Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations

Chin-San Lin, Yu-Ling Lu*, Chi-Jui Lien

Department of Science Education, National Taipei University of Education, Taiwan

Received March 1, 2021; Revised March 30, 2021; Accepted May 16, 2021

Cite This Paper in the following Citation Styles

(a): [1] Chin-San Lin, Yu-Ling Lu, Chi-Jui Lien, "Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations," *Universal Journal of Educational Research*, Vol. 9, No. 6, pp. 1323-1332, 2021. DOI: 10.13189/ujer.2021.090622.

(b): Chin-San Lin, Yu-Ling Lu, Chi-Jui Lien (2021). Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations. *Universal Journal of Educational Research*, 9(6), 1323-1332. DOI: 10.13189/ujer.2021.090622.

Copyright©2021 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract The study applies machine learning (ML) algorithms to investigate the association between the length of a test item written in Chinese (through word count), item difficulty, and students' item perceptions (IPs) in science term examinations. For Research Question 1, items for grade 7 students aged 12–13 in a Taiwanese secondary school from 2014 to 2019 were analyzed. For Research Question 2, the study included 4,916 students from the said population. For RQ3, perceptions were gathered from 48 students of the same school in 2020. The study's results showed that first, the average word count of the 611 items was 88.81, with an average stem word count of 41.16, average options word count of 47.66, and stem-to-options word count ratio (S-O ratio) of 1.27. Second, given that the ML M5P categorization algorithm affirms the items' predictive power, the length of an item is a key factor in determining its difficulty. As a result of this algorithm, 3 categories of the length of science term examination items were classified (<57.5 words, 57.5–91.5 words, and >91.5 words), and 3 linear prediction models of item difficulty (LM1, LM2, and LM3) were generated. From these models, it was found that as the length of an item increases, so does its difficulty. In the prediction analysis of students' IPs, the J48 prediction result was better and convertible into understandable rules. IP was the root node of the decision rule, indicating the importance of this variable. Therefore, students would more likely answer an item correctly when 1) it was perceived to be easy or

normal, 2) the students had high or ordinary learning achievements in science, and 3) it contained less than 71 words. The study's results can be used as a reference for educators, examiners, and researchers in practical science term examination design. Moreover, it can guide the further research method and direction of applying machine learning in analyzing the difficulty of items in scientific assessments.

Keywords Difficulty, Machine Learning, Item Length, Item Perception, Item Analysis

1. Introduction

A good test not only evaluates students' learning outcomes but also provides teachers with an insight as to whether the defined teaching goals were reached. However, Taiwanese society at large, including school teachers, has recently found that multiple-choice test items in nationwide tests (written in Chinese) have become lengthier and more difficult. For example, at the end of the Comprehensive Assessment Program for Junior High School Students in Taiwan in May 2019, many critics immediately pointed out the trend [1]. Furthermore, various countries have recently included the development of students' literacy as a national curriculum goal. In line

with this, the increasing length of items has been attributed to such an assessment of students' literacy, creating a social-educational issue [2]. As some studies have found, the lengthier the items, the more difficult students find them [3]–[5]. Therefore, the researchers have further studied the items' length so that students may be able to answer them under a reasonable cognitive load. This improves the performance and reliability of the assessment, which are vital for its effectiveness.

In the past, studies on item difficulty and the length of a scientific assessment were mostly based on the characteristics of the item's text and option set, cognitive demand, and level of knowledge required [6], [3]–[5]. Although the application of machine learning (ML) algorithms for analysis is one of the automated and potential estimation methods, it has scarcely been done for the concerned topic. Based on existing literature, this study intends to use ML algorithms, including OneR, REPTree, J48, and LMT, to establish item difficulty models based on the length of a multiple-choice test item, item difficulty, and students' item perceptions (IPs). It seeks to reveal hidden and useful information that educators, test designers, and researchers can use to improve assessments based on students' needs.

Three research questions (RQs) are formulated to guide this study.

1. Regarding multiple-choice test items in secondary school science classes, what are the lengths of their elements (i.e., item word count, stem word count, and options word count)?
2. What are the item word count classification and item difficulty predictive model?
3. What is the predictive model of students' IPs in science term examinations?

2. Literature Review

2.1. Item Difficulty

The traditional item difficulty index (P) is used to indicate the degree of difficulty of each test item, usually expressed as the percentage of all examinees who correctly answered each one (percentage pass). The higher the P value, the higher the number of examinees who answered the item correctly, and the easier the item [7]–[8].

The P value is calculated as $P = R/N$, where R is the number of examinees who answered correctly, and N is the total number of examinees.

For example, in a class of 50 students where 40 got a certain mathematics test item correct, the P value for the said item is $40/50 = 0.80$, which means that 80% of the students got the item correct. In general, in academic assessments, test designers usually set the difficulty range and distribution of items based on the assessment's

purpose. For general-purpose tests, a P value of 0.40–0.80 is appropriate [9], i.e., $P < 0.40$ is a difficult item, and $P > 0.80$ is an easy one.

2.2. Length and Difficulty of Items in Science Assessments

Term examinations are mechanisms for evaluating subjects' learning statuses at a particular stage. Science, a core subject in schools, relies on paper-based assessments for understanding students' learning achievements [10]. In Taiwan, secondary schools generally administer science term examinations thrice per semester, for a total of six times in a school year.

The length and difficulty of the assessments in schools have changed in recent years, in response to the direction of educational reform and international assessment trends that emphasized the preparation of students' literacy. Lin et al. [11] conducted a study on the design of science literacy test items, stressing the future need to streamline the word count of examination papers to reduce the cognitive load on students.

Enright et al. [4] used 44 items from the life sciences subscale of the National Assessment of Educational Progress (NAEP) science assessment to study item attributes and item difficulty. By analyzing those attributes, which included the length of items, they reported that the lengthier the items are, the more difficult they are. Rosca [5] used a linear logistic test model to analyze 104 items from the Trends in International Mathematics and Science Study (TIMSS). They found that longer words, lengthy response options, and items in which the key is longer than the distractors constitute difficult items. Meanwhile, studies on the readability of test writing in Chinese are relatively rare [12], and there are no appropriate readability indicators that can readily be used. Thus, some studies use word count as a predictive variable for difficulty. For instance, Chang et al. [3] used word count to predict 103 translated items from the Programme for International Student Assessment (PISA), a science literacy assessment. They reported that the Chinese word count positively and significantly predicts difficulty. In addition, Hu and Twu [6] used the latent logistic latent trait model and categorical regression to find out the efficient predictors (cognitive components) for item difficulty. The results showed that more words in a stem and options increase the item's difficulty.

The studies mentioned are summarized in Table 1 to point out the gaps in their respective research areas. For example, the studies explored the effects of the word counts of a test stem and options, as well as item difficulty, albeit with mixed results. Regarding language, studies in these areas also focus more on tests written in English, seldom on non-English tests. Moreover, previous studies applied methods such as two-way itemization, expert opinion, actual analysis, regression analysis, and linear

logic models to analyze the difficulty and word count of the items, rarely applying machine learning (ML). Previous studies also did not include factors (e.g., students' item perceptions; IPs) that might explain the conditions in which students answer the item correctly, thereby affecting the P value of a particular item. Therefore, this study attempts to apply ML to analyze the associations between the item word count, item difficulty, and students' IPs.

2.3. Application of ML to Length and Difficulty Level of Items in Science Assessments

Machine learning (ML) is an emerging computerized technology that relies on algorithms constructed by "learning" from training data, with the potential to revolutionize science assessments. Hsu et al. [13] developed an ML method for automatically estimating the item difficulty of social studies tests that constructed a semantic space based on item elements of multiple-choice items, i.e., test stem, answer, and alternative options. With this, they showed that the new method outperformed the traditionally widespread yet resource-consuming pretesting method. A review study [13] included 49 machine-based studies on science assessments, finding that many studies (24 of 49) directly embedded ML in pedagogical activities. However, no study about the difficulty level concerning the lengths of an item's elements was reported, indicating a research gap in using ML techniques in science assessments.

In ML, the central tasks are to obtain a classification

function or model (and classifier) and then classify data into the model [14]–[15]. Commonly used classification algorithms include linear regression, decision trees, decision rules, and support vector machines. Decision trees can analyze a large number of multidimensional data to create simple and understandable classification rules. ML techniques could be used for building models representing the classification. Moreover, such could be used for making predictions [16]–[17].

In the past, studies on the sources of item difficulty have been mostly focused on IQ test items. Such knowledge has been gradually applied to several major achievement tests, such as NAEP, TIMSS, Taiwan Assessment of Student Achievement, and the Basic Competence Test for Junior High School Students [6].

Waikato Environment for Knowledge Analysis (WEKA) is a widely used open-source machine learning platform [18] that can establish and determine the applicable classification and regression models based on data distribution [16], and analyze a significant amount of multidimensional data to construct easy-to-interpret rules. It also supports multiple functions for data mining. Because the P value and word count are continuous data with nonlinear properties [6], the M5P technique which is capable of making prediction at each branch of the tree have significant advantages over traditional linear regression. Furthermore, M5P produces results that are easier to interpret [19]. Therefore, in the study, the built-in M5P of WEKA was used to analyze and establish the classification model for each answer to RQ2.

Table 1. Related Studies on Length and Difficulty of Items in Science Assessments

Researchers (Year)	Source of Test Items	Test Type	No. of Test Items	No. of Test Participants	Average Word Count of Test Items	Research Findings
Enright, Allen, & Kim (1993)	1985, 1986 NAEP in Life Science	Achievement test	44	Not available	Not available	There are item attributes, including length. The more words there are, the more difficult the question.
Rosca (2004)	1999 TIMSS in Science	Literacy test	104	Not available	70.72	Longer words, lengthy response options, and a key longer than the distractors make the item more difficult.
Chang, Hsieh, & Shyu (2009)	2006 PISA in Science	Achievement test	103	8,815	294.91	The number of Chinese words positively and significantly predicts difficulty.
Hu & Twu (2012)	2005–2008 Taiwanese science test item database	Achievement test	133	4,918	Not available	More words in a stem make the item more difficult.

Note 1. Researchers' Compilation (in Chronological Order)

WEKA also provides an Auto-WEKA function that can automatically test different machine learning techniques and compare their results for suggesting the best classifier [18]. It also offers many decision tree algorithms, such as REPTree, LMT, J48, ZeroR, OneR and so on, showing visualized representations of the classification [16], [20]. In terms of comparing the performances of decision trees, Huang et al. [21] used the results of ZeroR as a benchmark. The method is regarded as adequate when the performance of the said decision tree model is better than that of ZeroR. In addition, a model with understandable rules is used instead when the results generated by the algorithm are too complex for interpretation. As for the measurement of performance, common metrics include "accuracy" and "coverage." For example, the accuracy and coverage index of the decision tree rule R can be calculated as follows [22].

Num.covers=Number of cases covered by rule R

Num.correct=Number of cases corrected covered by rule R

Thus, this study will use Correctly Classified Instances, Kappa Statistic, F-Measure, and ROC Area features produced by WEKA to represent the performance of each decision tree prediction model. Higher values of each of the four metrics represent better model performance, which enables researchers to select the most appropriate algorithm for answering RQ3.

3. Research Method

This study aims to apply machine learning (ML) techniques to analyze the length and difficulty of a test item and students' item perceptions (IPs) for the development of an item word count classification, item difficulty model, and students' IP model.

3.1. Participants

This study was conducted in a secondary school in northern Taiwan that is located in a suburban area. RQ1 was formulated to analyze the midterm and term examinations of grade 7 students aged 12–13 from 2014 to 2019. As for RQ2, the item word count classification and difficulty prediction were determined through a study population of 4,916 students. In RQ3, students' IPs were analyzed through 48 students in 2020.

3.2. Research Tools

3.2.1. Research Material of RQs 1 and 2

The source of the items was the past term examinations of the said school in northern Taiwan. A total of 1,200 multiple-choice test items from 24 test papers (50 items each) were taken from grade 7 science term examinations from 2014 to 2019. The researcher and one research collaborator examined the items one by one and removed

those containing pictures, tables, and group test items. In the end, 611 text-only items were included as the study material in this analysis.

3.2.2. RQ 3's Research Tools

The research tools were 9 items (multiple-choice questions) selected from the secondary school's grade 7 science term examination papers from 2014 to 2019. Below each item, students were asked to answer four questions with checkboxes about their IP. These were designed with reference to the literature and included students' perceptions toward the item word count, item difficulty, their attention, and whether the item was previously encountered. An example of one of the items with its four IP questions was as follows.

() Which of the following is not a characteristic of all living things? (A) need nutrition (B) can reproduce (C) can grow (D) can move

1. I think the length of this item is:
☐Short ☐Medium ☐Long
2. I think this item is:
☐Easy ☐Normal ☐Difficult
3. When I answered this item, my attention to this item was:
☐High ☐Medium ☐Low
4. Have I read this item before?
☐Yes ☐No

The questions were refined by two science education experts and two science teachers through a focus group discussion to ensure reliability and effectiveness. Following the first draft, the completeness of the question design was verified by experts. Then, the tool was tested with a class of 28 students. The data were obtained and analyzed.

3.3. Data Analysis

3.3.1. RQ1: Word Count and Difficulty Level Calculation of Items

The researchers analyzed the data of the 611 text-only items based on the perspectives of the literature and compiled statistics on the item word count, stem word count, options word count, and stem-to-options word count ratio (S-O ratio). Test items containing pictures, tables, or group test items were removed. Punctuation marks were included in the word counts. The options word count also included its sequence letter/number (e.g. [A]).

The word count calculation uses the LEN formula function in Microsoft Excel to return the number of characters in the text string for the word count statistics. The word count analysis of sample item B1041120 is

explained below.

The Chinese item is as follows. “動物和植物的構造差異很大，但基本都是由什麼為單位組成的？ (A)細胞 (B)組織 (C)器官 (D)個體。”

This item is translated as such: “Although the structures of animals and plants are very different, what are they all composed of? (A) Cells (B) Tissues (C) Organs (D) Individuals”

The total Chinese word count is 41. Its stem word count is 28; the options, 13; and the S-O ratio, 2.154.

The item difficulty index (P), the percentage of all student examinees who correctly answered the item, is calculated by a computerized reading machine program that reads the computerized answer cards of the students' answers. It then evaluates and outputs the option analysis report to present the percentage of correct answers of individual items. For example, if the percentage of correct answers to the sample item is 80%, its item difficulty is 0.80. The word count data of the items were then registered, entered into the computer, and saved for analysis.

3.3.2. RQ2: Predictive Analysis of ML in Item Word Count Classification and Item Difficulty

The predictive variable in this part of the study is item difficulty, and the predictor variables are the item word count, stem word count, options word count, and S-O ratio, comprising four variables. Ten-fold cross-validation data analysis was performed with the built-in M5P algorithm in WEKA Ver. 3.8 as a data mining method to extract the classification rules for establishing the item word count classification and item difficulty.

3.3.3. RQ3: Predictive Analysis of ML to Survey Students' IPs

There are five attributes for this part of the research data, as shown in Table 2. The evaluation variable of the research data is the “correct or incorrect” answer of the students, and the predicted variables are the item word count and student's IP (learning achievement, length, difficulty, attention, and whether the item has been read before), comprising six variables.

After inputting the research data into a computer file, WEKA Ver. 3.8's Auto-WEKA, ZeroR, REPTree, LMT, and J48 algorithms were used to find a suitable and more predictive algorithm. Furthermore, these were used to analyze the data for comparing the advantages and disadvantages of various decision trees, determine the appropriate model for prediction based on the performance measure of the prediction results, and interpret the prediction rule. Decision tree rules with a higher coverage (>10%) can be chosen for the explanation [22].

Table 2. Attributes and Categories of ML Predictive Analysis

Attributes		Categories	
Learning achievement	Low	Ordinary	High
Length	Short	Medium	Long
Difficulty	Easy	Normal	Difficult
Attention	Low	Medium	High
Read before?	Yes	No	

4. Research Results and Discussion

4.1. Length of Items of Science Term Examinations in Secondary School

The results of the study for RQ1 are shown in Table 3. The average length of the 611 items from grade 7 science term examinations in a secondary school was 88.81 words. Moreover, the average word counts were 41.16 for the stem and 47.66 for the options, while the stem-to-options word count ratio (S-O ratio) was 1.27.

Table 3. Descriptive Statistics of Word Count in Grade 7 Science Term Examinations in Secondary School from 2014 to 2019 (611 Items)

Variable	Minimum	Maximum	Average	SD
Item word count	29	209	88.81	32.25
Stem word count	9	183	41.16	26.02
Options word count	8	147	47.66	26.78
S-O ratio	0.08	8	1.27	1.25

Comparing this result with international assessments, the literature shows that the average number of Chinese characters in the 218 items in the international PISA 2006 science assessment is 294.91 [3]. That is, the word count of the PISA science test items is 3.23 times that of the school. This shows that the reading load of the science test items in an actual school environment is significantly lower than that of the PISA test items. One explanation is that the term examination items of the school contained some basic scientific concepts. Such items have relatively lower word counts, reducing the average word count. However, if students were more accustomed to test items in school with shorter lengths or fewer words, their reading load with the PISA test items would be greater than that in a school term examination, so they may not perform well in PISA. This is consistent with the results of Chang et al. [3] wherein the reading load of PISA test items is heavy. Therefore, increasing the item word count in term examinations to raise the quantity of scientific text read by the students is an option to increase the proportion of literacy components in the items, thereby improving students' performance in international assessments.

4.2. Word Count Classification and Difficulty Prediction Model of Items in Science Term Examinations Obtained through Machine Learning (ML)

For the RQ2 study, 5 attributes of 611 items were analyzed: item word count, stem word count, options word count, S-O ratio, and item difficulty index (P). The performance of the M5P tree model was then tested by ten-fold cross-validation. Prediction models of word count classification and difficulty of the items were obtained as follows.

4.2.1. Word Count Classification of Items

The calculation results of the M5P (smoothed linear models) revealed that the word count of items in science term examinations can be divided into 3 categories: <57.5 words, 57.5–91.5 words, and >91.5 words.

Among them, 57.5 characters and 91.5 characters are 2 sub-nodes, and 3 linear prediction models are generated accordingly. LM1 (92/81.002%) in the formula below is an example. The value 92 is the number of items; 81.002% is the root relative squared error. This is shown in Figure 1.

4.2.2. Difficulty Prediction Model for Items

Three difficulty linear prediction models (LM1, LM2, and LM3) generated by ML on the word count (words) in science term examination items are shown below.

$$\text{LM 1 } P = -0.0002 * \text{words} + 0.7293$$

$$\text{LM 2 } P = -0.0001 * \text{words} + 0.6693$$

$$\text{LM 3 } P = -0.0001 * \text{words} + 0.5765$$

From these three difficulty prediction models, it is known that the higher the word count of the items, the more difficult they are. The P value of LM1, LM2, and LM3 is the word count (words) multiplied by the coefficient plus the intercept. Given that the coefficient of the word count is negative, as the word count increases,

the more difficult the items are. Likewise, an increase in the word count of items in secondary science term examinations will increase item difficulty.

From Figure 1 and the difficulty prediction model, it was found that (a) when the word count of items is <91.5 words as in the 2 models (LM1 and LM2), it has no effect on item difficulty, and the P value is 0.66–0.72; (b) when the number of words is >91.5 (LM3), and the P value is <0.57, the increase of the word count increases item difficulty.

According to the above difficulty prediction model, this study uses five attributes: word count, stem word count, options word count, S-O ratio, and P value. The ML decision tree M5P classification was used to generate the results of the item difficulty prediction model. M5P only uses the word count of items (words), an independent variable, for prediction, but it also shows that the word count of items is the key factor in determining their difficulty. From this, it was found that as the word count increases, the items will become more difficult. This is the difficulty model of the word count of items in science term examinations in secondary schools. Such a result is similar to the findings of Enright et al. [4], who suggested that the word count affects the difficulty of the items, and those of Chang et al. [3] and Hu and Twu [6], who proposed that the higher the word count of the items is, the more difficult the items are. This is also in line with the current situation reported in the news that lengthier items translate to greater difficulty.

The difficulty prediction model found in this study can provide a predictor of item difficulty in addition to the time-consuming and labor-intensive pretest. This finding is similar to the studies by Chang et al. [3] and Hu and Twu [6], who suggested that word count has significant predictive power on item difficulty. Therefore, this study serves as a reference for the design of scientific tests and assessments, such as local and international students' academic ability, competence, and literacy tests.

Figure 1. Word Count Classification Diagram of Difficulty Level of Items in Science Term Examinations Obtained by ML

4.3. Prediction Analysis of Students' Item Perceptions (IPs) in Science Assessments

For the RQ3 study, nine items were selected that fit the word count classification obtained from M5P and the scope of units that students studied. Among them, <57 words (item 1.3 with 45 words, item 2 with 54 words); 58–91 words (item 6 with 63 words; item 4.5. with 71 words), >92 words (item 7 with 103 words, item 8 with 136 words, item 9 with 174 words). At the bottom of each item, students were provided with questions with checkboxes to select their IPs. Then, ML algorithms, such as Auto-WEKA, ZeroR, REPTree, LMT, and J48, were applied for prediction analysis. The results are described below.

4.3.1. Performance of Various Predictive Models

Table 4. Performance of Various Predictive Models

Measure Model	Correctly Classified Instances (%)	Kappa Statistic	F-Measure	ROC Area
ZeroR	70.83	0	0.829	0.5
OneR	71.06	0.206	0.832	0.583
REPTree	75	0.234	0.846	0.738
J48	80.09	0.509	0.861	0.832
LMT	81.48	0.528	0.874	0.87

Table 4 shows a comparison of the performance of various decision tree models, where OneR is the best classifier obtained by Auto-WEKA. According to Table 4, in terms of the four measurement methods, Correctly Classified Instances, Kappa Statistic, F-Measure, and ROC Area, the values of OneR, REPTree, J48, and LMT are all higher than that of ZeroR. Among them, the prediction results of LMT and J48 are better, but the LMT model is a set of values, which is difficult to interpret. Meanwhile, the J48 model can be converted into understandable prediction rules.

4.3.2. Predictive Results of Applying ML to Explore Students' IPs in Science Assessments

The resulting tree structure of WEKA decision tree J48 on the item word count, science learning achievement, perceived difficulty, perceived attention, perceived word count, whether the item has been read before, and correct/incorrect answer is shown in Figure 2.

The performance of the J48 classification is checked by ten-fold cross-validation. The calculation result indicates the classifier's performance and model's acceptability (true positives [TP] rate: 0.846; false positives [FP] rate: 0.675; precision: 0.753; recall: 0.846; F-Measure: 0.797; Matthews correlation coefficient: 0.194; ROC Area: 0.683; PRC Area: 0.813).

Each terminal node in Figure 2 represents a decision rule. The root node is the difficulty level in students' IPs, indicating that this variable is an important attribute.

In terms of the conditions of the classification decision rules, the rules with a higher coverage rate (>10% coverage) were retained from the tree diagram in Figure 2 and tabulated as below. According to Table 5, the attributes of students who answered the item correctly are as follows. First, they find the item easy (R1, R2) or normal (R3); second, they have a high science learning achievement (R1) or ordinary (R2); and third, the word count of the item is <71 words (R2, R3).

From the J48 decision rule, it was found that the item word count affects the correct answer rate when students perceive that the item is of normal difficulty. In other words, items with fewer words (<71 words) are more likely to be answered right. Although students with ordinary academic achievement find the items easy, test items with fewer words (<71 words) are more likely to be answered right. It is observed that if students find the items easy or normal, and the length of the item is relatively short, they can easily get the item right. This is consistent with the prediction result of the difficulty model on the item word count found in this study: the higher the word count of the item is, the more difficult the item is. In addition, the higher number of incorrect answers by low-achieving students may remind test designers to consider the students' level and not to have too many lengthy or high word count items or to set an appropriate ratio of word count categories when designing an item.

Table 5. Decision Rule Table on Students' IPs in Science Term Examinations (>10% Coverage)

Decision Rule	Respective Rules	Coverage	Accuracy
R1	IF (Perceived difficulty = E, easy) & (Science learning achievement = H, high) THEN answered right (R)	13.4% (58/432)	96.6% (56/58)
R2	IF (Perceived difficulty = E, easy) & (Science learning achievement = O, ordinary) & (Word count <= 71) THEN answered right (R)	12.0% (52/432)	92.3% (48/52)
R3	IF (Perceived difficulty = N, normal) & (Word count <= 136) & (Science learning achievement = H, high) & (Word count <= 71) THEN answered right (R)	16.0% (69/432)	81.2% (56/69)

Note. Perceived difficulty (easy=E, normal=N, difficult=D); Word count (item word count); Science learning achievement (low=L, ordinary=O, high=H); Perceived attention (High=A, Medium=B, Low=C); Perceived length (perceived word count; short=S, medium=M, long=L); Read before (Yes=Y, No=N); R (answered right); F (answered wrong); R1 to R3 (respective rules in Table 5)

Figure 2. Tree Structure Diagram Obtained through J48 to Explore Students' IPs in Science Term Examinations

5. Conclusions and Recommendations

The study aimed to apply machine learning (ML) on the item word count, difficulty level, and students' item perceptions (IPs) in science term examinations in secondary schools. The results of the study could be summarized in three points. First, the average word count of 611 items in a grade 7 science term examination in a secondary school was 88.81. The average stem word count was 41.16, and options had 47.66 words on average. As for the stem-to-options word count ratio (S-O ratio), it was 1.27. Second, the result of the M5P algorithm classifies the length of science term examination items into 3 categories: <57.5 words, 57.5–91.5 words, and >91.5 words. Three linear prediction models of item difficulty (LM1, LM2, and LM3) were generated. From these models, it is found that as the length of items increases, the items become more difficult. Third and last, in the prediction analysis of students' IPs, the prediction result of J48 was better and could be converted into understandable rules. Rules with a >10% coverage rate show that the characteristics or conditions of students who answered the items correctly are the following. First, they find the item easy or normal (R3). Second, their science learning achievement is high (R1) or medium (R2). Third and last, they answer an item with a length of <71 words (R2, R3). According to the J48 decision-making rules, when an item is perceived to be of normal difficulty, its word count affects the correct answer rate. Thus, items with a shorter length (<71 words) are more likely to be answered correctly.

This study uses five attributes [i.e., item word count, stem word count, options word count, S-O ratio, and item difficulty index (P)] to perform ML M5P classification calculation and produce a resulting prediction model on item difficulty. M5P uses only the item word count for prediction, showing that such is a key factor in determining its difficulty. Moreover, according to the model results, as the length of the item in a science term examination increases, the more difficult the item is. Therefore, when teachers compile items, they can make the proportion of the number of items according to the word count. For example, for 50 items, 10 items can have <60 words; 30 items, 61–90 words; and 10 items, >91 words. This way, the difficulty distribution of the items can be roughly matched to the distribution of students' abilities in a class of mixed-ability groupings, and the number of longer items can be adjusted to provide a way of predicting the difficulty of the items. Furthermore, although teachers can prepare items of different difficulty levels, if the difficulty model of the items constructed by ML in this study can be used, it can also provide an assessment of the difficulty of the items compared to teachers' subjective perception or objective analysis. This also serves as a reference for future research in analyzing the difficulty of items in other subjects. Based on the results of this study using ML J48 to analyze students' IPs, the IP is the root node of the decision rules,

indicating that this variable is an important attribute. The 3 decision rules for getting the items correct (>10% coverage) are as follows. First, students with high academic achievement find the item easy. Second, students with ordinary academic achievement find the test easy, and the word count of the item is <71 words. Third and last, students with high academic achievement find the test question normal, and the word count of the item is <71 words. This result shows that students can easily answer the items correctly if they find the items easy or normal with fewer words. Although students with ordinary learning achievement find the items easy, they are more likely to answer longer items incorrectly. As such, students with low learning achievement have more circumstances for incorrect answers. Therefore, a science assessment will measure students' academic level more accurately if it is modeled on TOEFL's automatic question generation mechanism. That is, it could provide a varied combination of items with an appropriate amount of item word count, stem word count, and options word count, each according to whether a student answers the first item correctly.

Based on the results of this study, the researchers make the following recommendations.

First, the word count of items in schools' term examinations should be increased, or the proportion of longer items should be increased as appropriate. Because the average word count of science test items in international assessments is higher than that in school term examinations, the average item word count in school term examinations can be raised to >100 words, or the proportion of longer items can be adjusted. ML algorithms should also be set up to assign items that are appropriate for students' academic level for improving students' assessment performance and correct answer rate. This way, an assessment may serve its functions of assessing students' learning performance, promoting students' learning, monitoring academic ability, and providing feedback on teaching.

Second, in this study, 48 students' IPs were used to perform the decision rule analysis. For data mining, the number of participants in the future can be increased to further investigate the decision rules of students with different academic achievements under various conditions of the item word count, difficulty level, and attention given. Besides, the test items used to measure student's IP in this study were limited to a biology learning unit in the seventh grade. Future research should explore the possible impact of topic or grade level on students' IP, so that a deeper understanding on the tree structure can be achieved. In addition, this study used only the word count to represent the difficulty of the items, not the vocabulary level. Therefore, if the vocabulary level used in the items is also considered and the Chinese readability index is added to the ML algorithm, a higher explanatory power might be obtained.

This study is based on the concept of "applying ML

algorithms to develop an item word count classification and difficulty model that are helpful for item design in science subjects, and using ML algorithms to explore students' item perceptions." It aligns with the current trend of applying ML in science education. Therefore, this study significantly serves as a reference for science subject assessment design in practice and direction for research methods on item difficulty analysis.

Acknowledgments

This work was supported by Ministry of Science and Technology, Taiwan [MOST 108-2511-H-152-013-my3]. We also thank the students who participated in the study.

REFERENCES

- [1] C. Y. Lee. "The new syllabus has implemented for half a year, Baixi Chen, Director of Research Center for Psychological and Educational Testing: Don't rush your children to learn literacy" Online available from <https://www.parenting.com.tw/article/5078890>
- [2] Z. Weishan, Y. Yi. "2019 syllabus interview / Literacy tests lengthier and more difficult?" Online available from <https://udn.com/news/story/121690/4863655>
- [3] M. C. Chang, H. Y. Hsieh, C. Y. Shyu. A Cognitive Component Analysis for PISA Science Literacy, Curriculum & Instruction Quarterly, Vol. 13, No. 1, 1–19, 2009.
- [4] M. K. Enright, N. L. Allen, M. Kim. A complexity analysis of items from a survey of academic achievement in the life sciences. Educational Testing Service Rep. No. ETS-RR-93–18, 1993.
- [5] C. V. Rosca. What makes a science item difficult? A study of TIMSS-R items using regression and the linear logistic test model. Unpublished doctoral dissertation, Boston College, MA, 2004.
- [6] C. C. Hu, B. Y. Twu. The Cognitive Components of Item Difficulty of a Science Achievement Test, Journal of Education, Vol. 28, 155–191, 2012.
- [7] M. C. Hsieh, B. L. Wu. Applying of Fuzzy Statistics in Item Difficulty, Bulletin of Educational Psychology, Vol. 44, No. 2, 207–228, 2012.
- [8] S. M. Sim, R. I. Rasiah. Relationship between item difficulty and discrimination indices in true false-type multiple choice questions of a para-clinical multidisciplinary paper, Annals-Academy of Medicine Singapore, Vol. 35, No. 2, 67–71, 2006.
- [9] S. J. Osterlind. Constructing test items, Kluwer Academic Publisher, Boston, 1989.
- [10] C. S. Lin, Discussion from the paper-pencil test on problems and reflections of natural science literacy assessment in elementary and secondary schools, Taiwan Educational Review Monthly, Vol. 6, No. 3, 67–70, 2017.
- [11] P. L. Lin, C. C. Pan, S. Z. Su, P. H. Chen. Development of Assessments for Scientific Literacy Based on Curriculum Guidelines for 12-Year Basic-Education in Science Domains, Jiaoyu Kexue Yanjiu Qikan, Vol. 63, No. 4, 295–337, 2018.
- [12] Y. T. Sung, J. L. Chen, Y. S. Lee, J. H. Cha, H. C. Tsehng, W. C. Lin, T. H. Chang, K. E. Chang. Investigating Chinese text readability: linguistic features, modeling, and validation, Chinese Journal of Psychology, Vol. 55, No. 1, 75–106, 2013.
- [13] F. Y. Hsu, H. M. Lee, T. H. Chang, Y. T. Sung. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques, Information Processing & Management, Vol. 54, No. 6, 969–984, 2018.
- [14] X. Zhai, Y. Yin, J. Pellegrino, K. Haudek, L. Shi. Applying machine learning in science assessment: A systematic review. Studies in Science Education, Vol. 56, No. 1, 111–151, 2020.
- [15] M. Y. Yuan. WEKA Machine Learning and Big Data, 3rd ed., TopTeam Information, Taipei, 2017.
- [16] F. Hutter, L. Kotthoff, J. Vanschoren. Automated machine learning: methods, systems, challenges. Springer Nature, Switzerland, 2019.
- [17] S. C. Horng, F. M. Hsu. Applying decision tree to investigate the job satisfaction of elementary school teachers, International Journal of Advanced Information Technologies, Vol. 10, No. 1, 54–61, 2016.
- [18] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, K. Leyton-Brown. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA, The Journal of Machine Learning Research, Vol. 18, No. 1, 826–830, 2017.
- [19] Y. Wang, I. H. Witten. Induction of model trees for predicting continuous classes, working paper, 1996. Online available from <http://researchcommons.waikato.ac.nz/handle/10289/1183>.
- [20] G. Holmes, A. Donkin, I. H. Witten. Weka: A machine learning workbench, In Proceedings of ANZIIS '94 - Australian New Zealand Intelligent Information Systems Conference, 357–361, 1994.
- [21] J. L. Huang, Y. F. Yeh, Y. S. Hsu, Y. H. Tseng. Critical Factors of Becoming Secondary School Teachers: Mining the Process of Teacher Education by Decision Trees, Journal of Research in Education Sciences, Vol. 62, No. 2, 89–123, 2017.
- [22] J. J. Shuai, H. C. Chen, Y. B. Chen. Characteristics and Behavior Analysis of Online Fortune-Telling Users – A Data Mining Approach, Survey Research – Method and Application, Vol. 37, 97–154, 2017.