: Engine Window. (**A**): options related to general item information; (**B**): options related to the general test information; (**C**): options related to the IRT models; (**D**): core part of the adaptive system; (**E**): the examinee's ability estimation area; (**F**): area for selecting data files or conducting a simulation.

The (A) area comprises the options related to general item information. It consists of (1) the number of items in the item bank, (2) the maximum number of items for the administered test, (3) the minimum number of items for the test, and (4) the stopping criterion for the adaptive test. In addition, this part provides the output files related to the item characteristics including item difficulty, discrimination, and information curve. The (B) area consists of the options related to the general test information, which is more connected to implementing adaptive formative assessments. It contains (1) the number of tests that the sequential formative assessment will have, (2) the maximum number of tests in a set of tests, and (3) the desired outputs related to the formative assessment including a learner's learning progression and test usage trace. The (C) area provides the options related to the IRT models, namely, Rasch, 1PL model, 2PL model, and GRM, which would be chosen depending on the item characteristics.

The (D) area is the core part of the adaptive system. It contains different adaptive algorithms that a user can choose from. Moreover, it contains the methods for the first item selection, which determines what method is used to generate the first item from the item bank. The user can choose one of the options for implementing adaptive item selection. The first method is maximum Fisher information (MEI), which is used for finding items that maximize the Fisher's information for a learner with the interim proficiency estimate and the number of items administered [33]. The next method is likelihood weight information criterion proposed by Veerkamp and Berger. In the likelihood weight information method, the information function is summed through the ability and weighted by the likelihood function after item administration has been performed. The last item selection option is the minimum expected posterior variance (MEPV) [34]. This method uses predictive distribution, and it selects items that yield the minimum predicted posterior variance given previous responses. The system picks the item i_k at stage k remaining in the item bank that minimizes the expected posterior variance.

The (E) area is about the examinee's ability estimation. It contains the method options for ability estimation and standard error calculation. This software provides two methods for estimating a learner's ability, including the maximum likelihood estimate and the expected a posteriori estimate. The maximum likelihood estimate is used to estimate a learner's ability by maximizing a likelihood function. The likelihood function of the responses to all items administered is computed as follows:

$$\hat{\theta}^{ML}_{u_{A_k}} \equiv \mathrm{arg}max_{\theta}\{L(\theta|u_A,\ \xi_A) : \theta \in (-\infty, \infty)\}, \tag{5}$$

where a learner's ability = θ, a bank of items ≡ B, a set of items administered after stage $k \equiv A_k$, a set of items remaining at stage $k \equiv R_k = B - A_{k-1}$, a response to item i (m_i possible categories): $u_i \in \{1,\ 2,\ 3,\ \ldots,\ m_i\}$, a single item administered at stage $k \equiv i_k$, a response to item i at stage $k \equiv u_{i_k}$, and ξ refers to the item characteristics, such as item difficulty and item discrimination, depending on the IRT models.

Meanwhile, the expected a posterior estimate (EAP) combines prior information about learners with likelihood information from the data [34]. EAP assumes that proper prior information about learners exists. The EAP can be computed by the following equation:

$$\hat{\theta}^{EAP}_{u_{A_k}} \equiv \mathrm{E}[\theta] = \int \theta g\left(\theta \middle| u_{A_k},\ \xi_{A_k}\right) d\theta \tag{6}$$

Additionally, this software computes the standard error of measurement for selecting and stopping CAT, for which posterior variance is used.

Lastly, the (F) area is for selecting data files or conducting a simulation. The data files, including the response data, item characteristics data, and examinees' abilities data, can be uploaded, and a simulated data file can be generated.

3.5.2. Dynamic Bayesian Network for Diagnostic Learning Progression under the Adaptive Formative Assessment System

Since CAT is an adaptive system, a psychometric method should be developed in order to estimate each student's learning progression over multiple measurement time points. The estimation of a learner's learning progression over multiple time points in CAFT is produced by DBNs [35]. DBNs are a way to extend a static Bayesian network to model probability distributions over multiple time points [36]. DBNs offer a real-time updating method to estimate a learner's previous, current, and future states of a system over a specific period of time [37].

The DBNs for a formative assessment system contain prior information on the hidden state, $\mathrm{P}(\mathrm{X}_1)$, a transition function of the hidden states over multiple time points, $\mathrm{P}(\mathrm{X}_t|\ \mathrm{X}_{1:t-1})$, and an observation function given the hidden state, $\mathrm{P}(\mathrm{Y}_t|\ \mathrm{X}_t)$. The detailed expression about three probability matrices are as follows:

Initial probability matrix of the hidden state at the first time point, $P(X_1)$;
Transition probability matrix, $P(X_t| X_{1:t-1})$;
Conditional probability matrix, $P(Y_t| X_t)$.

Another aspect of DBNs supports the monitoring of learning progression over a specific period of time in a formative assessment. Once an observation has been made on a subset of the variables in the network at a certain point in time, educators are able to make inferences about the remaining unobserved variables in the network at any given time point. In other words, the DBNs reflect the states at previous and future points in time, as well as the current state, because the states at the current point in time will impact the state in the future and are impacted by the states in the past. Therefore, there are three main inferences that can be performed using DBNs:

Smoothing: The process of monitoring states at previous time $t - 1$ given evidence at time t;
Filtering: The process of monitoring states at time t given evidence at time t;
Prediction: The process of monitoring states at future time $t + 1$ given evidence at time t.

4. Results

4.1. Item Bank Evalutioan

First, the quality of the item bank was evaluated to see if all of the items were appropriate to be used in this study. The unidimensionality, model fit, local independence, and item fit were assessed (Table 2). The variance explained by the first factor was 69% in the exploratory factor analysis, which means that there was a dominant factor and unidimentionslity was assumed. In addition, the model fit statistics of the one-factor model were computed using confirmatory factor analysis. CFI was 0.95 and RMSEA was 0.04, indicating that the one-factor model is acceptable. For the local independence evaluation, the Q3 values of all of the items were computed, all of which were below 0.36, indicating that all of the items are locally independent based on the criterion that Q3 values below 0.36 represents local independence [38]. Lastly, the item fit statistics were computed using S-X^2 statistics. The p-value of the S-X^2 statistics of all items was above 0.01; hence, all items were appropriate enough to be used. Therefore, 40 items in the item bank were qualified to be used in CAT. Figure 4 shows the information and standard error of measurement curve of the item bank.

Table 2. CBIAS, CRMSE, and test overlap rate (T) calculation.

Stopping Rule	CBIAS	CRMSE	T
None	−0.004	0.114	1
SE (theta) < 0.2	0.002	0.147	0.69
SE (theta) < 0.3	0.006	0.217	0.39
SE (theta) < 0.4	0.008	0.311	0.28

Note. CBIAS, conditional measurement bias; CRMSE, conditional root mean squared error; T, test overlap rate; SE, standard error.

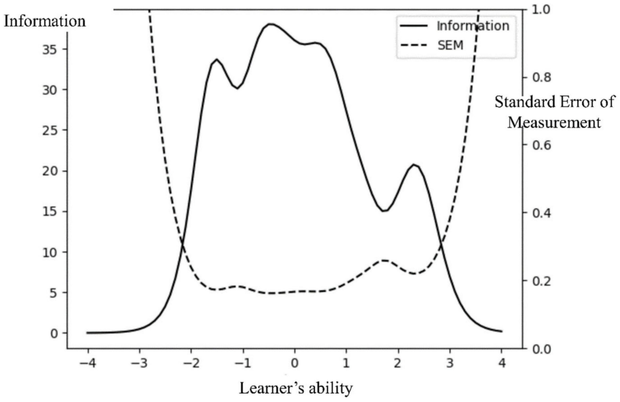


Figure 4. Information and standard error of measurement curve of the item bank.

4.2. Simluation Study

4.2.1. Validity and Efficiency Measures

The validity of the CAFT system was evaluated by computing the measures of the conditional measurement bias (CBIAS), conditional root mean squared error (CRMSE), and test overlap rate (T) across all areas of ability (Table 2). The values of CBIAS, CRMSE, and T across all θ areas are compared in terms of four different stopping rules (i.e., the adaptive testing is stopped if the criterion of the stopping rule is met). The algorithms of the stopping rules are based on the standard error of a learner's ability. Since the value of the stopping rule corresponds to the value of the standard error, a smaller stopping rule means that there is less standard error in the learner's estimated ability, and therefore higher measurement precision. The CBIAS and CRMSE were computed by the discrepancy between the learner's abilities as estimated by CAFT and the simulated learner's true abilities. The values of CBIAS, CMSE, and CRMSE decreased as the stopping rules became smaller. This means that the measurement precision of CAFT was higher as the stopping rules of CAFT were made stricter. In addition, the test overlap rate was computed to count the number of items used in CAFT. The test overlap rate increased as the stopping rules decreased. This is because the stricter the stopping rule became; the more items were used from the item bank. Therefore, the results found that CAFT functions that as the stopping rules decreased, the measurement accuracy and precision increased based on the measures of the CBIAS and CRMSE and the number of items used from the item bank also increased.

4.2.2. Reliability Measures and Number of Items Used

The mean and SD of the number of items used, reliability values, and correlation values were computed in the simulation study (Table 3). The mean and SD of the items used increased when the stopping rule decreased (i.e., measurement precision was made stricter). The average reliability values by Cronbach's alpha were computed using the used items used in the simulation study. The reliability values increased as the stopping rules decreased.

Table 3. Reliability measures.

Stopping Rule	**Number of Items Used**		**Reliability**	**r**
	Mean	**SD**		
None	40	0	0.98	1
SE (theta) < 0.2	27.67	12.01	0.95	0.97 **
SE (theta) < 0.3	15.80	8.93	0.92	0.94 **
SE (theta) < 0.4	11.23	4.92	0.89	0.91 **

Note. ** is *p*-value < 0.01.

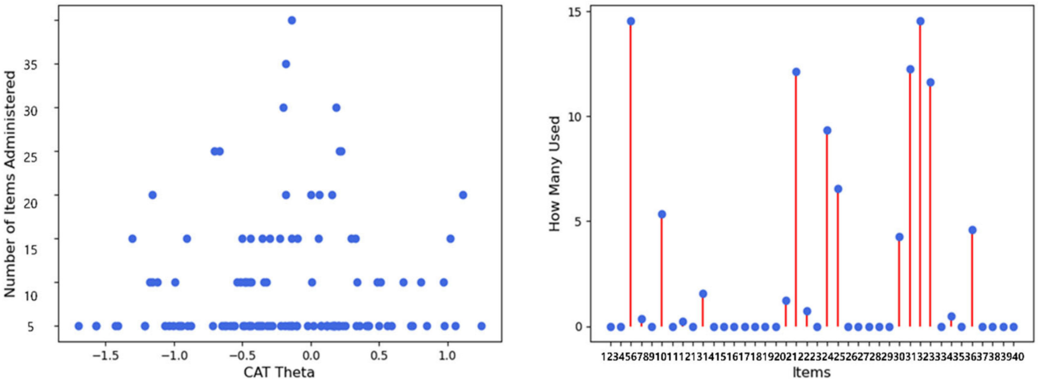
That is because more items were used as the stopping rule decreased and the measurement precision increased. These results show that the reliability was already above 0.9 by using only an average of 15.8 items. This indicates that CAFT is efficient, which means that CAFT can estimate a learner's ability while using less items without loss of measurement precision. Correlations between the estimated learner's ability and the no stopping rule or the remaining stopping rules were computed. The values of Pearson's correlation ranged from 0.91 to 1, indicating that CAFT functioned well under different stopping rules. In more detail, the correlation reached 0.97 between the estimated learner's abilities by CAFT using only 27.67 items from the item bank and the estimated learner's abilities when using all items in the item bank. Therefore, CAFT reached similar reliability while using 31% less items than a fixed test in this case.

4.3. Real Data Study

The purpose of the application study was to evaluate the efficiency, validity, and reliability of CAFT. CAFT was applied to real assessment system for the introduction to statistics course of an online university. Overall, 1000 learners were simulated for the simulation study and 120 learners were used in the real application study.

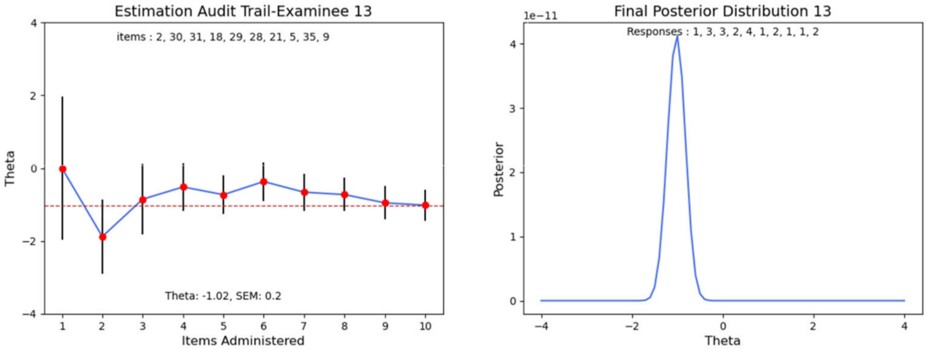
4.3.1. CAT Process in Application Study

The application study was conducted with 120 learners. All learners took the same statistical assessment as was used in the simulation study. Table 1 displays the descriptive statistics of the subjects in the application study. Half of them were given all of the items from the item bank, meaning there was no stopping rule, while the rest of them took adaptively chosen items where the stopping rule (SE) was less than 0.3. After this, the abilities estimated by CAFT, with a stopping rule of 0.3, were compared with the abilities estimated by CAFT with a no stopping rule. The average number of items used with the 0.3 stopping rule in CAFT was 21.3 items, while all items in the item bank (40 items) were answered with a no stopping rule. The Pearson's correlation between the learners' estimated abilities with a stopping rule of 0.3 and a rule of none was 0.97. This means that the CAFT assessment provides efficient and accurate ability estimates while using approximately half as many items in this case. Figure 5 shows the adaptively used items in terms of a learner's ability change. The X-axis indicates the estimated abilities of the different learners, and the Y-axis indicates the number of items used. Since the items were adaptively selected, as Figure 5 shows, the number of items used was different in terms of a learner's ability. Moreover, the number of times a particular item was used was not the same for all items. Figures 6 and 7 show two examples of the estimation of an examinee's ability. Figure 6A indicates the estimated ability change by administering different items, while Figure 6B shows the final estimated ability of the examinee. Figure 7 contains the same information as Figure 6; however, the examinee in Figure 7 was given more items than the examinee in Figure 6. In Figure 7, the examinee had 25 items in order to estimate their final ability, while in Figure 6, the examinee had 10 items. This means that each examinee received adaptively chosen items from an item bank in order to estimate their ability accurately and efficiently. Therefore, we can confirm CAFT functioned well with this information.



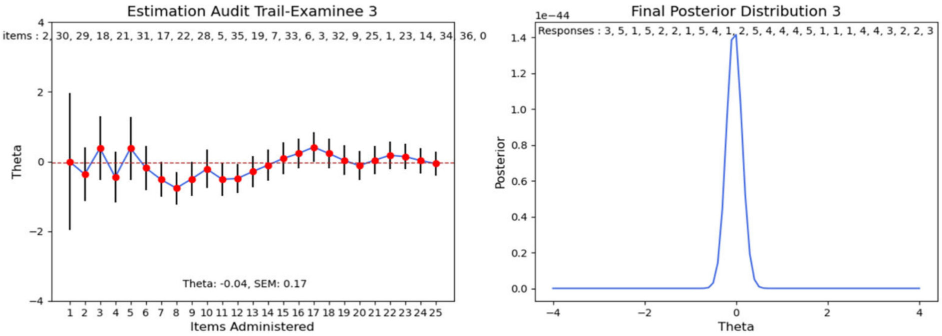
(**A**) (**B**)

Figure 5. Items were adaptively chosen based on a leaner's ability. (**A**) The number of items used in terms of a learners' abilities. (**B**) The number of items selected from an item bank.



(**A**) (**B**)

Figure 6. An example of an examinee's estimated ability change using CAFT. (**A**) The examinee took 10 items and his estimated ability changed in terms of the adaptively selected items. (**B**) The final estimated ability of the examinee.



(**A**) (**B**)

Figure 7. An example of an examinee's estimated ability change using CAFT. (**A**) The examinee took 25 items and his estimated ability changed in terms of the adaptively selected items. (**B**) The final estimated ability of the examinee.

4.3.2. Validity Measures Using Real Data

We evaluated the concurrent validity and predictive validity of CAFT. The correlation values were computed between the abilities estimated by CAFT for a midterm exam and a final exam that covered the same domain (Table 4). The final exam included more advanced statistical knowledge in addition to the content of the midterm exam. The correlation between the estimated abilities from CAFT and the sum scores of the midterm exam was 0.83 under a no stopping rule and 0.81 under a 0.3 stopping rule. Such high correlation values indicate that CAFT has high concurrent validity. In addition, the correlation between the estimated abilities from CAFT and the sum scores of the final exam was 0.69 under the no stopping rule and 0.67 under the 0.3 stopping rule. Therefore, the predictive validity of CAFT was confirmed.

Table 4. Correlation between the estimated ability by CAFT and the sum scores of the exams.

Stopping Rule	Midterm Exam	Final Exam
None	0.83	0.69
SE (theta) < 0.3	0.81	0.67

4.4. Estimation of Individual Learning Progression Using DBNs

An instructor can build a formative assessment system with several tests, and learners take the tests in terms of their previous ability levels. Then, the DBNs estimate the probability of a learner's level change over multiple testing time points. Figure 8 shows a test process/testing path for each learner through the adaptive formative assessment system. It is shown that each learner has a different number of tests adaptively administered to them across different time points.

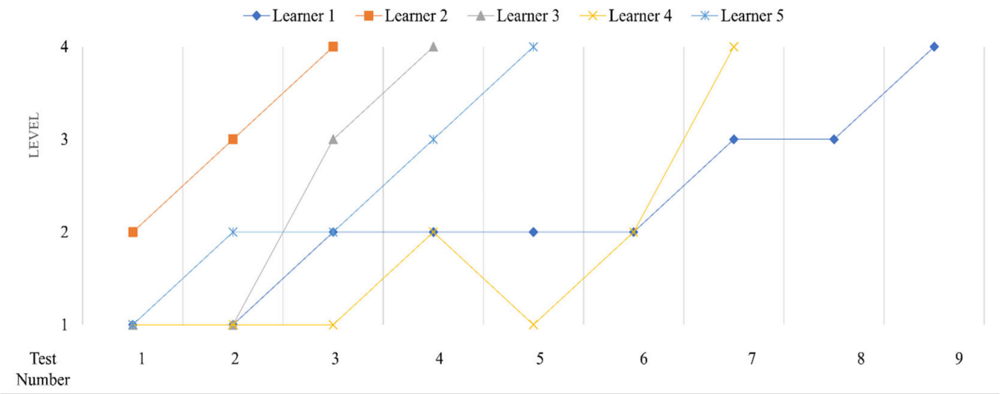
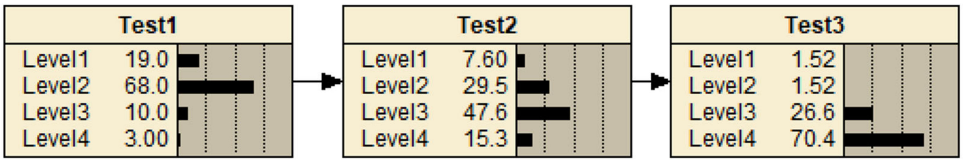


Figure 8. An example of a testing path from five learners in the adaptive formative assessment.

Figures 9 and 10 display examples of the representation of DBNs for two learners' level status changes in multiple tests over different measurement time points. DBNs estimate the probability of a learner's learning status changing across tests. Figure 9 shows the learning progression of the second learner. The second learner took three tests, and the levels changed from level 2 to level 3 and from level 3 to level 4. The probability that the learner was at level 2 in the first test was 0.68, while the probability that the learner was at level 3 in the second test was 0.47. Lastly, the probability that the learner was at level 4 in the third test was 0.70. After the third test, the test was stopped because the learner had reached the targeted level 4 within three tests. Figure 10 shows the testing process of the third learner, which is the same as that of the second learner, while the second learner took four tests because he did not reach level 4 until test 4.



Test1	
Level1	19.0
Level2	68.0
Level3	10.0
Level4	3.00

Test2	
Level1	7.60
Level2	29.5
Level3	47.6
Level4	15.3

Test3	
Level1	1.52
Level2	1.52
Level3	26.6
Level4	70.4

Figure 9. Level change of the second learner in CAFT.

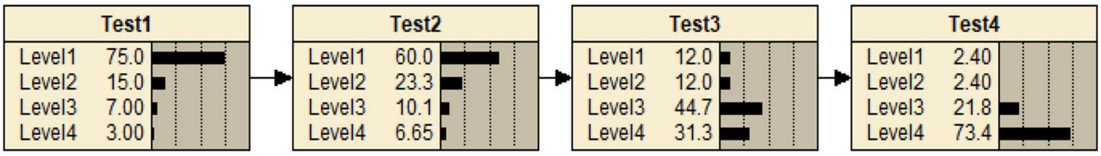


Figure 10. Level change of the third learner in CAFT.

5. Discussion

Current educational situations necessitate developing and implementing alternative strategies to replace the previous functions of learning evaluation settings due to COVID-19. The integration of artificially intelligent components into educational systems greatly helps to efficiently provide meaningful information for evaluation and learning [39–41]. On the one hand, adaptive testing technology offers an efficient and reliable personalized evaluation system. On the other hand, formative assessment systems provide information about the knowledge, skills, and abilities (KSAs) that a student may have obtained, their learning progression over time, and diagnostic feedback relative to their instruction and curriculum. The combination of adaptive testing technology and a formative assessment system is a promising means for effectively generating data that provide a teacher insight into where their students are struggling collectively and where particular students might need more help in accordance with the curriculum.

While substantive theory and task design aspects in online formative assessments have been investigated, few analytic systems that allow for adaptively implementing formative assessments have been proposed [42,43]. The linking of an adaptive algorithm as a psychometric method with an online formative assessment system allows the assessment to be efficient, accurate, and personalized [44,45].

This study developed CAT for an online formative assessment system, called CAFT (computerized adaptive formative testing). CAFT is an evaluation technology that integrates seamless artificial intelligence elements into education for remotely delivering a customized diagnostic learning service. CAFT takes advantage of the adaptive functions of CAT and the statistical estimation method from DBNs. The adaptive functionality of CAT allows for adaptively selecting an item for a test and selecting a test for a sequential test aligned with a curriculum. In addition, DBNs estimates the real-time change of a learner’s ability across multiple tests. Therefore, a combination of CAT as an adaptive selection method and DBNs as a real-time estimation method is the core ingenuity of CAFT as an integration of artificial intelligence into an e-learning system. This paper addressed the detailed specification and functions of CAFT. The performance of CAFT was examined using a simulation and application studies. The results showed that CAFT is a reliable and efficient testing system.

However, the current system has a few limitations. First, the system does not contain detailed functions related to item development. A valid adaptive formative assessment using CAT should coherently consider substantive concepts, item design, and statistical analysis. The connection among substantive theories, item design, and analytic methods provides information about how students are progressing and where they are having difficulties solving items. This information is useful in the adaptive selection of items, assignments, or alternative instructional approaches in terms of a learner’s level. If the item can be generated by various item features linking the substantive theories in a discipline, assessment can be more targeted and adaptive. Since the change of just one of the item features can require students to use different KSAs, more diagnostic information can be obtained using exquisite item design. Second, while there is a possibility of using CAFT for distinguishing experts from novices in perception, procedures, acquisition, and learning progression, more research into a valid psychometric model must be conducted. Third, a connection between adaptive learning and adaptive evaluation provides a more comprehensive and meaningful learning management system. A systematic connection between an adaptive learning system from CAFT and adaptive learning will be the topic of future study.

Despite these limitations, the CAFT system has benefits not only for creating an effective and efficient test setting, but also for providing information that helps teachers better understand students' learning progression. Eventually, the results of the assessment can be effectively used for the selection of instructional strategies such as re-teaching, utilizing alternative instructional approaches, altering the difficulty level of items or assignments, or offering more opportunities for practice.

In this study, we addressed the system development, theoretical models, and system implementation that can support adaptive evaluation and automatic learning evaluation. Traditional testing relies on fixed and static assessments. CAFT is a progression of testing that utilizes modern measurement techniques to integrate artificial intelligence into the educational process. Instead of giving all students the same test, CAFT dynamically generates tests from an item bank in a manner that is unique to the students taking the test. In addition, CAFT, by integrating adaptivity into the formative assessment process, can create a testing system in which the entire testing process is tailored to the individual as opposed to only certain tests. The result is a testing experience that is more efficient at gauging students' progress over time and more informative for both students and instructors. As such, CAFT is a novel system well equipped to offer an adaptive personalized learning service.

Author Contributions: Conceptualization, supervision, methodology, writing: Y.C.; program development, editing: C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education of the Korea and the National Research Foundation of Korea (NRF-2019S1A5A2A03052192).

Acknowledgments: The authors wish to express thanks to Binna Lee for data collection.

Conflicts of Interest: No potential conflict of interest relevant to this article was reported.

References

1. Tiago, M.F.; Paula, F.L. Towards Next Generation Teaching, Learning, and Context-Aware Applications for Higher Education: A Review on Blockchain, IoT, Fog and Edge Computing Enabled Smart Campuses and Universities. *Appl. Sci.* **2019**, *9*, 4479.
2. Alberto, R.F.; Rafael, M.C.; Faraón, L.L. Computational Characterization of Activities and Learners in a Learning System. *Appl. Sci.* **2020**, *10*, 2208.
3. Conejo, R.; Guzmán, E.; Millán, E.; Trella, M.; Pérez-De-La-Cruz, J.L.; Ríos, A. SIETTE: A web-based tool for adaptive testing. *Int. Artifi. Intelli. Educ.* **2004**, *14*, 29–61.
4. McCallum, S.; Milner, M.M. The effectiveness of formative assessment: Student views and staff reflections. *Assess. Eval. High. Educ.* **2020**, 1–16.
5. Bennett, R.E. Formative assessment: A critical review. *Assess. Educ.* **2011**, *18*, 5–25. [CrossRef]
6. Black, P.; Wiliam, D. Assessment and classroom learning. *Assess. Educ.* **1998**, *5*, 7–74. [CrossRef]
7. Briggs, D.C.; Ruiz-Primo, M.A.; Furtak, E.; Shepard, L.; Yin, Y. Meta–analytic methodology and inferences about the efficacy of formative assessment. *Educ. Meas.* **2012**, *31*, 13–17. [CrossRef]
8. Choi, Y.; Rupp, A.; Gushta, M.; Sweet, S. *Modeling Learning Trajectories with Epistemic Network Analysis: An Investigation of a Novel Analytic Method for Learning Progressions in Epistemic Games*; National Council on Measurement in Education: Philadelphia, PA, USA, 2020; pp. 1–39.
9. Walker, D.J.; Topping, K.; Rodrigues, S. Student reflections on formative e–assessment: Expectations and perceptions. *Learn. Media Technol.* **2008**, *33*, 221–234. [CrossRef]
10. Brown, L.I.; Bristol, L.; De Four-Babb, J.; Conrad, D.A. National Tests and Diagnostic Feedback: What Say Teachers in Trinidad and Tobago? *J. Educ. Res.* **2014**, *107*, 241–251. [CrossRef]
11. Havnes, A.; Smith, K.; Dysthe, O.; Ludvigsen, K. Formative assessment and feedback: Making learning visible. *Stud. Educ. Eval.* **2012**, *38*, 21–27. [CrossRef]
12. Schez-Sobrino, S.; Gmez-Portes, C.; Vallejo, D.; Glez-Morcillo, C.; Miguel, A.R. An Intelligent Tutoring System to Facilitate the Learning of Programming through the Usage of Dynamic Graphic Visualizations. *Appl. Sci.* **2020**, *10*, 1518. [CrossRef]

13. Khan, R.A.; Jawaid, M. Technology Enhanced Assessment (TEA) in COVID 19 Pandemic. *Pak. J. Med. Sci.* **2020**, *36*, S108. [CrossRef] [PubMed]
14. West, P.; Rutstein, D.W.; Mislevy, R.J.; Liu, J.; Choi, Y.; Levy, R.; Behrens, J.T. A Bayesian Network Approach to Modeling Learning Progressions and Task Performance. CRESST Report No 776. National Center for Research on Evaluation, Standards, and Student Testing. Available online: https://files.eric.ed.gov/fulltext/ED512650.pdf (accessed on 19 August 2010).
15. Lee, H.; Choi, Y. The Influence of Human Resource Management Strategy on Learning Achievement in Online Learning Environment: The Moderated Mediating Effect of Metacognition by Extraneous Cognitive Load. *J. Korean Assoc. Educ. Inf. Media* **2019**, *25*, 853–872.
16. Nagandla, K.; Sulaiha, S.; Nallia, S. Online formative assessments: Exploring their educational value. *JAMP* **2018**, *6*, 51. [PubMed]
17. Han, K.T.; Simul, C.A.T. Windows software for simulating computerized adaptive test administration. *Appl. Psychol. Meas.* **2012**, *36*, 64–66. [CrossRef]
18. Kingsbury, G.G.; Zara, A.R. Procedures for selecting items for computerized adaptive tests. *Appl. Meas. Educ.* **1989**, *2*, 359–375. [CrossRef]
19. Han, K.T. An efficiency balanced information criterion for item selection in computerized adaptive testing. *J. Educ. Meas.* **2012**, *49*, 225–246. [CrossRef]
20. Embertson, S.E.; Reise, S.P. *Item Response Theory for Psychologists*, 1st ed.; Psychology Press: Hove, East Sussex, UK, 2000.
21. Veerkamp, W.J.; Berger, M.P. Some new item selection criteria for adaptive testing. *J. Educ. Behav. Stat.* **1997**, *22*, 203–226. [CrossRef]
22. Han, K.T. Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Appl. Psychol. Meas.* **2016**, *40*, 289–301. [CrossRef]
23. Chang, H.H.; Ying, Z.A. Global information approach to computerized adaptive testing. *Appl. Psychol. Meas.* **1996**, *20*, 213–229. [CrossRef]
24. Chang, H.H.; Ying, Z. A-stratified multistage computerized adaptive testing. *Appl. Psychol. Meas.* **1999**, *23*, 211–222. [CrossRef]
25. Choi, Y.; Mislevy, R. Dynamic Bayesian Inference Network and hidden Markov Model for Modeling Learning Progression over Multiple Time Points. Ph.D. Thesis, University of Maryland at College Park, College Park, MD, USA, 2012.
26. Revy, J. Two-Phase Updating of Student Models Based on Dynamic Belief Networks. In *Intelligent Tutoring Systems: 4th International Conference ITS'98*; San Antonio, TX, USA, 16–19 August 1998, Springer: Berlin/Heidelberg, Germany, 1998; pp. 274–283.
27. Reye, J.A. *Belief Net Backbone for Student Modeling Intelligent Tutoring System: Intelligent Tutoring Systems*; Frasson, C., Gauthier, G., Lesgold, A., Eds.; Springer: Berlin/Heidelberg, Germany, 1996; pp. 596–604.
28. Yen, W.M. Scaling performance assessments: Strategies for managing local item dependence. *J. Educ. Meas.* **1993**, *30*, 187–213. [CrossRef]
29. Orlando, M.; Thissen, D. Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* **2003**, *27*, 289–298. [CrossRef]
30. Samejima, F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *ETS Res. Rep. Ser.* **1968**, *1*, i-169.
31. Barrada, J.R.; Olea, J.; Ponsoda, V.; Abad, F.J. Incorporating randomness in the fisher information for improving item-exposure control in CATs. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 493–513. [CrossRef]
32. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
33. Owen, R.J. A Bayesian approach to tailored testing. *ETS* **1969**, *1969*, 69–92. [CrossRef]
34. Stocking, M.L.; Lewis, C. Controlling item exposure conditional on ability in computerized adaptive testing. *ETS* **1995**, *23*, 57–75. [CrossRef]
35. Almond, R.G.; DiBello, L.V.; Moulder, B.; Zapata-Rivera, J.D. Modeling diagnostic assessments with Bayesian networks. *J. Educ. Meas.* **2007**, *44*, 341–359. [CrossRef]
36. Almond, R.G.; Mislevy, R.J.; Steinberg, L.S.; Williamson, D.M.; Yan, D. *Bayesian Networks in Educational Assessment*; Springer: New York, NY, USA, 2015.

37. Choi, Y.; Cho, Y.I. Learning Analytics Using Social Network Analysis and Bayesian Network Analysis in Sustainable Computer-Based Formative Assessment System. *Sustainability* **2020**, *12*, 7950. [CrossRef]
38. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Academic Press: Abingdon, VA, USA, 2013.
39. Nguyen-Thin, L. A Classification of Adaptive Feedback in Educational Systems for Programming. *Systems* **2016**, *4*, 22. [CrossRef]
40. William, V.C.; Milton, R.C.; Xavier, P.P. Improvement of an Online Education Model with the Integration of Machine Learning and Data Analysis in an LMS. *Appl. Sci.* **2020**, *10*, 5371. [CrossRef]
41. Shaojie, Q.; Kan, L.; Bo, W.; Yongchao, W. Predicting Student Achievement Based on Temporal Learning Behavior in MOOCs. *Appl. Sci.* **2019**, *9*, 5539. [CrossRef]
42. Swanson, L.; Stocking, M.L. A model and heuristic for solving very large item selection problems. *Appl. Psychol. Meas.* **1993**, *17*, 151–166. [CrossRef]
43. Van Der Linden, W.J. A comparison of item-selection methods for adaptive tests with content constraints. *J. Educ. Meas.* **2005**, *42*, 283–302. [CrossRef]
44. Nichols, P.D.; Chipman, S.F.; Brennan, R.L. *Cognitively Diagnostic Assessment*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1995.
45. Corcoran, T.; Mosher, F.; Rogat, A. *Learning Progressions in Science: An Evidence-Based Approach to Reform*; Consortium for Policy Research in Education: Philadelphia, PA, USA, 2009.