Received: 3 August 2017 | Revised: 5 April 2018 | Accepted: 22 June 2018

DOI: 10.1111/exsy.12331

ORIGINAL ARTICLE

WILEY Expert Systems

Discovering the associations between test items in an examination

Ligang Dong | Liujun Tang | Shihuan Liu | Xian Jiang

School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, China

Correspondence
Ligang Dong, School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, China.
Email: donglg@zjgsu.edu.cn

Funding information
the Graduate Scientific and Technological Innovation Projects Fund Project in ZJGSU, Grant/Award Number: 16020000293; the Key Research & Development Plan of Zhejiang, Grant/Award Number: 2017C03058; Zhejiang Provincial Key Laboratory of New Network Standards and Technologies (NNST), Grant/Award Number: No.2013E10012

Abstract

An examination is .an important approach to comprehensively evaluate students' abilities. An examination paper is composed of test items. Most studies assume that test items are independent, but there may, in fact, be associations among test items. How to select test items with low correlations into a test paper is our concern. Meanwhile, for diagnostic learning, students must perform highly relevant exercises to help them efficiently understand the concepts included in the test items. People usually find the associations among test items based on properties of individual test items. This paper proposes a test item association model to represent the correlations of test item pairs by mining the response data of examinees. Then, a hypothesis test and an empirical analysis were used to verify and analyse the relationship between the test item correlation and similarity. Next, the effects of test item parameters on the correlation are investigated by applying the model to the simulated response data which are produced by Item Response Theory (IRT) simulation data generation software. Finally, three applications of this test item association model are listed. After verification, the proposed model can be used for precise teaching.

KEYWORDS

correlation coefficient, data mining, item pairs, precision teaching

1 | INTRODUCTION

The widespread application of mobile smart devices, the rapid development of Internet devices, and rapid reductions in traditional storage costs have resulted in the creation of massive datasets. The whole Internet industry gradually takes "data" as the core of constructing the new era of data technology. Techniques such as data analysis, data mining, and data visualization have laid a solid foundation for the prosperity of big data (Chen & Zhang, 2014). The number of studies on and applications of big data are growing rapidly (Chen, Mao, & Liu, 2014). Big data is also infiltrating many non-computer domains such as medicine, business, and education (Nepal & Bouguettaya, 2013).

In the field of education, numerous scholars have used big data techniques to analyse education data (Al-Razgan, Al-Khalifa, & Al-Khalifa, 2014). Education data are formed by or related to human learning activities in educational backgrounds. Test items and examinee response data are components of education data. Therefore, finding associations between test items is necessary for educational data mining. It is a simple fact that the associations between test items could be found by analysing the contents of the test items themselves. As the number of test items increases, finding the associations among them manually becomes impossible. The efficiency of the manual method is quite low; therefore, data mining techniques appear to be suitable for addressing this problem in the modern machine intelligence era. Among the data mining techniques, association rules can be used to analyse the associations between test items. If the test items satisfy the user-specified minimum support (min_sup) and minimum confidence (min_conf), the test items are supposed to be related to each other, which is called association rules mining (ARM; Sharma et al., 2010a). However, few research studies are specifically concerned with test item associations (TIAs). Instead, such studies are always combined with other topics such as test construction (Cen et al., 2010; Chu, Hwang, Tseng, & And, 2006; Hwang, Hsiao, & Tseng, 2008; Liu & Chen, 2012; Tseng, Sue, Su, Weng, & Tsai, 2007).

Liu, Ma, and Wong (2001) pointed out that some defects exist in the ARM. Also, in the analysis of associations between test items, the support and confidence are not suitable for judging the correlation between test items. Therefore, the associations between test items are described by the Pearson correlation coefficient in this paper, whereas this coefficient was mainly used to describe species correlation (Zhang, 2005); user correlation (Sheugh & Alizadeh, 2015); and others before.

In this paper, we propose a new model to discover the correlations between test items based on examinees' response data. In this model, we use a data mining algorithm to generate all test item pairs and use a modified Pearson correlation coefficient equation (Xiong, Shekhar, Tan, & Kumar, 2004) to quantify the correlation between two paired test items. Then, the model is applied to the simulated examinee response data based on item response theory (IRT; Han, 2007) to identify how the parameters of the test items affect the correlation. Additionally, by using this model, we determine the correlations of different test item pairs from real examinee response data and apply those correlations to the diagnostic learning. For example, this information can help students enhance their performance in their weaker subjects. Finally, we consider that knowing the correlation between the test items could function as an impact factor for evaluating examination paper quality after using the model for real data.

The remainder of this paper is organized as follows. First, related work is presented (Section 2) before the TIA model is explained (Section 3). Second, Section 4 uses the hypothesis test and the empirical analysis to verify and analyse the relationship between the test item correlation and similarity. Third, Section 5 identifies the parameters of test items that affect the correlation. Then, Section 6 gives some applications based on the TIA model. Finally, conclusions and suggestions for future researches are drawn.

2 | RELATED WORK

One important method for finding associations between test items is the ARM. ARM is a type of data mining method to analyse and describe the potential, valuable relationships between patterns in a dataset. ARM was first postulated for analysing market baskets to determine buying patterns in the process of selling goods. Additionally, based on ARM, the Apriori algorithm was proposed (Agrawal, Imieliński, & Swami, 1993). Continued developments after ARM was proposed have resulted in numerous ARM algorithms, such as the famous FP-growth algorithm (Han, Pei, & Yin, 2000) and the Eclat algorithm (Zaki, 2000).

A large number of studies have applied association rules to find the relationships between test items. For example, Liu and Chen (2012) used a modified Apriori algorithm to provide the detail process of ARM for test item substitution. First, the rules of test item pairs from the same subject are found by ARM. It is evident that the test items in a pair have a strong association. Then, the test item can be substituted by the other item in the same pair to make backup test papers. Pechenizkiy, Calders, Vasilyeva, and Bra (2008) used ARM to analyse interesting and unexpected patterns in student feedback preferences or performance based on their response on a real test in order to determine how well test items are designed or tailored toward individual student needs. Romero, Zafra, Luna, and Ventura (2013) proposed a grammar-based algorithm to mine association rules from multiple-choice quiz data. Then, interesting rules were discovered to aid the instructors in decision making about how to improve both the quiz and the corresponding course containing the concepts evaluated by the quiz. The rules discovered are applied to show a list of updates in a specific quiz and course. Angeline (2013) used the Apriori algorithm to predict and improve student performance based on student participation in homework and evaluation tests.

Some studies have applied association rules to find the relationships between test items and concepts. For example, Hwang et al. (2008) proposed a concept-effect relationship model based on ARM to find the poorly learned and well-learned concepts of a student and how it generates learning guidance. The test response data collected from every concept are analysed by ARM to find the relationship between test items and concepts. In this way, the students' mastery of concepts can be known. Then, the guidance can be provided. Chu et al. (2006) proposed a concept-effect propagation approach to detect student learning problems. First, a concept-effect propagation table is used to model the relationships between concepts. Then, based on the response data from each concept, the degree of association between test items and concepts is obtained. According to the association, some targeted learning would be realized. In order to provide adaptive learning guidance for learners, Tseng et al. (2007) proposed a two-phase concept map-construction approach to automatically construct the concept map by learner historical testing records. First, a data mining approach is used to find fuzzy association rules between test items and concepts. Then, the multiple rule types are used to further analyse the above mined rules, and a heuristic algorithm is used to automatically construct the concept map. Panjaburee, Hwang, Triampo, and Shih (2010) proposed a set of weighting rules to define the weighting or degree of relevance for each concept to each test item. Abdullah, Herawan, Ahmad, Ghazali, and Deris (2014) used indirect least association rules to analyse student examination datasets as a basis to improve their teaching and learning strategies.

Additionally, other approaches have been proposed to describe relationships. For example, Barnes (2005) proposed the q-matrix method to create concept models to both understand student behaviours and direct learning paths for future students based on the student response data from each concept. Sun (2012) found the most dependent test items in student response data by adopting the concept of entropy from information theory. A distance metric is defined to measure the amount of mutual independency between two items, and it is used to quantify how independent two items are in a test.

From the studies above, we can find that a large number of studies used association rules to describe the relationships between test items. If test items satisfy an association rule, they are correlative. However, the exact value of correlation between them cannot be obtained, so the test

items with strong correlation also cannot be found. Some studies used a quantitative method to describe the degree of relationship between test items. In this paper, we propose that the test item correlation is described by the Pearson correlation coefficient, whereas this coefficient is mainly used to describe species correlation (Zhang, 2005); user correlation (Sheugh & Alizadeh, 2015); and others before. From the above, we design a model that performs the following tasks:

- Generating test item pairs from student response data based on ARM.
- Using a modified Pearson correlation coefficient to quantify the association between two different test items.

This model is applied to both simulated and real response data.

3 | THE TIA MODEL

The model introduced in this paper is called the TIA. The TIA is designed to find the correlation between two different test items based on examinee response records. The model requires a data-mining algorithm to generate test item pairs from real response data. Additionally, choosing a coefficient to quantize the correlation of test items in a pair is the key to applying the TIA model in practice. In section 3.1, the concept of test item pairs will be introduced. The procedure of selecting an appropriate coefficient is described in section 3.2. Section 3.3 proposes the algorithm in the TIA model and explains its principle.

3.1 | Test item pairs in response data

Assume there are m items in a test. Let $Q = \{q_1, q_2, \ldots, q_m\}$ be an itemset. Let T be a set of transactions, where the number of transactions depends on the number of examinees. This study does not consider the content of the test items themselves; instead, it focuses on the examinee response data. The default number of response categories is 2 in this paper. Therefore, the response data consist of correct and wrong answers, where 0 represents a wrong answer and 1 represents a correct answer. When there are n examinees, the dataset can be considered as a matrix of n rows and m columns. An example matrix is shown in Figure 1.

Taking out the indexes of correct or incorrect response data and using the indexes as the final dataset, the dataset can be a collection of all correct or incorrect responses. Suppose the data belong to all incorrect responses and $m = 10$ and $n = 10$. In this case, the response data are shown in Figure 2.

In Figure 2, the left column shows the number of 10 examinees in a test, whereas the right column is the index values of the examinees who answered incorrectly. By extracting all pairs from the first record according to the combinatorial number theory, the item pairs are (3, 4), (3, 5), (3, 6), (3, 7), (4, 5), (4, 6), (4, 7), (5, 6), (5, 7), and (6, 7). Therefore, all item pairs in a test are acquired from the incorrect response data of each examinee. The example clearly illustrates the concept of test item pairs. We will use these test item pairs for further research in the next sections.

3.2 | Selection of correlation coefficient

In traditional ARM, researchers typically use the support and the confidence to judge the association of different item sets. Let T be the total number of examinee response records. Suppose there are two test items, A and B. Let $A\&B$ be a test item pair. Then, the support is shown as

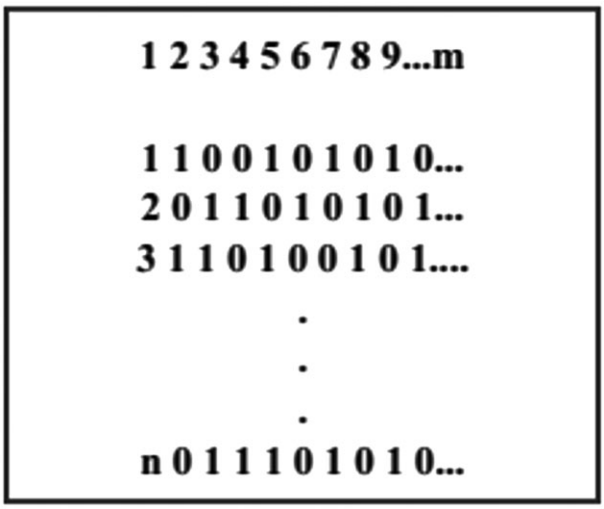

FIGURE 1 Matrix of response data

1	1100000111	3 4 5 6 7
2	0000011000	1 2 3 4 5 8 9 10
3	0000000101	1 2 3 4 5 6 7 9
4	1101101111	3 6
5	1010101001	2 4 6 8 9
6	1101001000	3 5 6 8 9 10
7	1111001111	5 6
8	1001001110	2 3 5 6 10
9	1111000100	5 6 7 9 10
10	1110100001	4 6 7 8 9

FIGURE 2 Response data when m = 10 and n = 10

$$\sup(A, B) = \mathrm{P}(A\&B). \tag{1}$$

In test item pair $A\&B$, the support is the count of $A\&B$ divided by T.

In traditional ARM, the support provides a threshold to check whether items can meet a requirement. Additionally, one of the standards is to judge the association among test items. However, in test items, only using the support to describe the association of two different items is not completely suitable. For instance, suppose that 2,000 examinees take part in a test. Items Q_1 and Q_2 are two test items in the test. Let us consider the following two cases.

- Half of the examinees (1,000) provide incorrect responses to Q_1, and half (1,000) provide incorrect responses to Q_2. Moreover, 500 examinees give incorrect responses to both Q_1 and Q_2. The support of $Q_1\&Q_2$ is 0.25.
- Of the 2,000 examinees, 600 examinees provide incorrect responses to Q_1, 600 examinees provide incorrect responses to Q_2, and 500 examinees provide incorrect responses to both Q_1 and Q_2. The support of $Q_1\&Q_2$ is still 0.25.

The correlation between Q_1 and Q_2 in the second case is obviously stronger than that in the first case. Based on these two cases, the support is not an appropriate coefficient for judging the correlations between test items; consequently, investigating the correlation between test items cannot rely solely on the support.

Suppose there are two test items, A and B. Let $A\&B$ be an item pair. Then, the confidence is shown as

$$\text{confidence}(B = >A) = \frac{P(A\&B)}{P(A)} = P(B|A). \tag{2}$$

In item pair $A\&B$, the confidence is the count of $A\&B$ divided by the count of item A. In Formula (2), the symbol of "=>" represents a rule for items A and B that if the item A is answered correctly, the item B will also be answered correctly (Zhou & Yau, 2007). And in the following paper, the meaning of "=>" is the same as that in Formula (2).

Consider the following two cases.

- Of the 2,000 examinees, 800 examinees provide incorrect responses to Q_1, and 1,000 examinees provide incorrect responses to Q_2. Among them, 500 examinees provide incorrect responses to both Q_1 and Q_2. If the rule for the items Q_1 and Q_2 is Q_1 => Q_2, the confidence of $Q_1\&Q_2$ is 0.625. If the rule for the items Q_1 and Q_2 is Q_2 => Q_1, the confidence of $Q_1\&Q_2$ is 0.5.
- Of the 2,000 examinees, 1,000 examinees provide incorrect responses to Q_1, and 1,000 examinees provide incorrect responses to Q_2. Among them, 500 examinees provide incorrect responses to both Q_1 and Q_2. If the rule for the items Q_1 and Q_2 is Q_1 => Q_2, the confidence of $Q_1\&Q_2$ is 0.5. If the rule for the items Q_1 and Q_2 is Q_2 => Q_1, the confidence of $Q_1\&Q_2$ is also 0.5.

The above two cases show that the confidences of a test item pair are not the same in different association rules. Only when the numbers of two test items in a pair that are answered incorrectly are the same, the confidence of this test item pair is unique. So the confidence is also not suitable for describing the correlation between test items.

The Pearson correlation coefficient is a statistical standard that reflects the degree of correlation between variables. The value scale is [−1, 1], where 1 indicates that the variables are completely related, 0 indicates that the variables are totally unrelated, and −1 indicates that the variables are completely negatively correlated.

The Pearson correlation coefficient is commonly represented by the Greek letter ρ (rho), and the formula for ρ is as follows:

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}. \tag{3}$$

In Formula (3), X,Y is a test pair, and N denotes the total number of transactions.

To apply the Pearson correlation coefficient to the analysis of test item correlation, we use a modified version of Formula (4), as shown below (Xiong et al., 2004):

$$\rho_{AB} = \frac{\sup(A\&B) - \sup(A)\sup(B)}{\sqrt{\sup(A)(1-\sup(A))\sup(B)(1-\sup(B))}}. \tag{4}$$

Formula (4) has the following properties:

- $\rho_{AB} > 0$ indicates that the correlation between A and B is positive.
- $\rho_{AB} = 0$ indicates that A and B are totally independent.
- $\rho_{AB} < 0$ indicates that the association between A and B is negative. In this paper, we do not consider the occurrence of this situation.

The relation between ρ_{AB} obtained in the correct response data and that obtained in the incorrect response data is shown in Lemma 1.

Lemma 1 : *ρ_{AB} in the correct response records is the same as ρ_{AB} in the incorrect response records.*

Proof : Set the group of examinees who provided an incorrect response to A as A' and the group of examinees who provided an incorrect response to B as B'. All the examinee response records on A and B comprise space T'. Set the group of examinees who provided incorrect responses to $A\&B$ as $A' \cap B'$, as shown in Figure 3. Therefore, $\sup(A) = P(A')$, $\sup(B) = P(B')$, and $\sup(A\&B) = P(A' \cap B')$. According to Formula (4), we can prove that ρ_{AB} in the set of incorrect responses is equal to the ρ_{AB} in the set of correct responses.

Suppose the data belong to the incorrect set of responses. The value of the numerator is $P(A' \cap B') - P(A')P(B')$. The numerator in the corresponding correct response part is $P\left(\overline{A'} \cap \overline{B'}\right) - P\left(\overline{A'}\right)P\left(\overline{B'}\right)$. Simplifying the numerator, we obtain

$$P\left(\overline{A'} \cap \overline{B'}\right) - P\left(\overline{A'}\right)P\left(\overline{B'}\right) = 1 - P\left(A' \cup B'\right) - \left(1 - P\left(A'\right)\right)\left(1 - P\left(B'\right)\right) = P\left(A' \cap B'\right) - P\left(A'\right)P\left(B'\right).$$

Additionally, the denominator values in the incorrect responses and the correct responses are, respectively,

$$P\left(A'\right)\left(1 - P\left(A'\right)\right)P\left(B'\right)\left(1 - P\left(B'\right)\right),$$

$$P\left(\overline{A'}\right)\left(1 - P\left(\overline{A'}\right)\right)P\left(\overline{B'}\right)P\left(1 - \overline{B'}\right).$$

Obviously, the two denominators are equal. Therefore, the ρ_{AB} of correct responses is equal to the ρ_{AB} of incorrect response.

Based on Lemma 1, for simplicity, we use only the incorrect response data in this study. In the rest of this paper, all the data refer to the incorrect response data.

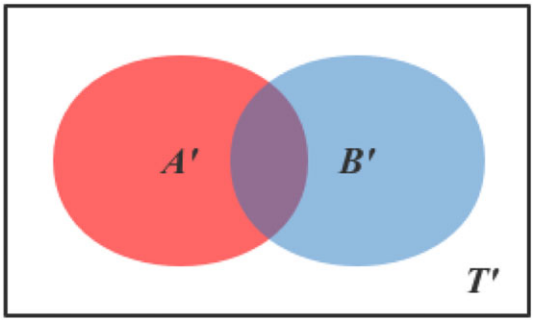

FIGURE 3 Venn diagram of A' and B'

Continuing to simplify Formula (4), we obtain Formula (5):

$$\rho_{AB^*} = \frac{TT_{AB} - T_A T_B}{\sqrt{T_A(T-T_A)T_B(T-T_B)}}, \tag{5}$$

where T_A refers to the number of examinees who provided an incorrect response to A, T_B refers to the number of examinees who provided an incorrect response to B, and T_{AB} refers to the number of examinees who provided an incorrect response to $A\&B$. According to Formula (5), we draw Figure 4 below.

In Figure 4a, we set T = 1,000 and $T_A = T_B$ = 500. T_{AB} and ρ_{AB^*} exhibit a significant linear increasing relationship. In Figure 4b, we set T = 1,000 and T_A = 500. T_{AB} is a random number that is no larger than the minimum value between T_A and T_B. As T_B becomes closer to T_A, ρ_{AB^*} approaches 1 or −1.

In the remainder of the paper, we use ρ to replace ρ*, which is calculated as Formula (5), for the purpose of simplicity.

3.3 | Algorithm for the TIA model

In this section, we propose an algorithm to generate all the test item pairs and calculate the ρ values based on examinee response data. The algorithm is divided into two steps.

Step 1:. Find all the test item pairs in the examinee response data.

- . Retrieve the total number of incorrect responses.
- . Count all the test items that appear in the examinee incorrect responses.
- . Count all the test item pairs of incorrect responses.

Step 2:. Calculate ρ for each test item pair using Formula (5). The details of the algorithm are shown below.

Algorithm for TIA model

Input: *D*, the indexes of examinee incorrect response data
Output: ρ, the correlation coefficient for each itempair
Variables: L, the size of *D*

Method:
Generate_TIA(*D*)

```
1. allItems = {}
2. allItemPairs = {}
3. for each response in D do
4.  for each item and itempair in response do
5.   add item, count(item) to allItems{item:count(item)}
6.   add itempair, count(itempair) to allItemPairs{itempair:count(itempair)}
7.  end for
8. end for
9. return allItems, allItemPairs
```

Calculate_ρ (allItems, allItemPairs)

```
10. for k, v in allItemPairs do
```

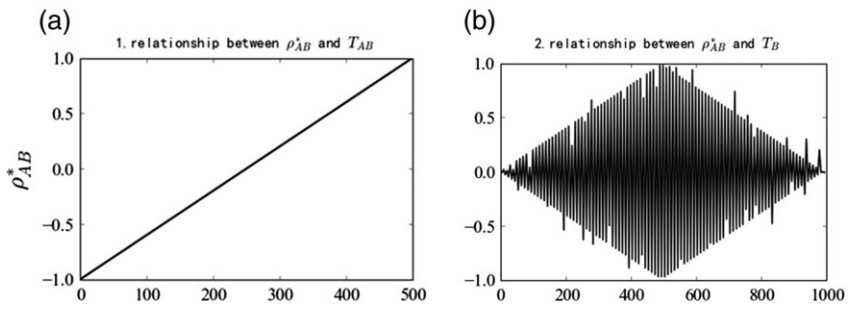

FIGURE 4 Relationships between ρ_{AB^*} and T_B, T_{AB}

```
11. if k[0] in allItems.key and k[1] in allItems.key then
12.     numerator = L*v - allItems[k[0]]*allItems[k[1]]
13.     denominator = sqrt(allItems[k[0]]*(L - allItems[k[0]])*
14. allItems[k[1]]*(L - allItems[k[1]]))
15.ifdenominator>0then
16. ρ=numerator/denominator
17. end for
18. return ρ
```

4 | VERIFICATION OF SIMILARITY

We use the hypothesis test and the empirical analysis to verify and analyse the relationship between the test item correlation and similarity in this section. Among them, the empirical analysis is a method of analysing the specific relationship between the test item correlation and similarity based on the similarity results of many sections, which are judged by experts. Based on the similar verification results, we can see that the TIA model has advanced performance.

4.1 | Data preprocessing

First, the student response records are obtained from the data structure course at a local university. Ten sections of response records are randomly selected as the experimental data from 46 sections. Then, all test item pair correlations are calculated by the TIA algorithm.

Second, experts are invited to make professional judgments on the content similarity of test item pairs in 10 sections. If the contents of two test items are very different, the similarity of this pair will be graded "0." If the contents of two test items are very similar, the similarity of this pair will be graded "2." Otherwise, the similarity of this pair will be graded "1."

Based on the data preprocessing, the test item correlation and the test item similarity are all obtained, which is prepared for the follow-up verification.

4.2 | Hypothesis test

This part uses the hypothesis test to illustrate that there is a relationship between the test item correlation and the test item similarity, which is achieved by the Minitab software (http://www.minitab.com/zh-cn/).

First, according to the above preprocessed data, we construct a data list that includes the values of the test item similarity and correlation.

Second, we formulate the null and the alternative hypotheses. In the Minitab, the null hypothesis is usually the opposite of the expected result and possibly rejected, whereas the alternative hypothesis is what we want to prove that it is true (Expósito-Ruiz, Pérez-Vicente, & Rivas-Ruiz, 2010). The null hypothesis in this paper is that the test item similarity is not related to the test item correlation.

Third, we select a suitable statistics to determine whether or not to reject the null hypothesis. The chi-square test for association in Minitab can be used to determine whether one variable is associated with another variable. Therefore, we use the chi-square test for association to judge whether the test item similarity is related with the test item correlation.

Finally, we create the hypothesis test. Among the experiments, the significance level α of the chi-square test for association is set to 0.05 (Expósito-Ruiz et al., 2010). The sample data are the data list constructed in the first step. The results are shown in Figure 5 below.

In Minitab, the choice of hypothesis is always determined by the *P* value. The *P* value is a probability that is defined as the minimum level of significance with which the null hypothesis is rejected. From the comparison with the significance level, if the *P* value is lower than α, we will have enough evidence to reject the null hypothesis in favour of the alternative hypothesis; otherwise, the null hypothesis will be selected (Expósito-Ruiz et al., 2010).

Figure 5 shows that the *P* value is less than 0.001, so we select an alternative hypothesis that the test item similarity is related with the test item correlation. In addition, the *P* value less than 0.001 has a significant statistical difference; that is, there is a significant difference in the test item correlation between test item pair with different similarities, which can be seen from the percentage profiles chart. In the percentage profiles chart, the left vertical coordinate represents three values of test item similarity, and the horizontal coordinate represents the proportion of different correlation in each similarity. In each similarity, the values of test item correlation are only selected the first six of data list constructed in section 4.2, that is, the correlation numbered 1 to 6. Furthermore, the proportion is calculated as that the test item correlation is divided by the sum of the first six correlation. As for the last chart in Figure 5, the left vertical coordinate represents three values of test item similarity, and the horizontal coordinate represents the proportion of disparity between expected and actual values. The positive values mean that the correlation occurs more frequently than expected, and the negative is the opposite.

To sum up, the test item similarity is really related with the test item correlation.

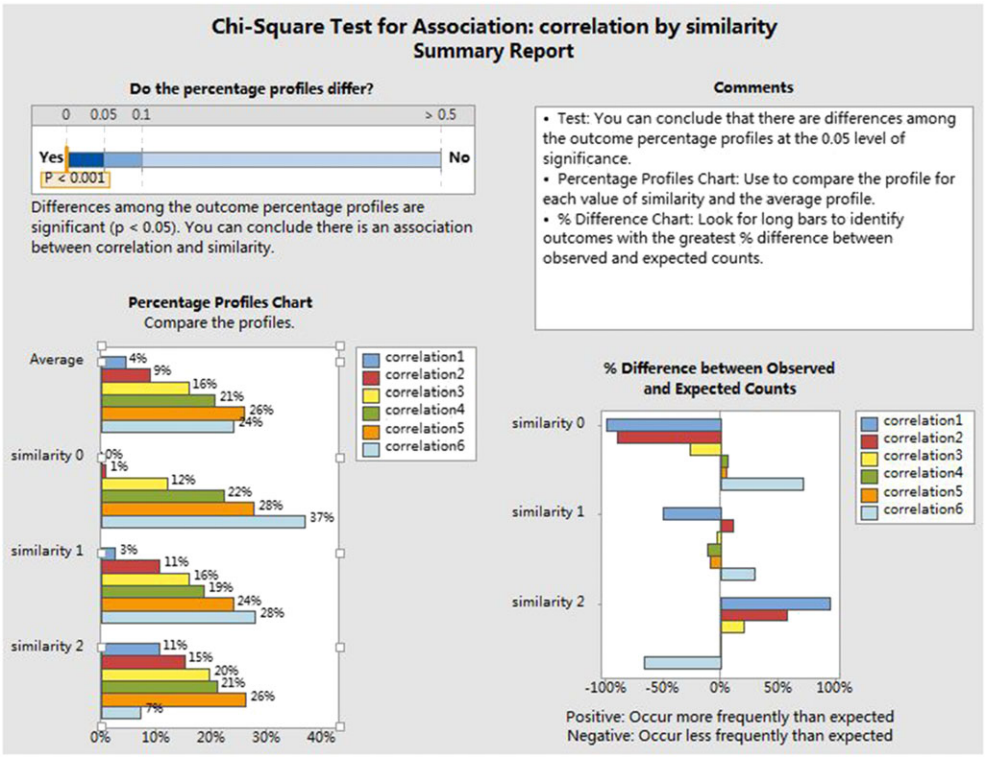

FIGURE 5 Summary report about the chi-square test for association

4.3 | Empirical analysis

According to the result of hypothesis test, the test item similarity is really related to the test item correlation. Furthermore, we use the empirical analysis to obtain the specific relationship between the test item similarity and correlation.

First, all test item pairs obtained from above preprocessed data in each section are arranged in ascending order of the test item pair correlation and classified into a number of correlation intervals.

Then, the average similarity of all test item pairs in each correlation interval is calculated. The results of 10 sections are shown in Figure 6, which is made by Highcharts (http://www.hcharts.cn/). In Figure 6, the left vertical coordinate represents the average similarity of test item pairs in each correlation interval. The right vertical coordinates represent the number of test item pairs in each correlation interval. The horizontal coordinates represent all correlation intervals.

Figure 6 shows that the similarity of test item pairs increases as the increase of the test item correlation in most sections. In particular, the last diagram, which represents the average similarity of 10 sections, clearly shows that the overall trend of similarity is upward. Therefore, the test item correlation can basically describe the similarity of the test item pair.

However, we also notice that high test item pair correlations will not necessarily reflect the high similarity of test item pairs. The first reason is the smaller amount of test item pairs in each correlation interval probably causes a phenomenon that happens by chance. The second reason is both of two test items have high discrimination, which results in high test item pair correlations, although these two test items have not similar contents.

In order to eliminate the above affects, students are allowed to judge the test item similarity to modify the test item correlation when they do exercises, so that the similarity of test items will be more accurate. From the above, the test item correlation can describe the test item similarity accurately.

For ARM, we also make an empirical analysis to obtain the relationship between the confidence and the test item similarity. The data processing is the same as section 4.1, and only the test item correlation is replaced by the confidence of test item pairs. In the process of calculating the confidence, all test item pairs are the same as that obtained in TIA algorithm, and we consider that the rule between two test items in a pair is Item 1 => Item 2, which is explained in section 3.2. The results of 10 sections are shown in Figure 7 below. In Figure 7, the horizontal coordinate represents all confidence intervals.

For the first six intervals (Intervals 1 to 6) of the TIA model, the slope of the test item similarity is 0.108. For the middle eight intervals (Intervals 2 to 9) of RAM, the slope of the test item similarity is 0.069. Based on the comparison of slope, for most intervals, we can find that the correlation curve (the last diagram in Figure 6) has better monotonicity than the confidence curve (Figure 7). Combining with the analysis results of section 3.2, we think that the TIA model has more advanced performance than ARM in the case of the similarity analysis between test items.

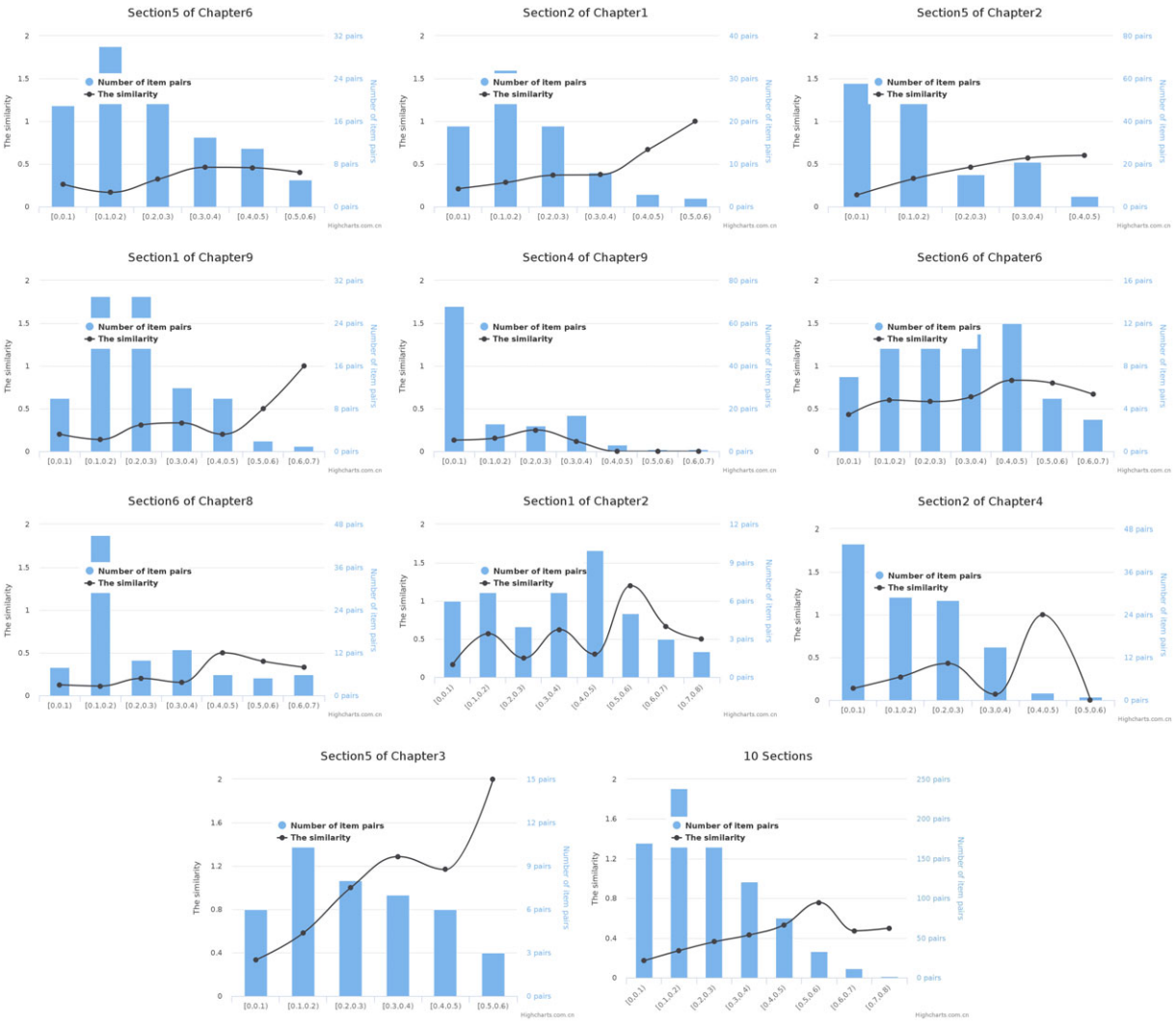

FIGURE 6 Test item pair similarity and correlation of 10 sections

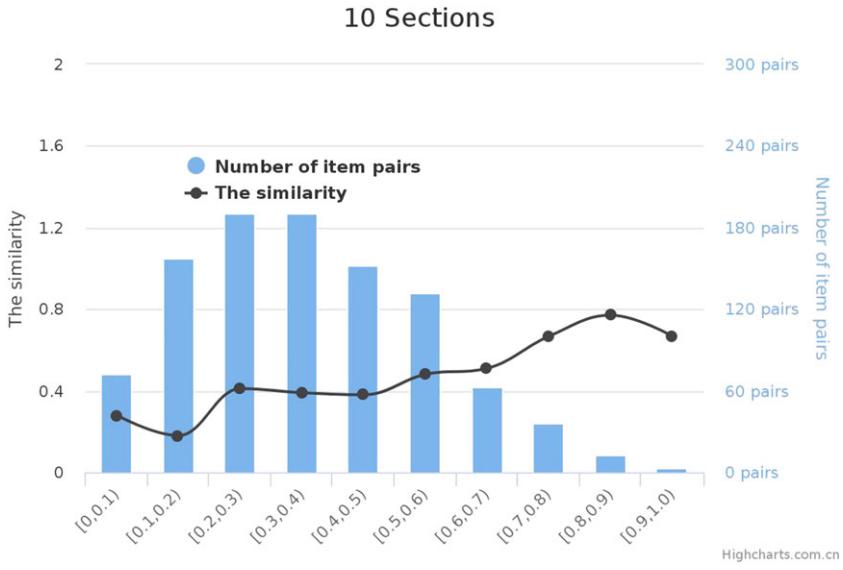

FIGURE 7 Test item pair similarity and confidence of 10 sections

5 | EVALUATION

In this section, we analyse how test item parameters such as discrimination, difficulty, and pseudo-guessing affect the correlation coefficient ρ. The analysis is based on IRT, and the analysis experiment uses WinGen to generate examinee response data. In section 5.1, we introduce IRT and explain the effects of the parameters on examinee response. Section 5.2 introduces WinGen and its operation. In section 5.3, we apply the TIA model to the generated examinee response data and find the relationships between ρ and the three parameters of the test items. Additionally, the impacts of the number of examinees and the number of test items are evaluated in this section.

5.1 | Introduction of IRT

IRT is a testing theory aimed at discovering the relationship between examinee abilities and various test item properties such as difficulty, discrimination, and pseudo-guessing. The three-parameter logistic model (3PLM) is one of the IRT models, and it can be described as shown below (Glas & Falćon, 2003):

$$P(\theta) = c + \frac{1-c}{1+e^{-Da(\theta-b)}}. \tag{6}$$

Formula (6) shows the probability of a test item being answered right. In Formula (6), θ indicates an examinee's ability; *a*, *b*, and *c* are test item parameters (discrimination, difficulty, and pseudo-guessing), respectively; and *D* is a constant equal to 1.7.

In order to discuss how three parameters affect the correlation, we use the 3PLM model to generate the stimulated examinee response data by the method of control variables. Then, the correlation coefficient ρ can be calculated by the TIA algorithm based on the response data. Finally, we find the relationship between the parameters and correlations.

5.2 | Experimental environment

We used WinGen (Windows Software that generates IRT parameters and item responses) to generate the response data (Han, 2007);

First, we set the number of examinees and the distribution of examinee abilities following a standard normal distribution. The number of test items in each record are set to a specific value and the response categories are defaulted to 2. Then, we select the 3PLM model to generate simulated response data that is closest to real response data.

Among the 3PLM, parameter *a* follows a log-normal distribution, parameter *b* follows a normal distribution, and parameter *c* follows a beta distribution (Kang & Cohen, 2007). Based on above settings, the stimulated response data can be generated for follow-up research.

5.3 | Experimental results

Because the test items in IRT are assumed to be independent, the simulated examinee data generated by WinGen based on IRT should cause a lower association between test items. However, certain experimental parameter values may prevent us from obtaining the expected smaller values of ρ. Hence, this section will focus on the influence of these parameters. We use ρ_mean and ρ_max to show the average and largest values of ρ, respectively.

Factor 1:. Discrimination

In this case, the number of examinees is set to 2,000; the number of test items is set to 100; the difficulty is following a standard normal distribution; and the mean of pseudo-guessing is 0.25. The reasons for parameter selection will be explained in the following experiments. The discrimination index should range from 0 to 3 (Probst, 2003); therefore, we ignore the discrimination index that its values are greater than 3. By changing the mean of the discrimination parameter *a* (*a*_mean), and keeping the standard deviation of the discrimination distribution at 0.5, we generate the stimulated data and apply it to the TIA model. The ρ_max and ρ_mean resulting from different circumstances are shown in Table 1 and Figure 8.

As shown in Figure 8, ρ_max and ρ_mean increase as *a*_mean increases. To explain this phenomenon, we need to know how discrimination affects *P*(θ). Figure 9 is an Item Characteristic Curve (ICC) which shows that students' probability of answering test items correctly in the different discriminations. As *a* approaches zero, the difficulty will have little impact on *P*(θ), which means *P*(θ) can almost be a constant. Therefore, the response data follow a binomial distribution, and the response is indeed randomly, which explains why the value of ρ is so low.

As the discrimination gradually increases, the performance of examinees on a test item will be affected by the test item's difficulty in a large extent. When *a* approaches infinity, *P*(θ) will be completely determined by the test item's difficulty, as seen in Figure 10.

In this condition, the test item's difficulty is 0. The probability of examinees whose abilities are below 0 that will answer this test item correctly is 0.25. However, others are likely to answer this test item correctly. Thus, for test items *A* and *B* of similar difficulty, examinees should have almost the same responses on the two test items, which means that T_A, T_B, and T_{AB} are closely and explains the phenomena in Figure 10.

TABLE 1 Results of ρ under various discrimination parameters

*a*_mean	*SD*	ρ_max	ρ_mean
0.1	0.5	0.08	0.02
0.3	0.5	0.11	0.02
0.5	0.5	0.18	0.04
0.8	0.5	0.23	0.06
1	0.5	0.33	0.09
1.5	0.5	0.44	0.13
2	0.5	0.42	0.15
2.5	0.5	0.49	0.17
3	0.5	0.53	0.2

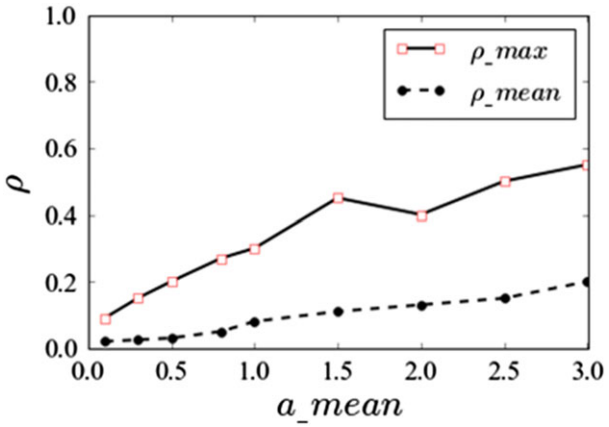

FIGURE 8 Relationship between discrimination and ρ

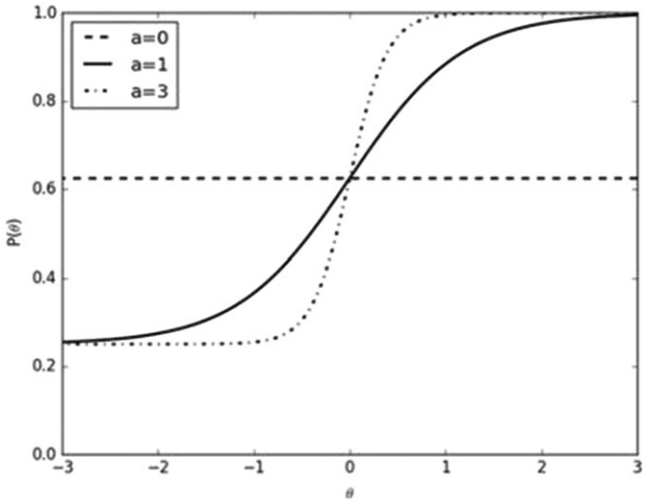

FIGURE 9 ICC of different *a*

To avoid the impact of discrimination on ρ, we should rectify the value of ρ, which is calculated according to Formula (4). From Figure 8, we can see that there is a general linear relationship between *a*_mean and ρ_mean. We obtain the Formula (7) after regression analysis, in which a_A refers to the discrimination of item *A* and a_B refers to the discrimination of item *B*.

$$\rho_r = \max(\rho - 0.032^*(a_A + a_B), 0) \tag{7}$$

After removing the test items whose discrimination is greater than 3 and using ρ_r to recalculate the associations among test items, the results are shown in Figure 11.

From Figure 11, we can see that the curves of ρ_max and ρ_mean are stable, and the value of ρ_r remains low. Therefore, we use ρ_r to show the correlation between two different test items and the mean of discrimination defaults to 1 in the following experiments.

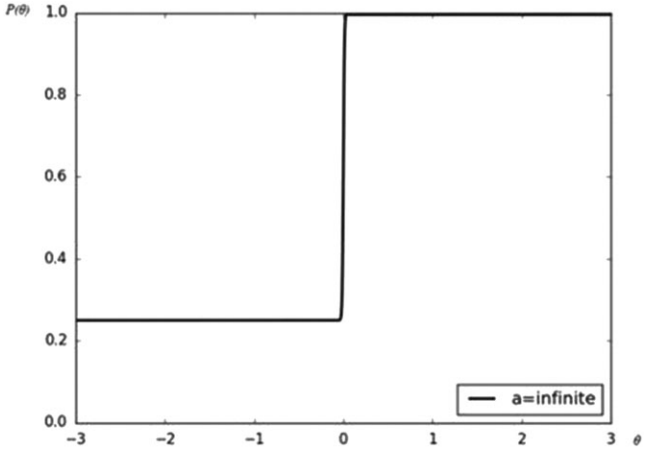

FIGURE 10 $P(\theta)$ as a approaches infinity

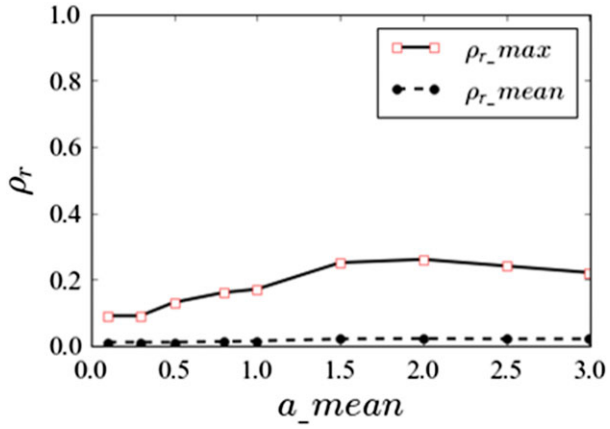

FIGURE 11 Relationship between discrimination and ρ_r

Factor 2:. Difficulty

In this case, the number of examinees is set to 2,000; the number of test items is set to 100; the mean of discrimination is 1; and the mean of pseudo-guessing is 0.25. By changing the mean of the difficulty parameter b (b_mean), and fixing the standard deviation of difficulty distribution to 1, we generate the stimulated data and apply it to the TIA model. The ρ_max and ρ_mean resulting from different circumstances are shown in Table 2 and Figure 12.

TABLE 2 Results of ρ_r under various difficulty parameters

*b*_mean	SD	ρ_r_max	ρ_r_mean
−6	1	0.44	0.009
−5	1	0.49	0.006
−4	1	0.19	0.005
−3	1	0.12	0.007
−2	1	0.16	0.016
−1	1	0.19	0.029
0	1	0.15	0.023
1	1	0.17	0.012
2	1	0.1	0.004
3	1	0.05	0.0006
4	1	0.04	0.0003
5	1	0.04	0.0003
6	1	0.03	0.0002

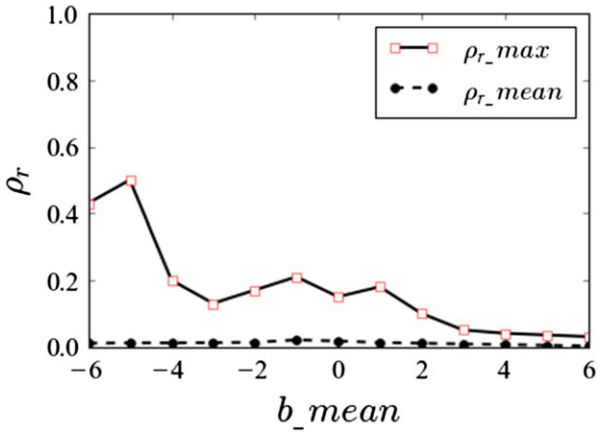

FIGURE 12 Relationship between difficulty and ρ_r

From Figure 12, the curve of ρ_r_mean almost parallels the horizontal axis, but ρ_r_max decreases as the mean of difficulty increases. To explore the reason, we analyse the occurrence of an item pair when *b*_mean is −6, as shown in Table 3. *Occu* refers to the occurrence number of a single item, and *Corr_oc* refers to the occurrence number of an item pair.

Table 3 shows that there are few wrong responses when an item's difficulty is too small. Most of the examinees provide correct answers in this situation. The examinees with higher abilities always answer the test items with the difficulties lower than their abilities correctly. Therefore, we omit the ρ_r where the occurrence number of test items is less than 5. The results are shown in Figure 13, in which ρ_r_max has an obviously decrease. There are sufficient wrong response data when the difficulty is in the middle; therefore, the mean of difficulty is set to 0. Additionally, for test items that the occurrence number is less than 5, the relative ρ_r is removed in the other experiments described in this paper.

Factor 3:. Pseudo-guessing

In this case, the number of examinees is set to 2,000; the number of test items is set to 100; the difficulty is set to a standard normal distribution; and the mean of discrimination is set to 1. By changing the number of test item options and the mean of pseudo-guessing parameter *c* (*c*_mean), we generate stimulated data and apply it to the TIA model. The ρ_r_max and ρ_r_mean resulting from different circumstances are shown in Table 4 and Figure 14.

TABLE 3 The occurrence number of an item pair when *b*_mean is −6

Item *A*	*Occu*	Item *B*	*Occu*	*Corr_oc*	ρ_r
30	1	91	4	1	0.44

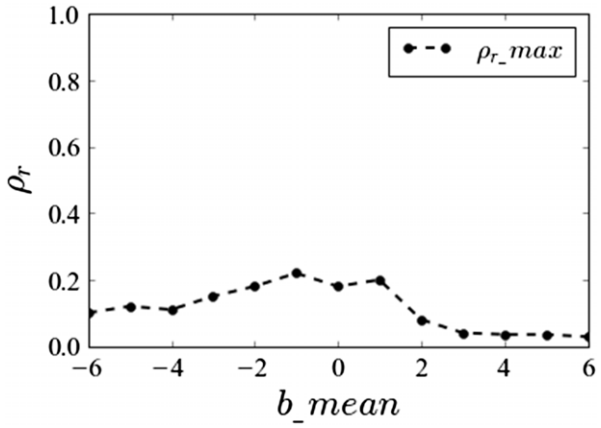

FIGURE 13 Result after removing items that occur less than 5 times

TABLE 4 Results of ρ_r under various pseudo-guessing parameters

Options	*c*_mean	ρ_r_max	ρ_r_mean
2	0.5	0.1	0.004
3	0.33	0.13	0.016
4	0.25	0.17	0.027
5	0.2	0.21	0.036

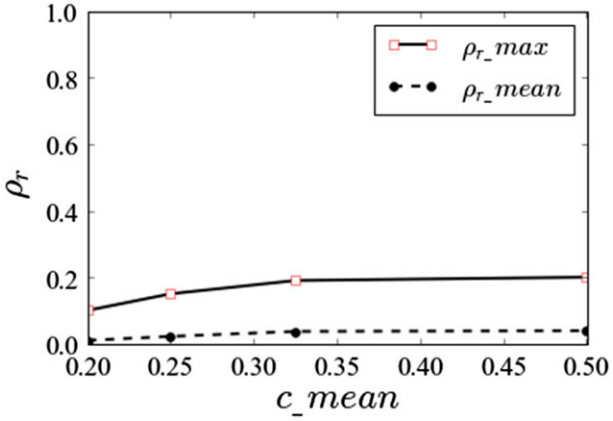

FIGURE 14 Relationship between pseudo-guessing and ρ_r

Pure pseudo-guessing relies on the number of test item options because all items are considered objectively in this study. Because most test items have four options, we use 0.25 to represent the value of pseudo-guessing in the other experiments of this paper. As shown in Figure 14, the pseudo-guessing has only a small effect on ρ_r.

Factor 4:. Number of examinees

In this case, the number of test items is set to 100, the difficulty is set to a standard normal distribution, the mean of discrimination is 1, and the mean of pseudo-guessing is 0.25. By changing the number of examinees, we generate stimulated data and apply it to the TIA model. The ρ_r_max and ρ_r_mean resulting from different circumstances are shown in Table 5 and Figure 15.

From Figure 15, ρ_r_max and ρ_r_mean decrease as the number of examinees increases when the number of examinees is below 500. However, when the number of examinees exceeds 500, the curves of ρ_r_max and ρ_r_mean tend to be stable.

TABLE 5 Results of ρ_r under various numbers of examinees

examinees	ρ_r_max	ρ_r_mean
10	0.56	0.25
20	0.75	0.16
30	0.65	0.13
50	0.41	0.08
100	0.44	0.07
200	0.31	0.05
500	0.21	0.03
1,000	0.20	0.03
1,500	0.14	0.02
2,000	0.15	0.02
3,000	0.17	0.02

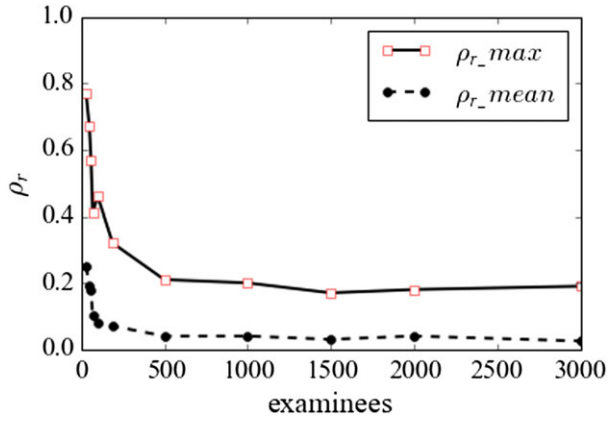

FIGURE 15 Relationship between the number of examinees and ρ_r

A small number of examinees can have a significant impact on ρ_r_max, but the number of examinees has less effect on ρ_r_mean because a small number of examinees cause a large random error. Therefore, IRT needs sufficient examinees. The recommended number of examinees should be no less than 1,000 in 3PLM (De Champlain, 2010). In the other experiments of this paper, the number of examinees is set to 2,000.

Factor 5:. Number of test items

In this case, the number of examinees is set to 2,000, and the attributes of three parameters in the 3PLM model are set the same as the settings of Factor 4. By changing the number of test items, we generate stimulated data and apply it to the TIA model. The ρ_r_max and ρ_r_mean resulting from different circumstances are shown in Table 6 and Figure 16.

As shown in Figure 16, the two curves are smooth because the number of test items has little impact on ρ_r. The number of test items is set to 100 in the other experiments of this paper.

5.4 | Discussion

Using the TIA model in the examinee response data helps find the correlation between two different test items. First, we find the linear relationship between ρ and discrimination. Next, we rectify the value of ρ based on Formula (7) and use ρ_r to represent the true association between two different test items. Then, we find that a small sample has a large impact on ρ_r under the experiments of Factors 2 and 4. The number of items and the pseudo-guessing have little effect on ρ_r. After these experiments, we suggest that the number of examinees should be over 1,000; the occurrence number of test items in a pair should be greater than 5; and the range of the test item discrimination should remain within [0, 3].

The TIA model can also be applied to real examinee response data. For example, a course has many topics, so not every student will master all these topics well. Students typically want to have more exercises in their weak topics. Therefore, we need to find the highly correlated test items based on a student's weak topics and present those items to students. The TIA model helps solve this problem.

At another example, an examination paper with high quality needs to cover all the knowledge points that examinees should have learned. Based on the correlation coefficient ρ_r, the test items with low correlation, difficulty distribution, and knowledge points have more checkpoints. In other words, we can use the correlation to generate high-quality examination papers. Therefore, the ρ_r can function as a metric to judge the quality of a real examination.

TABLE 6 Results of ρ_r under various numbers of test items

item_num	ρ_r_max	ρ_r_mean
10	0.14	0.024
20	0.13	0.025
30	0.12	0.024
50	0.10	0.016
100	0.15	0.024
150	0.15	0.021
200	0.20	0.021
300	0.18	0.020

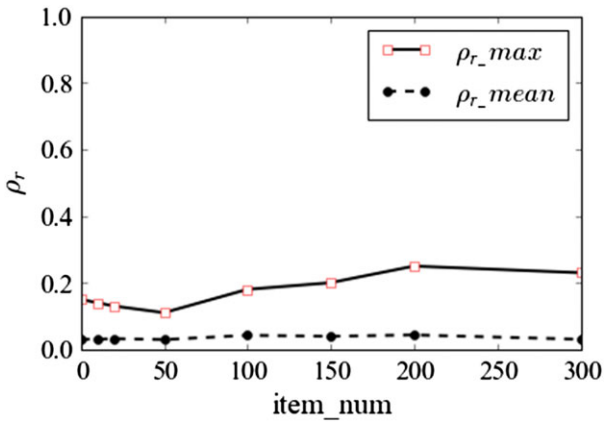

FIGURE 16 Relationship between the number of test items and ρ_r

6 | APPLICATION

Here, we focus on applying the TIA model to real examinee response data. The application will be divided into three parts. The first part mainly studies the feasibility of using the TIA model for diagnostic learning, and the second part generates an examination paper with comprehensive checkpoints based on the correlation. The last part judges the quality of examination papers by the correlation.

6.1 | Diagnostic learning

Diagnostic learning is a part of precision teaching; it provides efficient personalized instruction to students based on their learning. In the field of diagnostic learning, test items are always an important research objective. From section 4.3, we can see that the test item correlation can describe the test item similarity accurately. Based on the test item correlation, we can easily get the similarity between test items. According to high test item similarity, similar test items are pushed for students to exercise until students master this knowledge point. Thus, the TIA model can function as a method for applying diagnostic teaching.

The experimental data come from 1,727 examinee response records in an online learning platform of data structure course in a university. Ten test items are selected randomly from each section in every test. Based on examination response records, the correlation of all test item pairs are calculated with Formula (5). The correlation results are shown in Table 7 in a descending order.

According to Table 7, we find that the content of the test item with high correlation is actually quite relevant, indicating that we can use the TIA model to calculate the correlation between test items in the question bank and apply them to diagnostic learning. This is an efficient method for helping students master their weak topics in a course.

6.2 | Examination papers generation based on TIA

The examination papers should be of high quality and sufficient to assess student mastery of knowledge points. However, the test paper generated by the existing test-paper generation system has more similar test items and fewer checkpoints. Section 4.3 shows that the test items

TABLE 7 Top 20 of ρ

Item *A*	Item *B*	ρ
599	602	0.93
100	101	0.92
382	387	0.91
596	600	0.87
373	376	0.86
92	95	0.86
610	618	0.86
632	633	0.84
779	795	0.81
95	97	0.81
611	618	0.77
618	619	0.77
560	576	0.76
399	403	0.76
321	325	0.75
596	597	0.75
95	96	0.75
400	403	0.75
318	325	0.74
280	289	0.74

TABLE 8 Three examination papers

Paper	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	ρ
1	6	43	45	58	248	30	157	892	937	115	0
2	840	15	746	648	373	702	721	723	397	730	0
3	726	731	717	685	27	62	681	734	727	698	0

TABLE 9 Test item correlations

Item 1	Item 2	ρ
1	5	0.024
2	7	0.012
3	4	0.010
4	9	0.027
5	11	0.032
7	15	0.057
10	15	0.086
6	12	0.03
8	13	0.046
3	13	0.076
9	11	0.088
10	14	0.029

with low correlation have more checkpoints. Therefore, we use the method of test paper generation based on TIA to generate examination papers, which is composed of quantitative test items with low correlation, difficulty distribution, and comprehensive knowledge points.

Based on the correlation obtained in section 6.1, we generate three examination papers shown in Table 8.

The calculation results in Table 8 show that the average correlation coefficient of all selected items is 0. It means that the test papers generated by the method of test paper generation based on TIA have low correlation. Therefore, the test papers have a high quality.

6.3 | Judging the quality of an examination paper

This section used real examinee response data from a 10th-grade physics examination from Zhejiang Province in China. The data included 18,628 examinee response records. Sixteen objective items were selected in this experiment. By preprocessing these items, we obtain the response data that 0 represents the corresponding items that are answered incorrectly and 1 represents the opposite. All correlation coefficients ρ are obtained by the TIA algorithm based on above response data.

From the Table 9, we obtain that $\rho_{r\_}max = 0.11$ and $\rho_{r\_}mean = 0.03$, which indicate that all the objective items in this physics examination have low correlations. Based on the result of Table 9, those objective items in the examination are of fine quality.

7 | CONCLUSIONS

The TIA model aims to discover the correlations in test items automatically to replace traditional manual methods. The TIA model selects a proper formula for calculating the correlation between two different test items in item pairs. The correlation can be affected by parameters such as test item discrimination and the number of examinees. The hypothesis test and empirical analysis are to verify and analyse the relationship between the test item correlation and similarity. The results show that the test item correlation can describe the test item similarity accurately.

The test item similarity can be described by the correlation coefficient, so three applications are introduced. First, in diagnostic learning, quantitative similar test items are pushed for students to improve learning their weak parts. Then, the test items with low correlation have more checkpoints and less similar content. Therefore, the test-paper generation model based on TIA is used to select test items with low correlation into examination papers. Finally, we use the correlation as one of metrics to evaluate the quality of examination papers.

From Section 4, we can know that the test item correlation can be used to describe the test item similarity and the high accuracy. However, the accuracy is not absolutely right because of the chance and high discrimination illustrated in section 4.3. Therefore, students are allowed to judge the similarity of test items to modify the test item correlation when they do exercises.

The applications of the test item correlation are diagnostic learning and examination paper generation in Section 6. Based on the above theoretical research of TIA, an automatic test-paper generation system and a diagnostic learning platform are implemented. After multiple uses of these systems, we have found that the effect evaluation of the test item correlation is preferable. In the future work, we will search for more applications based on TIA model.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ORCID

Ligang Dong http://orcid.org/0000-0001-8429-9692

REFERENCES

Abdullah, Z., Herawan, T., Ahmad, N., Ghazali, R., & Deris, M. M. (2014). Mining indirect least association rule from students' examination datasets. *Lecture Notes in Electrical Engineering*, *285*(3), 159–166.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. *22*, No. 2, pp. 207–216). ACM.

Al-Razgan, M., Al-Khalifa, A. S., & Al-Khalifa, H. S. (2014). Educational Data Mining: A Systematic Review of the Published Literature 2006-2013. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)*.

Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. Sij Transactions on Computer Science Engineering & its Applications, *01*(01).

Barnes, T. (2005. The Q-matrix method: Mining student response data for knowledge. *In American Association for Artificial Intelligence 2005* Educational Data Mining Workshop, 1-8.

Cen, G., Dong, Y., Gao, W., Yu, L., See, S., & Wang, Q., et al. (2010). A implementation of an automatic examination paper generation system. *Mathematical & Computer Modelling*, *51*(11–12), 1339–1342.

Chen, C. L. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*(11), 314–347.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209.

Chu, H. C., Hwang, G. J., Tseng, J. C. R., And, G. H., & Hwang. (2006). A computerized approach to diagnosing student learning problems in health education. *Asian Journal of Health & Information Sciences*, *1*(1), 43–60.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117.

Expósito-Ruiz, M., Pérez-Vicente, S., & Rivas-Ruiz, F. (2010). Statistical inference: Hypothesis testing. *Allergologia et Immunopathologia*, *38*(5), 266–277.

Glas, C. A. W., & Falćon, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD International Conference on Management of Data* (Vol.29, pp.1–12). ACM.

Han, K. T. (2007). Wingen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459.

Hwang, G. J., Hsiao, J. L., & Tseng, J. C. R. (2008). A computer-assisted approach to diagnosing student learning problems in science courses. *International Journal of Distance Education Technologies*, *3*(4), 35–50.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358.

Liu, Y. C., & Chen, P. J. (2012). Discovering discriminative test items for achievement tests. *Expert Systems with Applications*, *39*(1), 1426–1434.

Liu, B., Ma, Y., & Wong, C. K. (2001). Classification using association rules: weaknesses and enhancements. *Massive Computing*, *2*, 591–605.

Nepal, S., & Bouguettaya, A. (2013). Big Data and Cloud. In: A. Haller, G. Huang, Z. Huang, H. Paik, & Q. Z. Sheng (Eds.), *Web Information Systems Engineering – WISE 2011 and 2012 Workshops. WISE 2011*, *WISE 2012*. *Lecture Notes in Computer Science* (Vol. 7652). Berlin, Heidelberg: Springer.

Panjaburee, P., Hwang, G. J., Triampo, W., & Shih, B. Y. (2010). A multi-expert approach for developing testing and diagnostic systems based on the concept-effect model. *Computers & Education*, *55*(2), 527–540.

Pechenizkiy, M., Calders, T., Vasilyeva, E., & Bra, P. D. (2008). Mining the student assessment data: Lessons drawn from a small scale case study. *Educational Data Mining 2008*, *the*, *International Conference on Educational Data Mining*, *Montreal*, *Québec*, *Canada*, *June 20–21*, *2008*. *Proceedings* (pp.187–191). DBLP.

Probst, T. M. (2003). Development and validation of the job security index and the job security satisfaction scale: A classical test theory and IRT approach. *Journal of Occupational and Organizational Psychology*, *76*(4), 451–467.

Romero, C., Zafra, A., Luna, J. M., & Ventura, S. (2013). Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Systems*, *30*(2), 162–172. https://doi.org/10.1111/j.1468-0394.2012.00627.x

Sharma, R., Shah, K., Shirahatti, Y., Patel, S., Sharma, R., & Shah, K., et al. (2010a). Data mining: Concepts and techniques. *Data Mining Concepts Models Methods & Algorithms Second Edition*, *5*(4), 1–18.

Sheugh, L., & Alizadeh, S. H. (2015). A note on Pearson correlation coefficient as a metric of similarity in recommender system. *Ai & Robotics* (pp.1-6). IEEE.

Sun, X. (2012). Finding dependent test items: An information theory based approach. *In Educational Data Mining 2012*.

Tseng, S. S., Sue, P. C., Su, J. M., Weng, J. F., & Tsai, W. N. (2007). A new approach for constructing the concept map. *Computers & Education*, *49*(3), 691–707.

Xiong, H., Shekhar, S., Tan, P. N., & Kumar, V. (2004). Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 334–343). ACM.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge & Data Engineering*, *12*(3), 372–390.

Zhang, Y. (2005). Quantitative analysis of the relationship of biology species using Pearson correlation coefficient. *Computer Engineering & Applications*, *41*(33), 79–81.

Zhou, L., & Yau, S. (2007). Efficient association rule mining among both frequent and infrequent items. *Computers & Mathematics with Applications*, *54*(6), 737–749.

Ligang Dong received his BS and master's degrees from Zhejiang University (Mixed Class), China, in 1995 and 1998, respectively. In September of 2003, he obtained a PhD degree from the Department of Electrical and Computer Engineering at the National University of Singapore, Singapore. He is currently a professor and the dean of the School of Information and Electronic Engineering at Zhejiang Gongshang University,

Hangzhou, China. Dr. Dong is a member of IEEE and IEEE–CS and a senior member of the Chinese Institute of Electronics. His research interests include various topics in smart networks and education.

Liujun Tang received her BS degree in Electrical and Information Engineering from Zhejiang Gongshang University, China, in 2016. She is currently pursuing a master's degree in Electrical and Information Engineering at Zhejiang Gongshang University, China. She focuses on educational data mining and deep learning.

Shihuan Liu received his BS degree in Telecommunications Engineering from Binzhou University, China, in 2011. He is currently pursuing a master's degree in Communication and Information Systems at Zhejiang Gongshang University, China. He focuses on the field of educational data mining.

Xian Jiang received his BS and master's degrees in Electrical and Information Engineering from Zhejiang Gongshang University, China, in 2012 and 2015, respectively. He is currently a teaching assistant of the School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, China. He focuses on big data and educational informatization.

How to cite this article: Dong L, Tang L, Liu S, Jiang X. Discovering the associations between test items in an examination. *Expert Systems*. 2019;36:e12331. https://doi.org/10.1111/exsy.12331