IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 July 2023, accepted 22 August 2023, date of publication 4 September 2023, date of current version 8 September 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3311615*

APPLIED RESEARCH

# The Development of an Automatic Test Assembly System for a Formative Assessment in Mastery Learning Instruction: Case of the SQL Mastery Course

**HUNG-YI CHEN[1], YING-CHIEH LIU[2], XUAN-QI LIU[1], AND VICTORIA CHIU[3]**

[1]Department of Information Management, Chaoyang University of Technology, Taichung 411030, Taiwan
[2]Department of Information Management, National Chin-Yi University of Technology, Taichung 413310, Taiwan
[3]Department of Accounting, Finance and Law, State University of New York at Oswego, Oswego, NY 13126, USA

Corresponding author: Ying-Chieh Liu (allanliu@ncut.edu.tw)

This work was supported in part by the Teaching Practice Research Program, Minister of Education, Taiwan, under Grant PBM1110104.

**ABSTRACT** The teaching method of mastery learning requires the lecturers to conduct their regular formative assessments so as to monitor each student's learning progress. However, regularly preparing quality test forms for these assessments has caused a tremendous workload on the lecturers and has prevented them from effectively executing the master learning instruction. To resolve the problem, this study has built an automatic test assembly system with the automatic and manual test assembly functions for preparing fixed-length testing forms. Our system is implemented in a Structural Query Language (SQL) mastery course that helps students prepare for the ORACLE SQL Certification (OCE SQL) examination. OCE SQL is a program offered by the Oracle Corporation to validate and recognize the skills of professionals in the SQL programming. By applying the Hambleton and De Gruijter's method, our system can automatically assemble items to a test form with a quality indication of the maximum probability of misclassifying mastery. Furthermore, our system can evaluate the prementioned probability for a manually assembled test form. The present study has also investigated the numeric methods for converting the domain cutoff score to the latent ability scale for items with heterogeneous parameters when applying the method. The empirical computing results showed that the secant method is superior to the bisection method for the conversion. This study contributes an automatic test assembly system to the engineering education practice so as to lighten the lecturers' burden of preparing quality assessments and to further facilitate the execution of the regular formative assessments in the mastery learning instruction.

**INDEX TERMS** Criterion-referenced test, engineering education, formative assessment, mastery learning instruction, test assembly systems, SQL programmer.

## I. INTRODUCTION

The mastery learning teaching method can effectively improve the learning outcomes [2] and increase the pass rate on licensure or certification examinations [3], [4]. When adopting the mastery learning instruction, lecturers must administer formative assessments so as to confirm the mastery of the learning topics and provide feedback and corrective actions to the students. After the corrective actions, the students who do not pass the assessment proceed with a second equivalent test. At the end of the learning stage, the lecturers perform the summative evaluation to ensure the students' mastery. The formative assessment is one of the crucial elements in the mastery learning instruction [5].

The formative assessment is a criteria-referenced valuation that designs the tests according to the requirements of the ability to measure the mastery level. The evaluation measures the behavioral side of the subject, not to compare

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Wang.

their achievements, but to determine their professional competence.

The tedious process of preparing the test forms and inaccurate assessments will affect the effective execution of the mastery learning instruction. To regularly implement the formative assessment of the mastery, the lecturers are required to set the ability criteria, design items, build item banks, assemble test forms, prepare detailed solutions, and administer tests. These entail a much larger workload for the lecturers, which can hinder them from devoting more precious time to teaching. Additionally, while assembling the test forms, the lecturers must consider the issue of misclassifying the mastery. There are two types of misclassifications: one is false mastery, in which a non-mastered student is classified as mastered, and the other is false non-mastery, in which a mastery student is classified as non-mastered. Excessive misjudgments prevent the teachers from identifying unskilled students and giving corrective activities, or investing unnecessary resources in the already proficient students.

Many mastery assessment methods or test assembly methods for formative assessments have been proposed in the literature for the mastery learning instruction. However, most focus on adaptive models, which generate different test forms for a subject. Although adaptive mastery assessments can measure the subject's latent ability more accurately than the fixed-length ones, they are not adept at being used in the mastery learning instruction. Lecturers suffer significant workloads because they must prepare a corrective action for each student with a different test form [6]. Conversely, fixed-length mastery assessment is easier to implement in teaching practice and can be flexibly used with paper-and-pencil testing or converted to online testing. Hence, a new system is commended for lecturers to prepare fixed-length test forms with measurement precision for practical mastery learning instruction.

This study has developed an automatic test form assembly system so as to assist the lecturers in performing the mastery learning instruction in the Oracle SQL certification course. The proposed system contains modules of the item bank and the test assembly. The item bank module can import, classify, and maintain items and their item response parameters (difficulty, discrimination, and guessing). Moreover, the test assembly module automatically generates the test forms. The module adopted the method proposed by Hambleton and De Gruijter [1] wherein items are selected from the item bank to assemble a test form. This method can generate the test forms with fixed lengths and high discrimination around the mastery cutoff score while considering the probability of misclassification. With the help of the proposed system, the lecturers would be able to quickly generate test forms with a high measurement precision for formative assessments in the mastery learning instruction. Moreover, because the test assembly procedure needs to use a numeric method to convert the cutoff score on the domain scale to the one on the latent ability scale in the Item Response Theory (IRT), the study compared the secant and bisection root-finding algorithms for the conversion with the criteria of the computation time and the solution quality.

We organize the remainder of the paper as follows. First, the study reviews the literature on the mastery learning, computer-based testing systems and automatic test assembly, and the mastery testing in the formative assessments. Second, the formalized system process, system functions, and the procedures to assemble and evaluate the test forms are discussed. Third, the study evaluates two root-finding algorithms for solving the conversion problem of the domain cutoff point. Fourth, the implementation of the system in a SQL mastery course is presented so as to demonstrate its usefulness. The conclusion is presented in the last section.

## II. LITERATURE REVIEW

### A. MASTERY LEARNING

Mastery learning was proposed by Bloom [7] to improve the students' learning outcomes through tutoring and individualized instructions. The teaching method uses the formative assessments so as to identify the difficulties in learning and then provides personal instructions or corrective actions as feedback for the students. After they complete the disciplinary actions, the lecturers perform the second parallel formative assessment to assess those who did not pass the first assessment. The theoretical basis of the mastery learning has three aspects [8]:

1) Corrective and enriching activities create an environment that supports the student's learning motivation.
2) Formative assessments require the students to retrieve the learned knowledge and apply it. The retrieval practice can improve the learning outcomes and enhance their metacognition.
3) Formative assessments continue to provide learning feedback for the students, which is one of the essential learning strategies.

This method can improve the teaching outcomes of the competency-based courses. Many studies have pointed out that mastery learning can increase the outcomes and enhance the learner's confidence in the subjects [2], [9], [10]. For example, applying the mastery learning in nursing education programs improves the pass rate of certificate examinations [3], [4]. Furthermore, since mastery learning has been applied to computer science education [11], it can also be applied to teaching the Structural Query Language (SQL) programming language to enhance the learning outcomes and the pass rate of the certification exam.

### B. COMPUTER-BASED TESTING AND AUTOMATIC TEST ASSEMBLY

Computer-based testing (CBT) has been widely used in high-stake testing, such as language proficiency tests, professional/licensure exams, achievement tests, etc., and in formative evaluation for proving learning feedback. Replacing paper-and-pencil tests, the CBT delivers tests online to make the large-scale tests quicker and more efficient. The

advantages include reducing the administrative costs and increasing the measurement efficiency and precision.

CBT, from the viewpoint of the system architecture, contains four components: item bank, automatic test assembly, test bank, and test delivery. First, the test bank contains test items that are authored according to the testing criteria and/or blueprints. The automatic test assembly component selects items from the item banks to assemble the test forms. Next, these test forms are stored in the test bank. Finally, the test delivery component selects the tests from the test banks and delivers them online to the examinees.

Building item banks involves categorizing items, estimating parameters, and calibrating new and old items. Items are categorized according to the assessment criteria. Pretests are required to estimate the item parameters [12] and ensure the parameters' invariance [13]. When new items are added to an item bank, they should be linked with the existing items in the bank so that the new and existing items have a standard measurement scale [13], [14].

The automatic test assembly component selects items from the bank to generate the test forms. The selection criteria depends on the testing purposes [15]. For the criterion-referenced testing, the selected items should have high information near the target ability level converted from the test passing score. In contrast, the test should have items with appropriate difficulties and high information for various ability levels for the norm-referenced testing. According to the IRT, high item information means a low measurement error [16]. In addition to the measurement precision, other factors must be considered to select items for the assembly of the test forms. For example, the need to include items from various topics in the assessment criteria so they will have the content validity. Alternatively, the number of items in a test form must be limited so that the examinees can complete the test within the time limit. These quality requirements on test forms become the constraints on maximizing the measure precision in the item selection process. Therefore, many automatic test assembly methods employ heuristics or optimization algorithms so as to generate the test forms [17]. As for the mastery learning, the automated test assembly must produce the parallel test forms with the same precision, but with different content to see how well the students learn after the corrective actions are introduced [18].

The time to execute the automatic test assembly component will depend on how the tests are administered. The component generates the test forms for the not-adaptive testing before they are administered. Conversely, the component is executed while the tests are administered for the adaptive testing [19], [20], [21]. The test delivery system should accordingly estimate the examinees' abilities and select the items with the appropriate difficulties. As for the item exposure, examinees would see the same items in the same non-adaptive test, but see different items in the same adaptive test.

### C. VARIABLE-LENGTH VS. FIXED-LENGTH TESTS IN FORMATIVE ASSESSMENTS

There are two methods when deciding the length of the test: variable and fix. Several pieces of literature have applied variable-length or adaptive mastering testing in the formative assessment. The variable-length mastery testing must be administered online. The advantage is that they can decide as to whether the subject is mastered with a shorter test length [22]. [23] proposed an adaptive formative assessment system for an e-learning environment. Their system is a type of Computer Adapted Testing (CAT) that select items according to the subject's ability estimated during testing. The test ends when the standard error reaches a certain level or no test item can contribute enough information [24], [25]. The system proposed by [26], based on the Sequential Probability Ratio Test (SPRT) method, also applies to the online formative assessment. The SPRT decides the mastery state each time the subject responds to a test item [27]. The likelihood ratio of mastery and non-mastery probabilities is computed for each respondence. When the likelihood ratio is less than the lower threshold; a, or greater than the upper one b, the procedure stops the test, in which $a = log(\beta/(1-\alpha))$ and $b = log(1-\beta/\alpha)$. Their paper did not mention the item selection method used in the testing. The system proposed by [28] is for the formative assessment in personalized online learning. Their system belongs to the CAT as well. Unlike others, they use the subject's memory cycle as the criterion to select items for assessments, which can improve the learning performance according to their experiments.

However, using variable-length formative assessments should be cautious for the high cognitive demand questions, which require memorizing, comprehending, inferring, and summarizing abilities. The variable-length assessments need to be administered by the online systems, as they bring extra cognitive loads and affect the students' testing scores [29]. Therefore, compared to the CBT, using paper-and-pencil tests is more helpful for students to organize their knowledge [30]. Nevertheless, a small amount of literature has addressed the high cognitive demand for solving complex questions in developing the formative assessment systems.

This study advocates the fixed-length method, instead of the variable-length method, for the formative assessment in the mastery learning teaching due to the characteristics of the OCE certificate test. An example of the OCE certificate multiple choice question is as shown in Fig. 1. To solve this question, students need to understand the meaning of GROUP BY, the interactive effects between GROUP BY, WHERE and SELECT, and the limitation of using GROUP BY in clauses. This demands the students to use their abilities of memorizing, comprehending, inferring, and summarizing all knowledge of the Oracle SQL and its applications in the Oracle database. To achieve these abilities and pass the certificate test, except the lecture, simulated tests blended with paper-based tests (PBT) and CBT should be held frequently. The PBT test practice is also crucial in the early

Q. Which two statements are true regarding the GROUP By clause in a SQL statement? (Choose two.)

Options:
a) You can use column alias in the GROUP BY clause.
b) Using the WHERE clause after the GROUP BY clause excludes the rows after creating groups.
c) The GROUP BY clause is mandatory if you are using an aggregate function in the SELECT clause.
d) Using the WHERE clause before the GROUP BY clause excludes the rows before creating groups.
e) If the SELECT clause has an aggregate function, then those individual columns without an aggregate function in the SELECT clause should be included in the GROUP BY clause.

**FIGURE 1.** An example of OCE certificate multiple choice question.

stage of study or when the lecture advances to new knowledge because the PBT enables the students to conduct critical and deep thinking by marking the keywords, writing down the comments, and creating and revising the SQL clauses on paper. Through these paper-bases assessments, the students are easier to organize their knowledge and finish these high cognitive demand questions [30], as this improves the students' learning performance [29].

The high cognitive demands in solving the OCE certificate questions drives our study in adapting the fixed-length method proposed by [1] for the formative assessment in mastery leaning instruction. The method starts with setting a mastery cutoff score with an in-difference zone in the domain scale. Next, the cutoff score in the domain scale is converted to the latent ability scale. Then, based on the IRT theory, their method selects items with high item information near the cutoff point in the latent ability scale to assemble a test form. Finally, their method determines the test length and the required number of items to pass the test so that the given limit on the probability of misclassifying the mastery is achieved. Because the test length is fixed, the generated test form can be administered through paper-and-pencil or online tests. Table 1 compares the proposed system with the reviewed ones.

Although the fixed-length mastery testing requires more items on average than the variable-length, it has several advantages when applied to the formative assessments in the master learning instruction. First, the test delivery component for the fixed-length tests is easy to implement because it does not require complex algorithms such as adaptive testing. Second, test administrators can review the test forms before being administered to the examinees to ensure their quality [19]. Third, fixed-length test forms can be easily administered online or through paper-and-pencil testing. Fourth, compared to the computer-based testing, paper-and-pencil testing can improve the student's knowledge and skills acquisition more effectively [29], [30]. Finally, since all the students have the same items for a test, lecturers have less of a workload [31] in preparing corrective actions for the students. Based on that, fixed-length testing is more applicable than variable-length testing to formative assessments in the mastery learning instruction in our study.

## III. SYSTEM DEVELOPMENT

### A. PROCESS FORMALIZATION FOR APPLYING ON OUR SYSTEM

An automatic test assembly system was built to help the lecturers administer the formative assessments in the mastery learning instruction. The process of applying our system to assist the lecturers in executing the mastery learning instruction needs to be formalized to identify the system's functional requirements. As shown in Fig. 2, three cycles were identified for the process: Item Content Development, Item Pretesting, and Test Form Assembly.

The Item Content Development cycle starts with the authoring items and their solutions by the role of the Item Author. Then, the role imports items into the system and classifies them according to the testing criteria. The last activity in the cycle involves the Item Author maintaining the items.

In the Item Pretesting cycle, the role of the Item Analyzer starts with pretesting and estimating the item statistics. These statistics are then imported into the system for item selection. The Test Form Assembling cycle is the last in the process, which is performed by the Test Form Assembler; its main responsibility is to assemble the test forms manually or automatically to the test bank. When using the automatic function, the Test Form Assembler must specify the topic range of testing, the cutoff score, and the acceptable maximum probability of misclassifying the mastery. With the execution of the process, the lecturers can select a test form from the test bank to perform a formative assessment.

### B. SYSTEM REQUIREMENTS

Based on the above process, the present study identified seven primary functions (F1∼F7) for the system, which were labeled with the activities in various cycles, as shown in Fig. 2:

(F1) Importing items: the Item Author can import items written in the Markdown format into the item bank. Images of the items, if provided, will also be imported simultaneously. The Markdown format is suitable for authoring items containing programming codes because the format supports syntax highlights for various programming languages.

(F2) Importing and editing item parameters: the Item Analyzer can import them into the item bank and edits them in the system.

(F3) Maintaining items: the Item Author can search, modify, or delete items.

(F4) Classifying items: the Item Author can create the hierarchical testing criteria and uses them to classify items.

(F5) Setting parameters for the automatic test assembly procedure: The Test Form Assembler can set the test range of topics, the cutoff score, and the acceptable maximum misclassification probability.

(F6) Assembling test forms automatically: the system can select items from the item bank to assemble a test form and saves it to the test bank according to the specified parameters.

(F7) Producing test forms: Generates test forms in a printable format, such as HTML, for paper-and-pencil testing; or exports them to a format of an online testing system.

The assumptions for the system requirements analysis include the following:

- Item parameter estimations are conducted outside the system. Then, these parameters are imported into the system. R packages such as mirt [32] and ltm [33] are available for the estimations.
- Online testing is conducted outside the system. Items in a test can be exported from the proposed system into another online testing system, such as the Moodle learning management system [34], to conduct online testing.

### C. AUTOMATIC TEST ASSEMBLY PROCEDURE

To implement the automatic test assembly procedure in the system, this study employed the method proposed by [1]. This method allows items to be selected and assembled into a test form that complies with the given limit of the probability of misclassifying the mastery. The misclassification probability can be used as an indicator to evaluate the test form quality. Thus, it facilitates the lecturers to exercise regular formative assessments with precision.

The procedure is summarized in Fig. 3. The inputs include the item bank, the domain cutoff score (between 0 and 1), the indifference zone surrounding the cutoff score, and the maximum misclassification probability. The indifference zone represents the extent to which the loss of misclassification can be negligible for the test designers [35]. The procedure steps are described as follows: First, the domain cutoff score and its indifference zone bounds are converted to the points in the latent ability scale. Second, the item with the highest measurement precision near the cutoff point in the latent ability scale is selected into the item set. Third, the domain cutoff and indifference zone are computed for the current item set, given that the cutoff points and indifference zone in the latent ability scale are set. Fourth, the cutoff item number to be classified as ''mastery'' for the current set is determined. Fifth, with the cutoff item number, the probability of misclassifying as mastery is obtained. If the misclassification probability is greater than the specified acceptable value and items are still available, the procedure continues to step 2. Otherwise, the procedure stops. The outputs include the item set whose misclassification probability is less than, or equal to, the specified acceptable value and the cutoff item number meant to classify a tester as having a mastery of the topic. The procedure details are provided in the Appendix.

### D. PROCEDURE TO EVALUATE THE PROBABILITY OF MISCLASSIFYING MASTERY FOR A TEST FORM

To evaluate the quality of manually edited test forms, this study revised the [1] method to compute the misclassification probability of a given item set as shown in Fig. 4. The revised procedure follows the steps in the original method, except for steps 2 and 6, wherein, an item set, cutoff scores, indifference zone endpoints, and the test range are given. The outputs of the revised procedure include the cutoff item number and the misclassification probability for the given item set.

## IV. CONVERTING THE DOMAIN CUTOFF SCORE TO THE LATENT ABILITY SCALE WITH HETEROGENEOUS ITEMS

The method of [1] requires converting the domain score to the latent ability scale; that is, given the domain cutoff score, $\pi_0$, and the three-parameter IRT model for each item represented by $P_i(\theta)$, the method needs to find the cutoff point in the latent ability scale, $\theta_0$, such that:

$$\pi_0 = \frac{1}{N}\sum_{i=1}^{N} P_i(\theta_0). \tag{1}$$

In the case of heterogeneous items with different model parameters, drawing an analytic solution to identify $\theta_0$ is not easy. Although finding an analytic solution for equation (1) is challenging, numerical methods can be applied to solve it. This scale conversion can be treated as a root-finding problem; thus, to find $\theta_0$, the following can be used:

$$F(\theta_0) = \frac{1}{N}\sum_{i=1}^{N} P_i(\theta_0) - \pi_0 = 0 \tag{2}$$

The bisection and secant algorithms are two typical root-finding algorithms. Based on the Intermediate Value Theorem, the bisection algorithms repeatedly divide the interval where the root might exist to approximate the root of a function. The secant algorithm, a variation of Newton's method, successively identifies the roots of the secant lines to converge to the root of the function. As for the reliability of finding a root, the bisection algorithm can find the root if the interval $[a, b]$ contains the root. To ensure its convergence, the secant algorithm needs two suitable starting points for the initial secant line. Consequently, determining the two starting points is critical in using the two root-finding algorithms.

### A. SIMULATED COMPUTATION EXPERIMENT ON SOLVING THE SCALE CONVERSION PROBLEM

The study designed a simulation experiment to compare the bisection and secant algorithms applied to the scale conversion problem. As shown in Table 2, the investigation considered two factors: the root-finding method and the item bank size.

#### 1) FACTOR 1: ROOT-FINDING METHODS WITH VARIOUS STARTING POINTS

The first factor, which contained five levels, could reveal the effects of using the bisection and secant algorithms with various starting points on the computation time and the solution quality. The first level, L1_B_F, uses the bisection algorithm with a fixed start point tuple (-3, 3). The setting for the second level, marked as L2_B_D, still uses the bisection algorithm but with two dynamic starting points. Below are how these dynamic values are set.

Let $\theta_i^0$ be the latent ability for the given domain cutoff score, $\pi_0$, for item $i$. Then, the latent ability cutoff point for

**TABLE 1. Comparisons of the proposed system with the cited literature.**

| ATTRIBUTE | PROPOSED SYSTEM | Choi and McClenen [23] | Sahin et al. [26] | Yang et al. [28] |
|---|---|---|---|---|
| Test Length | Fixed | Variable | Variable | Use fixed items in the experiments; but could be extended to variable-length |
| Based testing method | Hambleton and Gruijter [1] | Computerized Adaptive Testing (CAT) | SPRT | CAT |
| Criteria to determine the fixed-length | Cutoff score, an indifference zone, and an acceptable miss-classification probability. | - | - | Not mentioned in the paper (Use 25 items in the experiments) |
| Criteria to stop the variable-length testing | - | Measurement standard error meets the requirement | Likelihood ratio less than the lower threshold or greater than the upper threshold | - |
| Item Selection | High precision item near the target latent ability | High precision item near the subject's latent ability | Not mentioned in the paper | Item with shorter memory cycle |
| Test administering | Paper-pencil or online | Online | Online | Online |

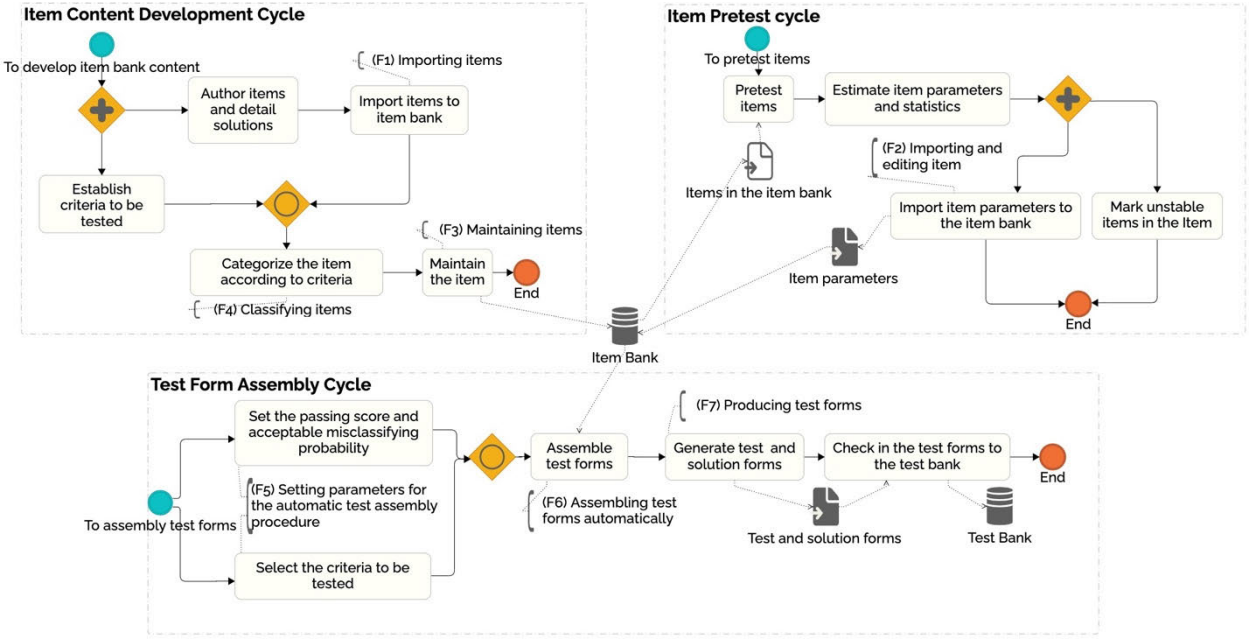



**FIGURE 2. Formalized process for applying the automatic test assembly system to assist formative assessments in mastery learning.**

the item $\theta_i^0$ can be obtained through:

$$\theta_i^0 = P_i^{-1}(\pi_0) = \frac{-1}{Da_i} \ln\left(1 - \frac{1 - c_i}{\pi_0 - c_i}\right) + b_i \qquad (3)$$

In equation (3), $\pi_0 - c_i > 0$, where $P_i^{-1}(\pi_0)$ is the inverse function of the three-parameters IRT model for item $i$. For an item bank of $N$ items, the set $\Theta = \{\theta_i^0, i = 1...N\}$ represent the distribution of the latent ability cutoff points for the items. Let $\mu_\Theta$ and $s_\Theta$ be the average and standard deviation for the elements in $\Theta$. Then the dynamic starting points for the second level are set to the tuple $(\mu_\Theta - \sigma_\Theta, \mu_\Theta + \sigma_\Theta)$.

From the third level, the focus moves to the secant algorithm and its starting points. The third level is set to the secant algorithm with the two starting points,$(\mu_\Theta, \mu_\Theta + \sigma_\Theta)$, denoted as L3_S_MS. The fourth level indicated as L4_S_R, changes the starting points to random numbers from [-3,3]. Finally, the last level, represented as L5_S_MM, sets the starting points to $(\mu_\Theta, \text{Mod}(\Theta))$, where Mod is the mode statistics of $\Theta$.

### 2) FACTOR 2: ITEM BANK SIZE

The second factor for the experiment was the size of the item bank which contained three levels of 100, 300, and 500 items.

### 3) EXPERIMENT IMPLEMENTATION

The experiment had $5 \times 3$ level combinations, and each combination replicated 150 trials. Therefore, there were 2,250 trials in the experiment. For each item bank, the IRT model's

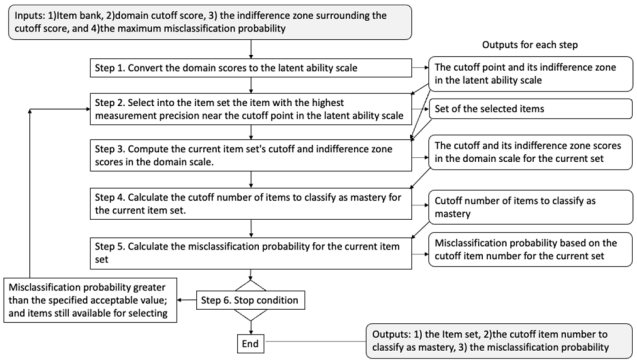



**FIGURE 3.** The procedure for assembling the items into a test form based on the method developed by [1]. The procedure generates the item set, cutoff item numbers, and the item set's probability of misclassifying as mastery.

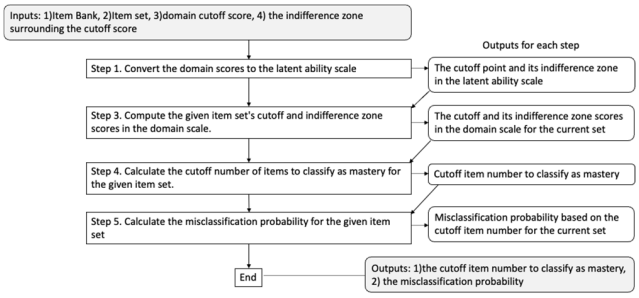



**FIGURE 4.** The modified procedure evaluates the misclassification probability of a given test form, adapted from [1].

three parameters for each item were set to random values from the following ranges: discrimination $a_i \in [0.01, 4]$, difficulty $b_i \in [-3, 3]$ and guessing $c_i \in [0.01, 0.15]$ .The experiment collected measurements of each trial's computation time and solution error. The algorithms in the experiment stopped and outputted the results if the solution error $F(\theta) \leq 10^{-5}$ was obtained. The experiments were carried out on a Mac M1 computer with a memory of 16G. This study used Python to implement the algorithms and analyze the experimental data.

### B. SIMULATED COMPUTATION RESULTS AND DISCUSSION

#### 1) ANALYSIS OF VARIANCE

The collected computation time is distributed as positive skewness, as shown in Fig. 5(a). For analyzing the variances (ANOVA), we employed the Tukey's Ladder of Powers method [36] to convert the data to a normal distribution. Fig. 5(b) shows the transformed distribution. The transformed computation time in the experiment slightly deviates from the normal distribution (Skewness - 0.036 $<$ 0, Kurtosis 2.25 $>$ 0).

Even if the data deviates slightly from the normality, using ANOVA and the t-test analysis is still appropriate. Although ANOVA requires the normality assumption, it is very robust to nonnormality data [37], [38]. The nonnormality data thresholds are |Skewness| $<$ 2 and kurtosis $<$ 7 [39]. ANOVA or even t-tests can be used within the thresholds [40]. The

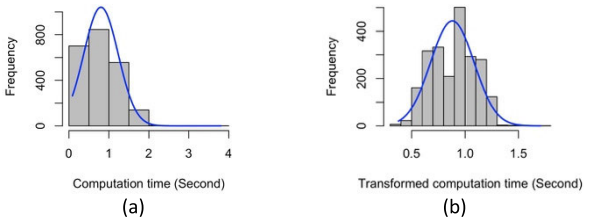



**FIGURE 5.** The original and transformed distribution of the computation time for the experiment. (a) The original distribution; (b) The transformed distribution by Tukey's Ladder of Powers method [36].

skewness and kurtosis of the transformed data were far less than the thresholds. In addition, the number of observations was quite large, with 150 trials for each factor combination. Based on the robustness of the ANOVA to the normality assumption and the sufficient number of observations, using the ANOVA and t-test for analyzing data is appropriate.

The experiment compared two ANOVA models so as to compare their fitness. The first, Model.1, only considered the effects of the two main factors: the root-finding method and the item bank size. Meanwhile, Model.2, considered the main effects and their interactions.

The Akaike information criterion (AIC) was used to select the best ANOVA model by using the AICcmodavg package [41]. The AIC prefers the model with high likelihood but penalizes the number of parameters to balance the model fitness and the number of parameters. The model with the lowest AIC value is selected. The Model.2 (AICc -3603.27) was selected as the winning model since it had a lower AIC value than Model.1 (AICc -3563.67), as shown in Table 3.

The winning Model.2, significantly fitted the data ($F(14, 2234) = 402.2$, $p < 0.01$), which explained 71.41% of the variance (adjusted R-Square0.7141). Table 4 shows the results of the ANOVA analysis and the effect sizes of factors. The $\eta^2$ was used to measure the factors' effect sizes. The effect sizes are small, medium, and large when the $\eta^2$ values are greater than the thresholds 0.01, 0.06, and 0.14, respectively [42]. Factor F1: Root-finding method ($F(4, 2234) = 44.622$, $p < 0.01$ ) was significant, and its effect size was small to medium ($\eta^2 = 0.052 < 0.06$). Factor F2: Item bank size ($F(2, 2234) = 481.8108$, $p < 0.01$) was significant as well and with a large effect size ($\eta^2 = 0.281 > 0.14$). Lastly, the interaction of two main factors was significant ($F(8, 2234) = 7.021$, $p < 0.01$), with a small effect size ($\eta^2 = 0.016 > 0.01$)

To ensure the results of the ANOVA analysis on the slightly non-normal data, the study used the nonparametric Scheirer-Ray-Hare Test [43] to verify again. Factors F1: Root-finding method ($H=286.87$, $p = .000 < 0.05$), F2: Item bank size ($H=1336.74$, $p = .000 < 0.05$) and their interaction ($H=19.78$, $p = .011 < 0.05$) were all statistically significant, which were consistent with the those of the ANOVA analysis.

The differences caused by the factor levels are further analyzed. The study used Tukey's HSD post hoc Test to identify the statistical differences between the factor levels. Then, Cohen's d was employed to measure the effect size between

**TABLE 2.** Factor level settings for the simulated computation experiment to compare the application of bisection and secant algorithms in solving the scale conversion problem.

| Factor 1 | Levels | | | | |
|---|---|---|---|---|---|
| Root-finding methods and their starting points: | L1_B_F | L2_B_D | L3_S_MS | L4_S_R | L5_S_MM |
| | Settings | | | | |
| Rooting-finding algo. | Bisection | Bisection | Secant | Secant | Secant |
| starting points | -3; 3 | $\mu_\Theta - \sigma_\Theta$; $\mu_\Theta + \sigma_\Theta$ | $\mu_\Theta$; $\mu_\Theta + \sigma_\Theta$ | Random numbers from [-3, 3] | $\mu_\Theta$; Mode($\Theta$) |
| Factor 2 | Levels | | | | |
| Item bank size (items) | 100 | 300 | 500 | - | - |

**TABLE 3.** Akaike information criterion (AIC) for comparing two ANOVA models.

| | K | AICc | Delta_AICc | AICcWt | Cum.Wt |
|---|---|---|---|---|---|
| Model.2 | 16 | -3603.27 | 0.0 | 1 | 1 |
| Model.1 | 8 | -3563.67 | 39.6 | 0 | 1 |

Model.1: Consider only the main effects.
Model.2: Consider the main and interaction effects.

**TABLE 4.** ANOVA table for the winning model considering the main and interaction effects. There were 150 replicate trials for each combination of factors.

| Response: Transformed computation time (Second) | | | | | | |
|---|---|---|---|---|---|---|
| | Sum Square | DF | F-value | p-value | $\eta^2$ | power |
| (Intercept) | 83.031 | 1 | 7093.2182 | < 2.2e-16 *** | | |
| F1: Rood-Finding Methods with Various Starting Points | 2.089 | 4 | 44.6224 | < 2.2e-16 *** | 0.052 | 1 |
| F2: Item bank size | 11.280 | 2 | 481.8108 | < 2.2e-16 *** | 0.280 | 1 |
| F1 * F2 | 0.657 | 8 | 7.0122 | 3.563e-09 *** | 0.014 | 1 |
| Residuals | 26.150 | 2234 | | | | |

Fitness of the linear model:
Residual standard error: 0.1082 on 2234 degrees of freedom
Multiple R-squared: 0.7159, Adjusted R-squared: 0.7141
F-statistic: 402.2 on 14 and 2234 DF, p-value: < 2.2e-16

the factor levels to indicate their practical significance. The thresholds for small, medium, and large effect sizes are 0.2, 0.5, and 0.8 [44], [45], respectively.

Firstly, the group means of all levels of the root-finding method factor were significantly different since all the group means had different group labels (GL). As indicated in Fig. 6(a), the secant algorithm with the random starting points (L4_S_R) had the fastest execution (GL = e) in terms of the transformed computation time, with the group means GM = 0.798, group standard deviation GSD = 0.171, and group size GN = 450. In contrast, the bisection with the dynamic starting points (L2_B_D) consumed the most time (GM = 0.987; GSD = 0.204; GN = 450; GL = a).

Table 5 shows the effect sizes between the L4_S_R level against others. The effect size between L4_S_R and L2_B_D levels was large (Cohen's d = 1.001 > 0.8). Furthermore, the table shows that the effect sizes of varying the starting point settings in the secant-based algorithm were small and negligible because the Cohen's d values of L3_S_M and L5_S_MM were 0.294 (> 0.2) and 0.175 (< 0.2), respectively. However, when using the bisection-based methods, the transformed computation time mainly increased with the

**TABLE 5.** Effect sizes for group mean differences of the root-finding method factor.

| Ref Level | Compared Level | Cohen's d | Cohen's SE | Practical Sig. |
|---|---|---|---|---|
| L4_S_R | L1_B_F | 0.781 | 0.0692 | Medium |
| | L2_B_D | 1.001 | 0.0707 | Large |
| | L3_S_MS | 0.294 | 0.0670 | Small |
| | L5_S_MM | 0.175 | 0.0668 | Negligible |

effect sizes of medium (Cohen's d 0.781 > 0.5) and large (Cohen's d 1.001 > 0.8). Therefore, the computation time can be significantly improved when using the secant-based algorithms and the secant algorithm with the random starting points (L4_S_R) can improve the most.

Secondly, regarding the effect sizes of differences in levels of the item bank size, the transformed computation time increased almost linearly, as shown in Fig. 6(b). Differences between the three levels of the factor were significant since all the levels had different group labels according to Tukey's post hoc HSD Test (alpha = 0.05) result. The level of 500 items consumed the most amount of time (GM = 1.069; GSD =

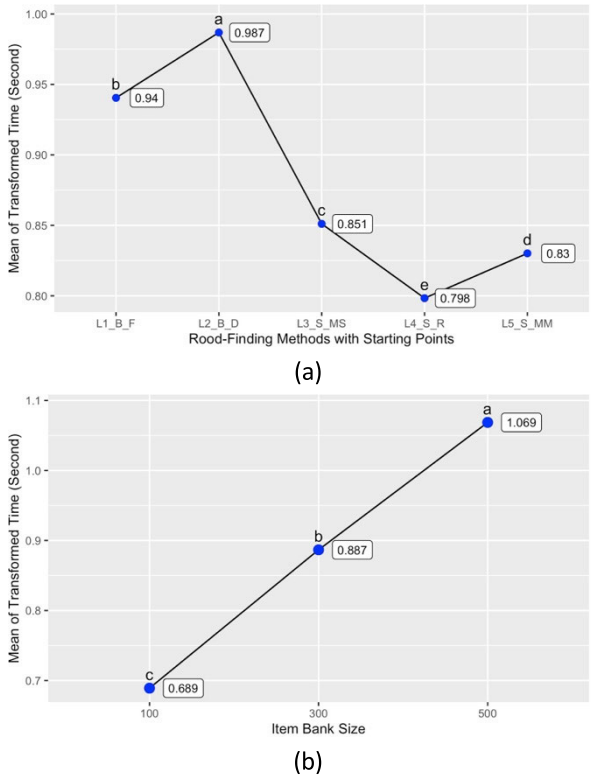



**FIGURE 6.** Plots of the two main effects. (a) Factor 1: Root-finding methods with various starting points; (b) Factor 2: Item bank size.

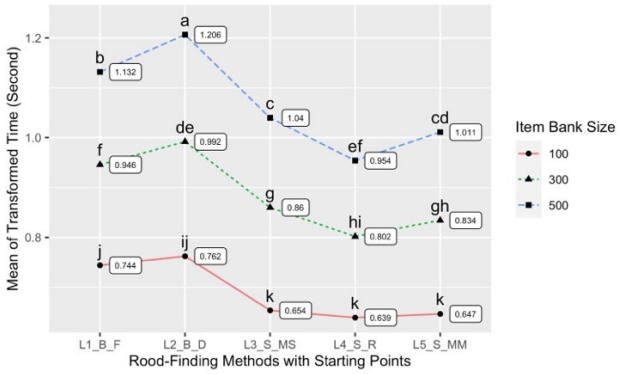



**FIGURE 7.** The plot of the interaction effects with the results of the HSD Test (alpha = 0.05) group labels on the factors of root-finding methods with starting points and the item bank size.

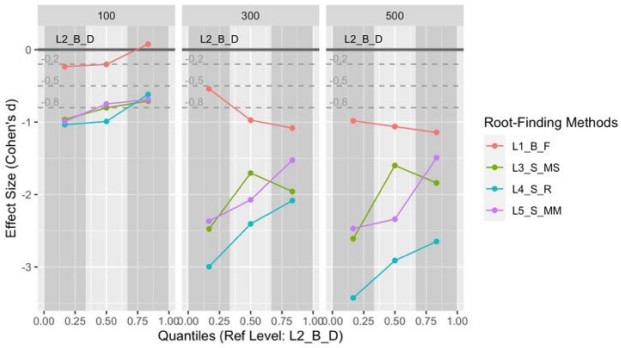



**FIGURE 8.** Effect sizes for differences in levels of the interactions between the root-finding method and the item bank size factors.

0.127; GN = 749). In contrast, the level of 100 items consumed the least (GM = 0.689; GSD = 0.156; GN = 750). The effect sizes caused by increasing the number of items from 100 to 300 and 500 were negligible (Cohen's d 1.501) and small (Cohen's d 2.665), respectively.

Thirdly, the effect sizes of differences in levels of interactions between the root-finding method and the item bank size are as shown in Fig. 7. With the post hoc test (alpha = 0.05) results, different group labels indicate significant differences in the group means. Two observations can be seen in Fig. 7. The first is that when the item bank size was 100, the setting of different starting points did not affect the transformed computation time of the secant-based methods (all the group means of L3_S_MS, L4_S_R, and L5_S_MM were marked as k). However, it increased the computation time of the bisection-based methods (GLs j and ij) in this instance. The second observation is that when the item bank size was greater than 100, different starting point settings did affect the root-finding methods' transformed computation time, and the secant-based method with random starting points was the fastest (GLs hi and ef). The slowest L2_B_D method is used as a reference point to show the effect sizes of different levels of the interactions, as shown in Fig. 8. The X-axis of the figure is the quantile of the data distribution of the L2_B_D method, and the Y-axis is Cohen's d value to measure the effect size. The negative Cohen's d value means decreasing the computation time. The dashed lines in the figure represent the thresholds for the effect sizes of small, medium, and large [44]. The figure shows that as the item bank size increases, the effect sizes of different search method levels will gradually increase. Among them, the effect sizes of L4_S_R increase the fastest.

The quality of the solution of all the root-finding methods was satisfactory, with all solution errors less than $10^{-5}$. However, the secant algorithm with starting points, $(\mu_\Theta, \text{Mode}(\Theta))$, which is the L5_S_MM level, could not find the root in one out of 150 replicate trials when the item bank was with 500 items.

### 2) DISCUSSION

According to the simulation computation results, the secant-based algorithms were superior to the bisection-based ones when the item bank size was more than 100 items. The result is consistent with the study of Ehiwario [46]. The two starting points for the secant algorithm can be set to random numbers within [-3, 3] to have satisfactory speed and reliability. No elaborate heuristics are needed to find the starting points to ensure the speed and reliability of the secant algorithm. Finally, the solution quality of all the root-finding methods was satisfied since all their errors were less than $10^{-5}$ in the experiment. Therefore, the study suggests the

secant algorithm with two random start points from [-3, 3] for solving the scale conversion problem.

## V. SYSTEM IMPLEMENTATION

### A. SQL MASTERY COURSE FOR ORACLE SQL CERTIFICATE

SQL is one of the essential programming languages for students in the Department of Information Management. Oracle Certified Association (OCA) SQL is one of the examinations with considerable difficulty and discrimination. Information companies widely recognize students with the OCA SQL certificate. A SQL mastery course has been offered by the department of one of the authors to help students prepare for the OCA SQL examination. As can be seen in Table 6, the course is organized into several modules including: basic query and DML, table joins and data set combination, advanced subquery, advanced DML, table creation and management, and creation of other schema objects.

The course administered regular formative assessments to monitor the learning progress of the students. Formative assessments using the paper-and-pencil test forms were administered when the lecturer completed a module or part of it. The test form contained question items wherein the exam test criteria related to the completed module. After the assessment, the lecturer discussed and provided detailed solutions to each question in the class as a corrective action for the students. There were eight formative assessments in the course.

In the implementation stage, this study chose paper-and-pencil testing to avoid the problems of authentication, cheating, or attrition that may be encountered in online testing. Another reason was that this testing with a fixed length is easier and more practical to implement in a course than online testing with variable lengths. If using variable-length testing, instructors must prepare personalized materials to give feedback to each student individually, which increases the teaching burden of instructors and is not easy to be implemented.

### B. IMPLEMENTING THE SYSTEM TO THE COURSE

We implemented the proposed system for the mentioned course. The implementation dramatically reduced the workload of assembling the test forms which allowed the lecturer more time to focus on teaching and giving students corrective instructions. Currently, the item bank contains about 150 items and their detailed solutions, which are organized according to the OCA SQL testing criteria [47].

The lecturer first created a framework to classify the items according to the OCA SQL test criteria by using the Topics function of the system, as shown in Fig. 9. After the classification framework was built, the lecturer collected past questions from the OCA SQL exam and authored their detailed solution in Markdown format using the text editor that supports the Markdown. These items were then imported into the system using the import function. Fig. 10(a) shows an example of using the Microsoft Visual Studio Code editor to author a question item in the Markdown format. The lecturer could preview and modify the item after it was imported, as shown in Fig. 10(b).

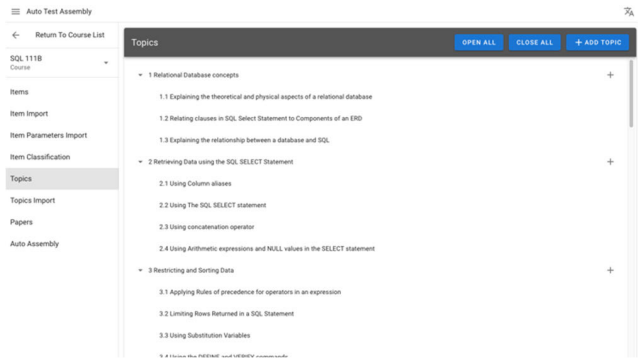
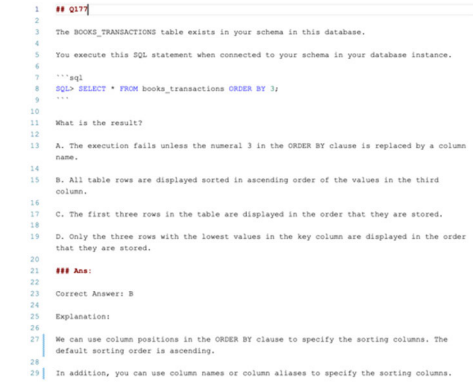
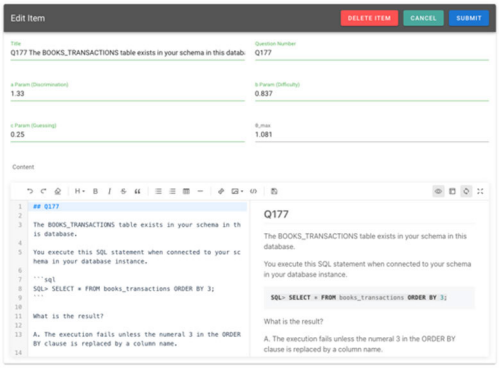



**FIGURE 9.** Creating and maintaining a framework according to the OCA SQL test criteria for classifying items.




(a)




(b)

**FIGURE 10.** Authoring a question item and its detailed solution and importing them into the system. (a) Using Microsoft Visual Studio Code editor to author a question item and its detailed solution in the Markdown format. (b) Previewing and editing the item after importing it into the system.

To classify items, the lecturer used the Item Classification function of the system. An item was dragged and dropped to a test criterion to specify its topic, as shown in Fig. 11. An item is allowed to have more than one topic.

The lecturer used the Auto Assembly function of the system to generate a test form for a formative assessment.

**TABLE 6. Modules and their chapters and related OCA SQL test criteria of the SQL mastery course.**

| Module | OCA SQL Test Criteria | Chapter | Formative Assessment |
|---|---|---|---|
| Basic query and DML | Relational Database concepts | Entity relationship model and table structure | 1 |
| | Retrieving Data using the SQL SELECT Statement | Basic SQL select statement | |
| | Restricting and Sorting Data | WHERE and ORDER BY clause | |
| | Using Single-Row Functions to Customize Output | Single-row functions for char, number and date data types | 2 |
| | Using Conversion Functions and Conditional Expressions | Data type conversion and conditional expression | |
| | Reporting Aggregated Data Using Group Function | GROUP BY and HAVING clauses | |
| | Managing Views | Creating customized data presentations via views | 3 |
| | Managing Tables using DML statements | Manipulating data by INSERT, UPDATE, and DELETE statements | |
| Table join and data set combination | Displaying Data from Multiple Tables | Join tables to combine columns from one or more tables | 4 |
| | Using SET Operators | Combining multiple query results into a data set | |
| Advanced subquery | Using Subqueries to Solve Queries | Single- and multiple-row subqueries | 5 |
| | | Multiple-column subqueries; Correlated subqueries | |
| | | Subquery factoring; recursive query | |
| Advanced DML | Managing Tables using DML statements | Using subqueries with DML statements | 6 |
| | | Using correlated subquery with DML statements | |
| Table creation and management | Managing Objects with Data Dictionary Views | Introduction to Data Dictionary View | 7 |
| | Use DDL to manage tables and their relationships | Introduction to Data Definition Language (DDL) | |
| | | Managing table constraints | |
| Creating other schema objects | Managing Indexes Synonyms and Sequences | Sequences, synonyms, and indexes | 8 |

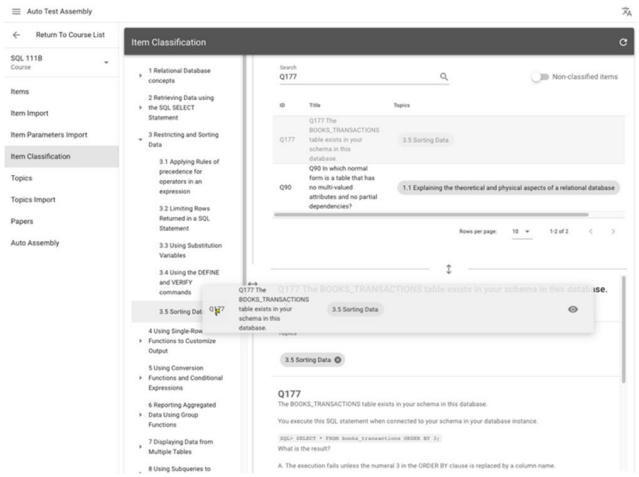



**FIGURE 11. Dragging and dropping an item to a testing criterion to specify the classification topic of the item.**

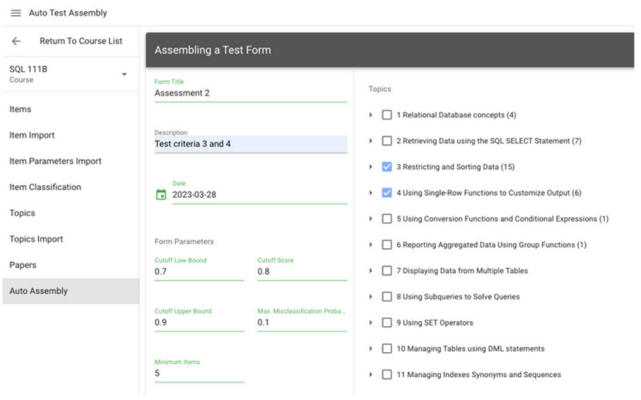



**FIGURE 12. Entering the cutoff score, indifference zone bounds, acceptable misclassification probability, and the test range to automatically assemble items to a test form.**

First, as shown in Fig. 12, the lecturer entered the cutoff score, indifference zone bounds, acceptable misclassification probability, and the test range. Next, the system assembled a test form according to the procedure in Section III-C. A list was given to show the test form lengths, cutoff item number, and misclassification probabilities for the different item sets during the procedure. The item set that met the specified misclassification probability condition was marked. As shown in Fig. 13, the resultant test form comprised seven question items and the cutoff item number six, whose misclassification probability was 0.098. Lastly, the lecturer selected the test form to be created; the system showed the selected question items and the covered test criteria of the test form for further editing.

To edit an existing test form, the lecturer used the topic navigator to select the items for a topic and added them to the test form, as shown in Fig. 14. The lecturer could further re-evaluate the test form. The system uses the procedure presented in Section III-D to compute the cutoff item number and the misclassification probability as a quality measure for

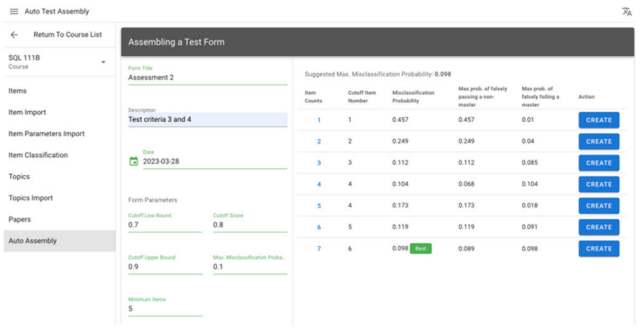



**FIGURE 13.** Generating a list showing the test form lengths, cutoff item numbers, and misclassification probabilities for different item sets during the automatic assembly procedure.

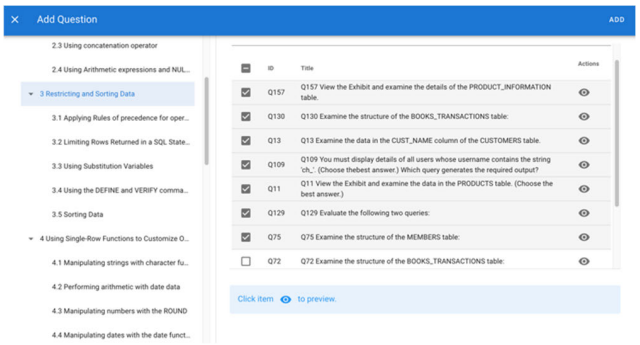



**FIGURE 14.** Selecting items by a topic and adding them to the edited test paper.

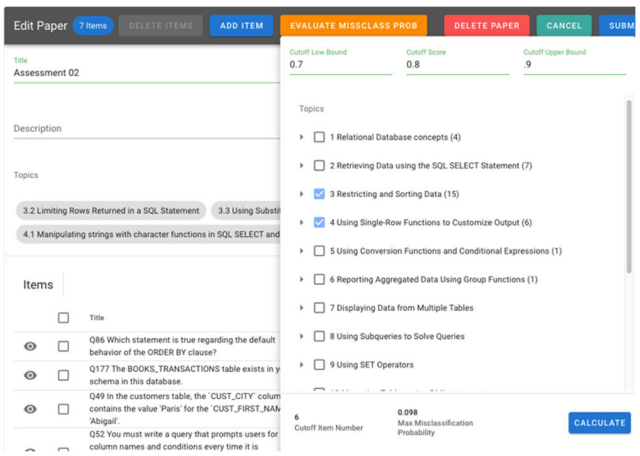



**FIGURE 15.** Entering the cutoff value and the indifference zone bounds to re-evaluate the modified test form.

the test form, provided that the cutoff score, the indifference zone endpoints, and the covered topics are given, as shown in Fig. 15. The above operations can also be applied when manually creating a test form.

To implement the proposed system practically in a course, the suggested steps are as follows:

I. Item content development

1. Create a classification structure according to the test criteria.

2. Collect or author test items.

3. Classify items according to the classification structure.

4. Import items and their classification to the system.

II. Pretest and item parameters estimation

1. Use the system to create test forms for pretesting and collect item response data.

2. Use the collected response data to estimate item parameters.

3. Import the item parameters to the system.

III. Assemble test forms and administer testing

1. Use the system to manually or automatically generate test forms for a given test range.

2. Administer tests either by paper-and-pencil or online testing. For online testing, export items in the assembled test form to an external online testing system.

According to our implementation experience, developing the test items and collecting item response data are the most time-consuming. Once the test items and the item parameters have been made ready, generating test forms becomes much easier. Using the system, test forms with measure precision can be quickly produced for given test scopes so as to assist the instructors in preparing their formative assessments for the mastery learning instruction.

## VI. CONCLUSION

### A. SUMMARY AND CONTRIBUTIONS

Using regular formative assessments to evaluate the students and give them prescriptive feedback is one of the crucial activities in the mastery learning instruction. However, preparing test forms with measurement precision for these regular formative assessments causes a burden on the lecturers and hinders the efficient execution of the mastery learning instruction.

Motivated by the need to efficiently facilitate the mastery learning instruction in teaching practice, this study developed an automatic test assembly system so as to assist the lecturers in preparing fixed-length test forms for the formative assessments in mastery learning. The system applies method [1] to automatically assemble a test form with a given misclassification probability. Also, the system can further evaluate the misclassification probability as a quality indication for a manually assembled test form. The fixed-length test forms are more appropriate for the formative assessments in the mastery learning instruction. Although they cannot adapt to the students' abilities during the assessment, the lecturers would not be overwhelmed by preparing the corrective instruction for each student with different test forms administered in one formative assessment.

During the assembly of a test form, the domain cutoff score must be converted to the latent ability scale to select items. Based on the IRT theory, the [1] method selects high-precision items near the cutoff point in the latent ability scale when the domain scale's cutoff score is given. Identifying an analytic solution to the conversion is difficult since each IRT model is heterogeneous. Instead, this study

employed numeric methods for the conversion problem. The secant and bisection root-finding algorithms for the conversion were evaluated by conducting simulation computation experiments. The empirical computing results of this study suggest that for the conversion problem, it is best to use the secant algorithm with the two starting points set to random values between [-3, 3].

The current study demonstrated the system's usefulness by implementing it in the SQL mastery course for the OCA SQL certificate. An item bank for the certification examination was built which contains about 150 multiple-choice questions collected from the Internet. These questions were arranged according to the testing criteria of the examination. When completing a few topics in the course, the lecturers could automatically and/or manually assemble items related to these topics to a test form and then administer a formative assessment to evaluate the student's mastery. The system can indicate the test form's quality by the maximum probability of misclassifying the mastery. This computerized process saved a lot of manual item selection and test form editing, which also reduced the burden of assembling the test forms. Hence, it assisted the lecturers and provided them with more time to execute the mastery learning instruction.

The contributions of this study to the engineering education practice are several folds. First, the study proposed an automatic test assembly system to computerize the process from the item bank management to assembling the test forms. The system can help lecturers exercise regular formative assessments with quality test forms in the mastery learning teaching. In addition, the study investigated the numeric methods of converting the domain cutoff score to the latent ability scale with items of heterogenous IRT models to apply method [1] in practice. Finally, the system source codes are open to the engineering education community and can be accessed from the following URL: https://github.com/hychen39/TestFormMakerAuto.

The implications of the study are three-fold. First, in addition to caring about students' learning effects, the methods or systems that can reduce the teaching burdens should be also addressed so as to improve the engineering education. In this way, the teachers have more time to devote to improving teaching quality. Second, when developing a formative assessment system, we should be careful as to whether the questions are suitable for online assessment. Questions requiring a high cognitive load are more appropriate on a paper-based test than a computer-based test regarding assessment and learning. Third, the implementation feasibility of the formative assessment systems in classrooms should be considered. Although personalized tests and learning feedback greatly benefit students, it causes a considerable burden on the teachers. Finding a balance between the two is a crucial issue worth considering.

### B. LIMITATIONS AND FUTURE WORKS

Nevertheless, there are a few limitations to the proposed system. First, the feature of the IRT parameter estimation is not included. Second, the feature of online testing is not implemented in the system. Not including these two features is because there are already existing packages or systems. For example, the R mirt package [32] or ltm package [33] can be used for the IRT parameter estimation. The Moodle learning management system provides online testing features. Items that meet its XML format can be imported to Moodle's quiz module for online testing [34]. Despite the limitations, the system is still beneficial to the mastery learning instruction. Estimated IRT parameters can be imported to the proposed system, and the test form in the system can be exported to the existing online testing systems. Hence, the system can collaborate with other systems of the IRT parameter estimation and online testing so as to facilitate the lecturers in exercising the mastery learning instruction.

Future works should consider more system features to facilitate the formative assessments in the mastery learning instruction for the lecturers. Firstly, one can include the ratio of items in the test form assembly procedure so as to ensure the diversity of item topics in a test form. Secondly, one could consider the ratio of items that overlap between the two test forms for linking them. Finally, controlling the number of exposures of items could be included to ensure that each item appears in at least one or more test forms.

## APPENDIX: DETAILED STEPS OF THE AUTOMATIC TEST ASSEMBLY PROCEDURE [1]

The following notation denotes the inputs to the method:

- $B = \{i | i = 1 \ldots N\}$ : an item bank containing $N$ items, where $i$ represents the index for an item.
- $\pi_0$ : the test's cutoff score in the domain scale. $0 \leq \pi_0 \leq 1$.
- $[\pi_l, \pi_u]$ : an indifference zone such that $\pi_l \leq \pi_0 \leq \pi_u$, where the loss of the misclassification can be neglected. $0 \leq \pi_{\mathrm{l}}, \pi_{\mathrm{u}} \leq 1$.
- $\bar{P}$ : the acceptable Misclassification Probability (MP) specified by the test designer.

The following notation denotes the outputs of the method:

- $S_T = \{i | i = 1 \ldots M\}$ : the set of the selected $M$ items.
- $n_0$ : the number of items answered correctly to classify as mastery.
- $P_k\,(n_0)$ : the MP for the current item set with $k$ items.

Additionally, this study set the calculation precision at five decimal places when implementing the method.

*Step 1*: Convert the domain cutoff score to the latent ability scale in IRT. Convert the scores of the cutoff and indifference zone boundaries to the latent ability scale for an item bank containing $N$ items. That is,

$$(\pi_l, \pi_0, \pi_u) \underset{N \text{ items}}{\rightarrow} (\theta_l, \theta_0, \theta_u) \tag{4}$$

The relationship between a domain score $\pi$ and a latent ability $\theta$ is defined as the following according to [48]:

$$\pi = \frac{1}{N} \sum_{i=1}^{N} P_i\,(\theta) \tag{5}$$

where $P_i(\theta)$ denotes the probability of correctly answering item $i$ for an examinee with latent ability $\theta$. The study employed the three-parameter IRT model as the probability model for $P_i(\theta)$. Equation (5) can be used to convert the cutoff score from the domain scale to the latent ability scale. For a given cutoff score $\pi$ in the domain scale, one can use a root-finding numeric method to solve (5) to find the corresponding latent ability $\theta$.

*Step 2*: Select an item. Select items with high measurement precision near the cutoff score in the latent ability score ($\theta_0$), then add them to the item set $S_T$. The measure precision for an item is computed as:

$$I_i(\theta_i) = \frac{D^2 a_i^2 (1 - c_i)}{\left(c_i + e^{Da_i(\theta_i - b_i)}\right)\left(1 + e^{-Da_i(\theta_i - b_i)}\right)^2} \tag{6}$$

where

$$\theta_i = b_i + \frac{1}{Da_i} \ln .5(1 + \sqrt{1 + 8c_i}). \tag{7}$$

*Step 3*: Compute the cutoff and indifference zone scores in the domain scale for the current item set $S_T$. Denote $\pi_0'$ as the cutoff score and $[\pi_l', \pi_u']$ as the indifference zone in the domain scale respectively. Assume $S_T$ has $k$ items. Then, $\pi_0'$ can be obtained by using equation (5) with then given $\theta_0$:

$$\pi_0' = \frac{1}{k} \sum_{i=1}^{k} P_i(\theta_0) \tag{8}$$

With $\theta_l$ and $\theta_u$, $[\pi_l', \pi_u']$ can be obtained similarly for the current item set. The $\theta_0$, $\theta_l$ and $\theta_u$ are obtained in Step 1.

*Step 4*: Calculate the number of items answered correctly to classify as mastery. Given $\pi_0'$, the cutoff item number for mastery, $n_0$, is an integer near $k\pi_0'$, which is based on [1].

*Step 5*: Calculate the misclassification probability for the current item set. Given $n_0$, the probabilities of correctly classifying not mastery and mastery can be computed using equations (9) and (10) respectively.

$$\alpha(n_0) = \sum_{x=0}^{n_0 - 1} \binom{k}{x} (\pi_l')^x (1 - \pi_l')^{k-x} \tag{9}$$

$$\beta(n_0) = \sum_{x=n_0}^{k} \binom{k}{x} (\pi_u')^x (1 - \pi_u')^{k-x} \tag{10}$$

Consequently, the MP for the current item set with $k$ items is obtained through:

$$P_k(n_0) = max 1 - \alpha(n_0), 1 - \beta(n_0) \tag{11}$$

*Step 6*: Stop conditions. Go back to Step 2 if the MP for the current item set is greater than the specified value, that is, $P_k(n_0) > \bar{P}$, and when items are still available in the item bank. Otherwise, end the procedure. Assume $M$ items have been selected at the end. If the procedure ends with the condition of $P_k(n_0) \leq \bar{P}$, the length of the test form is $|S_T|$, and the cutoff item number is $n_0$ with the MP value of $P_M(n_0)$.

## ACKNOWLEDGMENT

The authors thank the reviewers for taking the time and effort to review the manuscript. They sincerely appreciate all the valuable comments and suggestions that helped them improve the manuscript's quality.

## REFERENCES

[1] R. K. Hambleton and N. M. De Gruijter, "Application of item response models to criterion-referenced test item selection," *J. Educ. Meas.*, vol. 20, no. 4, pp. 355–367, 1983. [Online]. Available: https://www.jstor.org/stable/1434952

[2] C.-L.-C. Kulik and J. A. Kulik, "Mastery testing and student learning: A meta-analysis," *J. Educ. Technol. Syst.*, vol. 15, no. 3, pp. 325–345, Mar. 1987, doi: 10.2190/fg7x-7q9v-jx8m-rdjp.

[3] D. S. Roberts, R. R. Ingram, S. A. Flack, and R. J. Hayes, "Implementation of mastery learning in nursing education," *J. Nursing Educ.*, vol. 52, no. 4, pp. 234–237, Apr. 2013.

[4] M. S. Lipsky, C. J. Cone, S. Watson, P. T. Lawrence, and M. N. Lutfiyya, "Mastery learning in a bachelor's of nursing program: The Roseman University of Health Sciences experience," *BMC Nursing*, vol. 18, no. 1, p. 52, Dec. 2019, doi: 10.1186/s12912-019-0371-x.

[5] T. R. Guskey, "Lessons of mastery learning," *Educ. Leader.*, vol. 68, no. 2, pp. 52–57, 2010.

[6] H.-T. Lin, E. Z.-F. Liu, and S.-M. Yuan, "An implementation of web-based mastery learning system," *Int. J. Instr. Media.*, vol. 35, no. 2, pp. 209–221, 2008.

[7] B. S. Bloom, "Learning for mastery," *Eval. Comment*, vol. 1, no. 2, pp. 1–12, 1968.

[8] M. Winget and A. M. Persky, "A practical review of mastery learning," *Amer. J. Pharmaceutical Educ.*, vol. 86, no. 10, Dec. 2022, Art. no. ajpe8906, doi: 10.5688/ajpe8906.

[9] C.-L.-C. Kulik, J. A. Kulik, and R. L. Bangert-Drowns, "Effectiveness of mastery learning programs: A meta-analysis," *Rev. Educ. Res.*, vol. 60, no. 2, pp. 265–299, Jun. 1990.

[10] T. R. Guskey and T. D. Pigott, "Research on group-based mastery learning programs: A meta-analysis," *J. Educ. Res.*, vol. 81, no. 4, pp. 197–216, Mar. 1988, doi: 10.1080/00220671.1988.10885824.

[11] J. Garner, P. Denny, and A. Luxton-Reilly, "Mastery learning in computer science education," in *Proc. 21st Australas. Com. Educ. Conf.*, 2019, pp. 37–46.

[12] R. Flaugher, "Item pools," in *Computerized Adaptive Testing: A Primer*, 2nd ed., H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, Eds. Mahwah, NJ, USA: Lawrence Erllbaum Associates, 2000.

[13] R. M. Anzaldua, "Item banks: What, where, why and how," in *Proc. Annu. Meeting Southwest Educ. Res. Assoc.*, Austin, TX, USA, 2002, pp. 1–32. [Online]. Available: https://files.eric.ed.gov/fulltext/ED462435.pdf

[14] M. J. Kolen, R. L. Brennan, and R. L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd ed. New York, NY, USA: Springer-Verlag, 2004.

[15] N. A. Thompson and D. A. Weiss, "A framework for the development of computerized adaptiv tests," *Practical. Assess. Res. Eval.*, vol. 16, no. 1, pp. 1–9, 2011.

[16] A. Birnbaum, "Estimation of an ability," in *Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick, Eds. Reading, MA, USA: Addison-Wesley, 1968.

[17] W. J. Linden, *Linear Models for Optimal Test Design*. New York, NY, USA: Springer-Verlag, 2005.

[18] K. Fuchimoto, T. Ishii, and M. Ueno, "Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly," *IEEE Trans. Learn. Technol.*, vol. 15, no. 2, pp. 252–264, Apr. 2022, doi: 10.1109/tlt.2022.3163360.

[19] R. M. Luecht and S. G. Sireci, "A review of models for computer-based testing," College Board, New York, NY, USA, Tech. Rep., 2011. [Online]. Available: https://files.eric.ed.gov/fulltext/ED562580.pdf and https://eric.ed.gov/?q=A+review+of+models+for+computer-based+testing&id=ED562580

[20] R. M. Luecht and R. J. Nungester, "Computer-adaptive sequential testing," in *Computerized Adaptive Testing: Theory and Practice*, W. J. van der Lindenand G. A. W. Glas, Eds. Dordrecht, The Netherlands: Springer-Verlag, 2000, pp. 117–128.

[21] L. Xu, S. Wang, Y. Cai, and D. Tu, "The automated test assembly and routing rule for multistage adaptive testing with multidimensional item response theory," *J. Educ. Meas.*, vol. 58, no. 4, pp. 538–563, Dec. 2021, doi: 10.1111/jedm.12305.
[22] D. J. Weiss and G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *J. Educ. Meas.*, vol. 21, no. 4, pp. 361–375, Dec. 1984, doi: 10.1111/j.1745-3984.1984.tb01040.x.
[23] Y. Choi and C. McClenen, "Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks," *Appl. Sci.*, vol. 10, no. 22, p. 8196, Nov. 2020, doi: 10.3390/app10228196.
[24] B. G. Dodd, W. R. Koch, and R. J. De Ayala, "Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules," *Educ. Psychol. Meas.*, vol. 53, no. 1, pp. 61–77, Mar. 1993, doi: 10.1177/0013164493053001005.
[25] S. W. Choi, M. W. Grady, and B. G. Dodd, "A new stopping rule for computerized adaptive testing," *Educ. Psychol. Meas.*, vol. 71, no. 1, pp. 37–53, Feb. 2011, doi: 10.1177/0013164410387338.
[26] M. Sahin, F. Aydin, S. Sulak, C. T. Müftüoglu, M. Tepgeç, G. K. Yilmaz, R. Yilma, and H. Yurdugü, "Using adaptive mastery testing in assessment management systems," in *Proc. 18th Int. Conf. Corn. Explore. Learn. Digit. Age, Online*, 2021, pp. 205–211.
[27] G. G. Kingsbury and D. J. Weiss, "A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure," in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, D. J. Weiss, Ed., New York, NY, USA: Academic Press, 1983, pp. 257–283.
[28] A. C. M. Yang, B. Flanagan, and H. Ogata, "Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning," *Comput. Educ., Artif. Intell.*, vol. 3, Jan. 2022, Art. no. 100104, doi: 10.1016/j.caeai.2022.100104.
[29] R. Carpenter and T. Alloway, "Computer versus paper-based testing: Are they equivalent when it comes to working memory?" *J. Psychoeducational Assessment*, vol. 37, no. 3, pp. 382–394, Jun. 2019, doi: 10.1177/0734282918761496.
[30] L. Smolinsky, B. D. Marx, G. Olafsson, and Y. A. Ma, "Computer-based and paper-and-pencil tests: A study in calculus for STEM majors," *J. Educ. Comput. Res.*, vol. 58, no. 7, pp. 1256–1278, Dec. 2020, doi: 10.1177/0735633120930235.
[31] D. Bengs, U. Kroehne, and U. Brefeld, "Simultaneous constrained adaptive item selection for group-based testing," *J. Educ. Meas.*, vol. 58, no. 2, pp. 236–261, Jun. 2021, doi: 10.1111/jedm.12285.
[32] R. P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," *J. Stat. Softw.*, vol. 48, no. 6, pp. 1–29, 2012, doi: 10.18637/jss.v048.i06.
[33] D. Rizopoulos. (2018). *Package 'LTM': Latent Trait Models Under IRT*. [Online]. Available: https://cran.r-project.org/web/packages/ltm/ltm.pdf
[34] Moodle. *Import Questions—MoodleDocs*. Accessed: Jun. 15, 2023. [Online.] Available: https://docs.moodle.org/402/en/Import_questions
[35] R. R. Wilcox, "A note on the length and passing score of a mastery test," *J. Educ. Statist.*, vol. 1, no. 4, p. 359, 1976, doi: 10.2307/1164988.
[36] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.
[37] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, "Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption," *Methodology*, vol. 6, no. 4, pp. 147–151, Jan. 2010.
[38] M. J. Blanca, R. Alarcón, J. Arnau, R. Bono, and R. Bendayan, "Non-normal data: Is ANOVA still a valid option?" *Psicothema*, vol. 29, no. 4, pp. 552–557, 2017, doi: 10.7334/psicothema2016.383.
[39] P. J. Curran, S. G. West, and J. F. Finch, "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis," *Psychol. Methods*, vol. 1, no. 1, pp. 16–29, Mar. 1996.
[40] D. C. Howell, *Statistical Methods for Psychology*, 5th ed. Australia, Oceania: Duxbury-Thomson Learning, 2002.
[41] M. J. Mazerolle. (2023). *AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c)*. [Online]. Available: https://CRAN.R-project.org/package=AICcmodavg
[42] J. Miles and M. Shevlin, *Applying Regression and Correlation: A Guide for Students and Researchers*. London, U.K.: SAGE, 2001.
[43] R. R. Sokal and F. J. Rohlf, *Biometry*, 3rd ed. New York, NY, USA: W.H. Freeman, 1995.
[44] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.
[45] G. M. Sullivan and R. Feinn, "Using effect size—Or why the *P* value is not enough," *J. Graduate Med. Educ.*, vol. 4, no. 3, pp. 279–282, Sep. 2012.
[46] J. C. Ehiwario and S. O. Aghamie, "Comparative study of Bisection, Newton–Raphson and secant methods of root-finding problems," *IOSR J. Eng.*, vol. 4, no. 4, pp. 1–7, Apr. 2014, doi: 10.9790/3021-04410107.
[47] *Oracle Database SQL Exam Number: 1Z0-071*, Oracle University. Accessed: Sep. 10, 2022. [Online.] Available: https://education.oracle.com/oracle-database-sql/pexam_1Z0-071
[48] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1980.

**HUNG-YI CHEN** received the Ph.D. degree in industrial engineering and management from the National Yunlin University of Science and Technology, Yunlin, Taiwan.

He is currently an Associate Professor with the Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan. His current research interests include information systems, supply chain and logistics management, and engineering education.

**YING-CHIEH LIU** received the Ph.D. degree in management information systems from Edith Cowan University, Australia.

He is currently an Associate Professor with the Department of Information Management, Chin-Yi University of Technology, Taichung, Taiwan. His current research interests include electronic commerce, online learning, e-health, and virtual teams.

**XUAN-QI LIU** received the M.S. degree in information management from the Chaoyang University of Technology, Taichung, Taiwan.

He is currently a Web Developer with Digi-Win Software Company, Taichung. His current research interest includes full-stack web development by leveraging modern frameworks.

**VICTORIA CHIU** received the Ph.D. degree in accounting from the Rutgers Business School, NJ, USA.

She is currently an Associate Professor of accounting with the Department of Accounting, Finance and Law, State University of New York at Oswego. Her research has been published in multiple journals, including *International Journal of Accounting Information Systems*, *Issues in Accounting Education*, *Journal of Accounting Literature*, *Eurasian Journal of Business and Economics*, *Operations Research Perspectives*, *Management and Production Engineering Review*, *Journal of Applied Research and Technology*, *International Journal of Industrial Engineering Computations*, *International Journal of Mathematical, Engineering and Management Sciences*, and *Journal of Emerging Technologies in Accounting*.