Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique

Suttida Rakkapao,[1,*] Singha Prasitpong,[2] and Kwan Arayathanitkul[3]

[1]*Department of Physics, Faculty of Science, Prince of Songkla University,
Hat Yai, Songkhla 90110, Thailand*

[2]*Faculty of Education, Thaksin University, Muang Songkhla, Songkhla 90000, Thailand*

[3]*Department of Physics, Faculty of Science, Mahidol University, Ratchathewi, Bangkok 10400, Thailand*

(Received 6 November 2015; published 25 October 2016)

This study investigated the multiple-choice test of understanding of vectors (TUV), by applying item response theory (IRT). The difficulty, discriminatory, and guessing parameters of the TUV items were fit with the three-parameter logistic model of IRT, using the PARSCALE program. The TUV ability is an ability parameter, here estimated assuming unidimensionality and local independence. Moreover, all distractors of the TUV were analyzed from item response curves (IRC) that represent simplified IRT. Data were gathered on 2392 science and engineering freshmen, from three universities in Thailand. The results revealed IRT analysis to be useful in assessing the test since its item parameters are independent of the ability parameters. The IRT framework reveals item-level information, and indicates appropriate ability ranges for the test. Moreover, the IRC analysis can be used to assess the effectiveness of the test's distractors. Both IRT and IRC approaches reveal test characteristics beyond those revealed by the classical analysis methods of tests. Test developers can apply these methods to diagnose and evaluate the features of items at various ability levels of test takers.

DOI: 10.1103/PhysRevPhysEducRes.12.020135

I. INTRODUCTION

The test of understanding of vectors (TUV), developed by Barniol and Zavala in 2014, is a well-known standard multiple-choice test of vectors for an introductory physics course at the university level. The TUV consists of 20 items with five choices for each test item, and covers ten main vector concepts without a physical context. A source of strength for the TUV is that the choices were constructed based on responses to open-ended questions, posed to over 2000 examinees. The TUV assesses more vector concepts than other previous standard tests of vectors and its reliability as an assessment tool has been demonstrated by five classical test assessment methods: the item difficulty index, the item discriminatory index, the point-biserial coefficient, the Kuder-Richardson reliability index, and the Ferguson's delta test [1]. However, the framework of classical test theory (CTT) for test assessment has some important limitations. For instance, the item parameters depend on the ability distribution of examinees and the ability parameters depend on the set of test items. To overcome these shortcomings, the item response theory (IRT) was introduced [2–5].

Therefore, the purpose of this study is to explore the 20-item TUV based on the framework of IRT. We will first present the key concepts of IRT, focusing on the three-parameter logistic (3PL) model used in the study (Sec. II). Since IRT displays only the functionality of correct answers to items, we will also investigate distractors of the TUV items using the item response curves (IRC) technique (Sec. III). Section IV is data collection of the TUV Thai language version from first-year university students ($N = 2392$). Section V (results and discussion) will be divided into three main parts. In part *A*, we will present some limitations of CTT, advantages of IRT, and the significance of using 3PL-IRT analyzed by our data. Parts *B* and *C* address the results and discussion of 3PL-IRT analysis and IRC analysis, respectively. Last, in Sec. VI, we will summarize what we did and found in this study.

II. ITEM RESPONSE THEORY (IRT)

The IRT framework rests on the assumption that the performance of an examinee on a test item can be predicted from the item's typical features and the examinee's latent traits—often called *abilities* or *person parameters*. The relationship between the examinees' performance on an item and their ability is described by an *item characteristic function* (or *item response function*), which quantitates how the probability of a correct response to a specific item increases with the level of an examinee's ability. The graph of this relationship is known as the *item characteristic curve* (ICC). The empirical ICCs in prior published

[*] Corresponding author.
suttida.rak@gmail.com

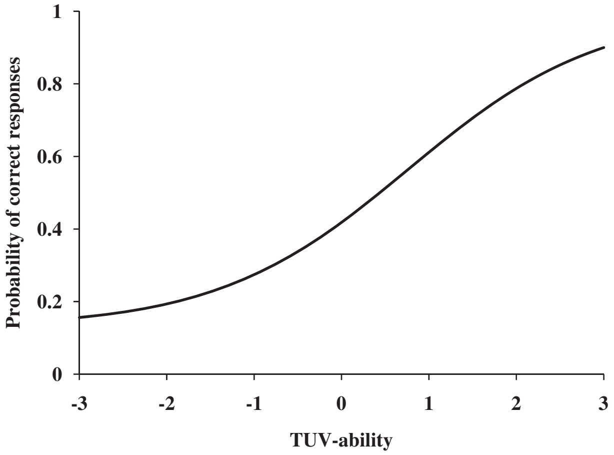

FIG. 1. Item characteristic curve (ICC) of item 1 in the TUV, modeled from data on Thai students ($N = 2392$).

research, of relevance to IRT, tend to be S shaped or sigmoidal. As the ability increases, the empirical ICC rises slowly at first, more sharply in the middle, and again slowly at very high levels of ability. In its early days, the normal ogive function was commonly used to model the ICC, while nowadays the logistic function is a popular alternative, as shown in Eq. (1). An example of the ICC from our data, established by fitting a logistic function, is shown in Fig. 1.

In item response function models, the ability parameter is usually standardized to zero mean and unit standard deviation (SD), and is commonly denoted by the Greek letter theta (θ). Theoretically, ability can range from $-\infty$ to ∞, but its values in practice are usually in the interval $[-3, 3]$. This interval would contain 99.7% of cases if the standardized variable is normally distributed, i.e., a z score. We assumed unidimensionality, meaning that a single dominant ability characteristic in the students influences their test performances. We then defined "*TUV ability*" as the single trait influencing a student's performance in the test of understanding of vectors, similar to the FCI ability reported in a study by Scott and Schumayer in 2015 [6]. Moreover, we made the local independence assumption that performance on one item is independent of that on another item, and only depends on the ability.

The TUV consists of multiple-choice questions with responses sometimes based on guessing, and the three-parameter logistic (3PL) model of IRT is appropriate for the investigation of the related item parameters (i.e., difficulty, discriminatory, and guessing parameters). In the item response function, the probability of answering item i correctly for an examinee with the ability θ is given by

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{[1 + \exp\{-Da_i(\theta - b_i)\}]}, \quad (1)$$

where b_i is the item difficulty parameter (θ at the inflection point of the ICC), and a_i is the slope of the ICC at that inflection point, called the item discriminatory parameter. The lower asymptote level c_i of the ICC, which corresponds to the probability of a correct answer at very low-ability levels, is referred to as the *pseudo-chance level* or sometimes as the guessing parameter. The constant D normally equals 1.7, and is used to make the logistic function as close as possible to the normal ogive function. The 3PL model can be reduced to the two-parameter logistic (2PL) model by setting $c = 0$. The 2PL model is most plausible for open-ended questions, in which responses are rarely guesses. Moreover, this can be further reduced to a one-parameter logistic (1PL) model by considering only the b parameter at which the probability of a correct response is 0.5, while holding a fixed [2,7–8]. The special cases with $D = 1.0$, $c = 0$, and $a = 1.0$, are known as Rasch models [9].

In this study, we applied the PARSCALE program, written by Muraki and Bock (1997), to fit the 3PL model to the dichotomous TUV data. PARSCALE uses the expected *a posteriori* (EAP) on estimating the ability, and marginal maximum likelihood (MML) to estimate the item parameters. The estimated abilities are scaled to mean 0 and unit S.D. The numeric search for optimum parameters uses Newton-Raphson iterations, and the program outputs both numerical results (parameters and diagnostics) and graphs [10].

III. ITEM RESPONSE CURVES (IRC)

Introduced by Morris and colleagues in 2006, IRC analysis is a simplification of IRT for evaluating multiple-choice questions and their options [11–12]. It relates the percentage of students (on the vertical axis) to each ability level (on the horizontal axis), separately for each choice in a given item of the test. Unlike IRT, which only considers whether the correct choice was made, IRC analysis displays the effectiveness of every choice. In other words, the information provided by wrong answers is also put to use. Moreover, it is easy to use and its results are easy to interpret. The IRC technique can help test developers improve their tools by identifying nonfunctioning distractors that can then be made more appropriate to the examinees' abilities.

This study approaches IRC analysis in a slightly different way from prior analyses [11–12], but without essential differences. These prior studies used the total score for the test as an estimate of the ability level across the students, and they are indeed strongly correlated. However, the strong correlation between total score for the test and ability is not necessarily equivalent for individual items of the test, nor for right or wrong responses to specific items—as also mentioned in Refs. [11–12]. Therefore, on applying the IRC technique, the ability of each examinee was estimated by the PARSCALE program, which uses optimality criteria instead of using the total score as a surrogate for ability.

IV. DATA COLLECTION

The 20-item TUV was translated into Thai and validated by a group of Thai physics professors. The professors were asked to perform the TUV in both English and Thai within a 2-month period. The test was revised based on suggestions from the professors, with regard to technical terms and translations of English to Thai. We applied the TUV to 2392 science and engineering first-year students at three universities. These students had learned vector concepts through lectures and integrated video demonstrations and group discussions, with approximately 300 students in each class. The participants took 25 min to complete the TUV, a month after the end of the classes. We used the Kuder-Richardson reliability index to measure the whole-test self-consistency of the TUV Thai version, and Ferguson's delta test for the discriminatory power of the whole test. The obtained indicator values, 0.78 for the KR-20 index and 0.97 for Ferguson's delta, are within the desired ranges [13]. The collected data were analyzed using the 3PL model in IRT and IRC plots.

V. RESULTS AND DISCUSSION

A. Significance of using 3PL-IRT

In the context of the data on Thai students, CTT presented some shortcomings and IRT analysis some advantages. We demonstrate these observations via the item difficulty index (P) and the point-biserial coefficient (r_{pbs}) of the CTT framework. The difficulty index of an item (P) measures the proportion of students who answer correctly. The point-biserial coefficient (r_{pbs}) measures the Pearson correlation of scores for a single item to the whole test scores, to reveal the discrimination information of that item. When we calculated P and r_{pbs} of our TUV data, we found that certain items showed outside the criterion ranges of $P = [0.3–0.9]$ and $r_{\mathrm{pbs}} \geq 0.2$ [1,13]. Only the results for items 1, 2, and 3 are shown in Table I, attached for discussion in this part. Item 2 ($P = 0.12$) and item 3 ($P = 0.24$) appear to have been somewhat difficult, and item 2 ($r_{\mathrm{pbs}} = 0.17$) was less useful in discriminating Thai students who know the cross product of vectors from those who do not.

To explore the invariance of item indices within the sample group when the ability of test-takers changes, we divided the Thai students into three groups by ability level. Overall, the Thai students' TUV ability (θ) varied from -1.7 to 3.1 with mean 0 and S.D. 1, as provided by the PARSCALE program for a single-trait model in IRT [10]. Using ranges below, within, and above the interval [mean ± 0.5S.D.] in TUV ability (θ), we partitioned the Thai students into low-ability ($N = 799$), medium-ability ($N = 833$), and high-ability ($N = 760$) groups. As shown in Table I, both P and r_{pbs} values of each item changed when ability levels of the students changed. The P and r_{pbs} values calculated from the low-ability group of examinees are likely to disagree with the norm values more than the other groups. This indicates that the item difficulty and discriminatory indices analyzed by the framework of CTT depend on a particular group of examinees. This is one important shortcoming of CTT. Moreover, in CTT, an examinee's ability being defined by the observed true score of the test depends on the item features. Simply, the item parameters depend on the ability distribution of examinees and the person parameters depend on the set of test items. Furthermore, CTT makes the assumption of equal errors for all ability parameters. There is no probability information available about how examinees of a specific ability might respond to a question. Generally, CTT focuses on test-level information, and depends on a linear model [2–5].

As mentioned earlier, IRT is an alternative that overcomes some disadvantages of CTT. The IRT is based on nonlinear models, makes strong assumptions, and focuses on item-level information. An ability parameter and its individual error are test independent, and are estimated from the patterns in the test responses. The item and ability parameters should be invariant if the model optimally fits the test data and the sample size is large enough [2–3,7–8]. These are theoretical advantages of IRT over CTT: however, some empirical studies have reported that the item and person parameters derived by the two measurement frameworks are quite comparable [4–5].

Using the IRT framework, we explored the invariance of item parameters relative to abilities of the test takers by applying the 3PL model to the data on Thai students taking the TUV test. Taking item 1 of the test as an example

TABLE I. Item difficulty index (P) and point-biserial coefficient (r_{pbs}) from CTT analysis for items 1, 2, and 3 in the TUV for overall and for three ability levels of Thai students.

	Item 1		Item 2		Item 3	
	P	r_{pbs}	P	r_{pbs}	P	r_{pbs}
Overall $\theta = [-1.7, 3.1]$ ($N = 2392$)	0.39	0.48	0.12[a]	0.17[a]	0.24[a]	0.36
Low ability $\theta = [-1.7, -0.6]$ ($N = 799$)	0.15[a]	0.17[a]	0.11[a]	0.18[a]	0.14[a]	0.25
Medium ability $\theta = [-0.5, 0.5]$ ($N = 833$)	0.34	0.13[a]	0.11[a]	0.25	0.20[a]	0.28
High ability $\theta = [0.6, 3.1]$ ($N = 760$)	0.70	0.24	0.16[a]	0.27	0.39	0.38

[a] Outside the criterion range.

(Fig. 1), the 3PL model fit to response data can be displayed as the item characteristic curve (ICC), representing the probability of correct response $[P(\theta)]$ across various TUV abilities (θ). The parameters for item 1, when fit to the entire data ($N = 2392$), are $a = 0.91$, $b = 0.76$, and $c = 0.13$, and these were computed by the PARSCALE program. When we separately fit the logistic model to subset data for the low-ability, the medium-ability, and the high-ability groups, the item parameters remained unchanged. This indicates invariance of the item parameters, relative to the subject population tested, which is desirable.

This can be simply explained by revising the logistic model of Eq. (1) into linear form. It can be rewritten as $\ln[\frac{1-P(\theta)}{P(\theta)-c}] = \alpha\theta + \beta$, where $\alpha = -Da$ and $\beta = Dab$. This linearization has slope α and intercept β, while $\ln[\frac{1-P(\theta)}{P(\theta)-c}]$ is the log odds ratio at given θ. Indeed, the same linear model should apply to any range for θ, giving the same values α and β, and therefore unchanged a and b. A single 3PL-IRT model for item 1 corresponds to a linear relationship, valid for any range of θ (low-ability, medium-ability, or high-ability groups) with fixed slope and intercept. However, this invariance property only holds when the model fits the data exactly in the population [2–3,7].

Several prior PER studies have applied the IRT framework to examine concept tests. For example, in 2010, Planinic and colleagues reported using the one-parameter logistic model in IRT to explore the function of the Force Concept Inventory (FCI) [14]. In the same year, the FCI was analyzed by Wang and Bao using the three-parameter logistic model in IRT, assuming the single-trait model [15]. However, in 2012, the study of Scott and colleagues showed that a five-trait model in IRT was suitable for analysis of the FCI [16]. Recently, the FCI has been analyzed using multitrait item response theory [6]. Aside from concept tests, IRT has also been applied to general online questions with large numbers of participants [17–18].

B. Analysis of 3PL-IRT

In applying IRT to the data gathered on Thai first-year university students ($N = 2392$), we used the PARSCALE program to fit the three-parameter logistic (3PL) models, one for each TUV item. We assumed a single ability, named the TUV ability, which represents the latent traits in each student that affect performance in the TUV. Each logistic model is determined by identifying its three parameters: discrimination a, difficulty b, and guessing c. These identified parameters are shown in Table II for the 20 TUV items, categorized by their concepts. Moreover, the item difficulty index (P) and the point-biserial coefficient (r_{pbs}) from the CTT framework are included as the last columns of Table II. The criterion ranges of the item

TABLE II. The model parameters identified in IRT analysis, namely, discrimination a, difficulty b, and guessing c, for the 20 items in TUV categorized by concepts, along with the item difficulty index (P) and the point-biserial coefficient (r_{pbs}) from CTT analysis.

		IRT			CTT	
		a	b	c	P	r_{pbs}
Vector concept	Item	[0,2]	$[-2, 2]$	[0,0.3]	[0.3,0.9]	≥0.2
1. Direction	5	0.84	0.18	0.40[a]	0.67	0.39
	17	0.93	1.54	0.14	0.26[a]	0.39
2. Magnitude	20	0.84	−0.26	0.10	0.61	0.49
3. Component	4	0.88	0.57	0.33[a]	0.57	0.41
	9	1.18	1.09	0.28	0.42	0.42
	14	0.69	0.56	0.26	0.53	0.40
4. Unit vector	2	1.12	2.82[a]	0.11	0.12[a]	0.17[a]
5. Vector representation	10	0.65	0.67	0.23	0.50	0.37
6. Addition	1	0.91	0.76	0.13	0.39	0.48
	7	0.96	0.25	0.24	0.57	0.47
	16	0.97	0.91	0.15	0.37	0.47
7. Subtraction	13	1.07	0.78	0.08	0.34	0.53
	19	1.12	0.86	0.05	0.30	0.54
8. Scalar multiplication	11	1.08	0.62	0.13	0.40	0.53
9. Dot product	3	1.04	1.70	0.14	0.24[a]	0.36
	6	0.75	−0.45	0.16	0.67	0.43
	8	0.59	0.83	0.00	0.26[a]	0.49
10. Cross product	12	1.02	1.74	0.08	0.17[a]	0.39
	15	0.60	0.74	0.00	0.29[a]	0.49
	18	1.17	1.16	0.12	0.28[a]	0.48

[a] Outside the criterion range.

2. The figure below shows vector $\vec{A}$. Choose the option that shows the unit vector in the direction of vector $\vec{A}$.

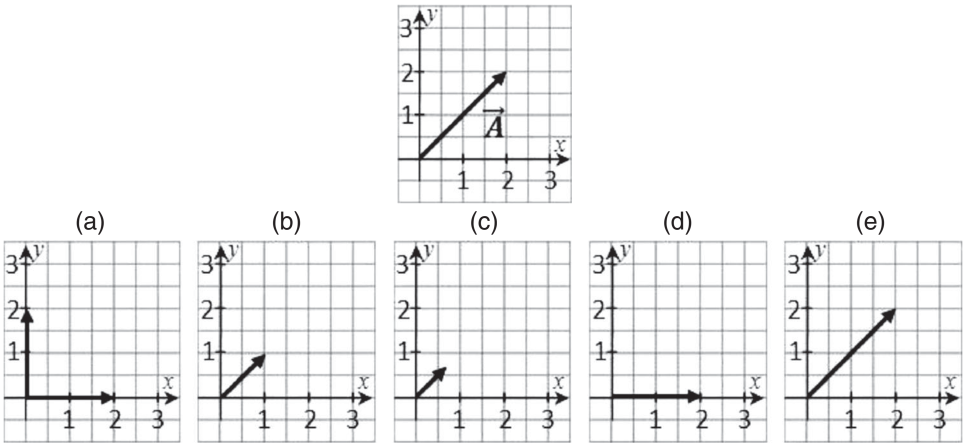

FIG. 2. Item 2 in the TUV, relating to the graphical representation of a unit vector in the direction of $\vec{A} = 2\hat{\imath} + 2\hat{\jmath}$.

parameters are shown by interval bounds in square brackets.

The item difficulty b is the ability θ at the inflection point of ICC. In the logistic model, at this point, the probability of correct answer is $(1 + c)/2$, midway between the asymptote levels, as seen by substituting $\theta = b$ in Eq. (1). When $c = 0$, as in the 1PL and 2PL models, the probability of a correct answer is 0.5 and this could be used to identify b. Parameter b is named "difficulty" because a harder test item requires higher ability b for probability 0.5 of a correct answer. The criterion range for b is chosen to be $[-2, 2]$ [2]. Clearly, an item with b close to -2 is very easy, while b close to 2 is very difficult for the sample population of examinees.

The results in Table II show that item 2, involving the unit vector concept, was the most difficult question for the group of Thai students, with the maximum $b = 2.82$. Item 2 in the TUV is displayed in Fig. 2. Only 12% of the students correctly answered item 2 (choice C) ($P = 0.12$). The most popular distractor for the students was choice B (61%). Interview results showed that these students understand that a unit vector of $\vec{A}$ has magnitude 1 and points to the same direction, but they thought the $\hat{\imath} + \hat{\jmath}$ vector in choice B has magnitude 1. It indicates that they did not have difficulties in the unit vector direction, but did not know the basic mathematics for calculating vector magnitude. The same misconception was reported in the study of Barniol and Zavala [19]. Since only a few students from the high-ability group understood both the direction and the mathematics to determine the magnitude of a vector, item 2 clearly presented the most difficulty to the normal-ability students.

Although difficulty b from the IRT analysis and the P value from the CTT analysis differ in their theoretical interpretations, they tend to be in good agreement. For example, item 12 for the cross product ($b = 1.74$) and item 3 for the dot product ($b = 1.70$) were somewhat difficult for the students, according to IRT. Their correct response rates were only 17% and 24%, respectively, indicating them as hard items according to the P values also. Further agreement was found in item 6, whose P value indicated an easy question, with about 67% correct answers, consistent with difficulty $b = -0.45$. However, the two measures of difficulty seem to disagree on items 8 and 15. Viewed through CTT, they appeared to be quite difficult items with correct answer rates below 30%, but their respective b values were 0.83 and 0.74, quite far from the upper bound of 2 for the difficulty parameter of IRT. As discussed earlier, this is one downside of the CTT approach, whose results depend on the examinees' ability. The Thai students had low-ability levels with two-thirds of them below ability 0.5, as shown in Table I. Then the P values are biased downwards, and this bias favors labeling items as hard or difficult.

The discriminatory parameter a for an item is related to the slope of ICC at point b: for the 3PL model, the slope of the ICC at $\theta = b$ is actually $Da(1 - c)/4$. Items with high a values, or steeper slopes, are more useful for distinguishing between examinees with closely similar abilities near b. The typical values of a are in the interval [0, 2] [2,8]. Several TUV items displayed very high a values, such as items 9, 2, 19, and 18, and these provide discrimination around the respective b values. In contrast, viewed through the CTT framework, the point-biserial coefficient of item 2 ($r_{\mathrm{pbs}} = 0.17$) indicates low discrimination power.

The guessing parameter c of an item represents the probability that an examinee with very low-ability level answers correctly. This may relate to the attractiveness of the answer choices and/or the guess behavior of the examinees. Its value is equal to the level of a lower asymptote for the ICC, and it ranges from 0 to 1. Typically, c should be less than 0.3 [8]. Table II shows that, overall, the very low-ability students had less than a

30% chance of choosing the correct option for most TUV items, with the exception of items 4 and 5. Item 5 involves selecting a vector with a given direction from among several in a graph, and very poorly performing students had a 40% chance to answer or guess correctly ($c = 0.40$). Possibly the correct choice was the most attractive to the very low-ability students. The most usual incorrect response (choice A) and the correct response (choice C) share the belief that vectors pointing to the same quadrant (northeast) have the same direction [1,20]. This may enable the low-ability students to score by chance in item 5 ($c = 0.40$).

In contrast, the probability that a student with very low-ability would correctly answer item 8 or item 15 was nil ($c = 0.00$). It is an indication that there are strongly held misconceptions represented in the distractors that appear as answer choices in these questions. Item 8 and item 15 involve the calculation of the dot product ($\vec{A} \cdot \vec{B}$), and the cross product ($\vec{A} \times \vec{B}$) of the vectors $\vec{A} = \hat{i} + 3\hat{j}$ and $\vec{B} = 5\hat{i}$, respectively. Our results show that choice C, $5\hat{i} + 3\hat{j}$, was the distractor most commonly chosen in both item 8 (41%) and item 15 (35%). The misconception is that the dot or cross product of two vectors with identical unit vector is the same unit vector and can be combined with other vectors [1]. Low-ability students, who have difficulties with calculations of the dot and cross product that involve unit vector notation, have no chance ($c = 0.00$) to score from both items. In fact, they have a better chance if they just guess and do not read the question, as this would give the correct choice with probability $1/m$ in a multiple-choice question with m options, as explained by the CTT framework. That probability is 0.2 in these TUV items with five choices in each.

Let us assess some sets of questions that measured the same vector concept, categorized previously in Table II. In items testing for the component of vector concept, items 4 and 14 gave quite similar model parameters. Item 4 had higher discrimination power than item 14, but larger guessing value. These items separate the examinees around the same ability level b, so they can be considered parallel questions for one concept at a fixed difficulty level. Items 1, 7, and 16, testing the addition of vectors in different contexts, display very similar a values, or sensitivities around ability $= b$, so we can select to use an item based on the b value (or the ability) we focus around.

To show how the probability of correct response for a specific item depends on the ability of an examinee, we build the item characteristic curve (ICC). The ICC of a well-designed question should have a sigmoidal S shape [2,7]. Then the probability of a correct response would consistently increase with ability, and a high slope at the inflection point would indicate sharp separation by ability around that point. As shown by the solid line in Fig. 3, item 19 of the TUV mostly agrees with these criteria in our data on Thai students. It has high discrimination power ($a = 1.12$) for separating examinees at medium-ability level near $b = 0.86$, and very low-ability students have a 5% chance to correctly answer the item ($c = 0.05$). This item asks students to choose the vector difference $\vec{A} - \vec{B}$ of two vectors ($\vec{A} = -3\hat{i}$; $\vec{B} = 5\hat{i}$) in the arrow representation. The most frequent error is adding instead of subtracting, and choosing the $\vec{A} + \vec{B}$ option (choice B). They just overlap two arrows, cancel a part of opposite direction and answer the remaining part. In some sense, many students seem to believe that the opposite arrow has already accounted for the subtraction, then they just add it with another arrow instead of subtracting it [1,21]. The low-ability students may hold such a misconception and have a lower scoring probability (5%) than they would have from a random guess, while the high-ability students easily master the concept of graphical subtraction of vectors in one dimension. Item 19 was then considered as the most appropriate question in the TUV for separating low- and high-ability students ($b = 0.86$).

Moreover, the steepness of ICC for items 9, 11, 13, and 18 (multiple choice items not shown here) was very similar to that of item 19, but with different ability thresholds or guessing parameters. Simply, a curve with greater b value is shifted further to the right (items 9 and 18), while a greater c value raises the bottom of the curve up (item 9). As shown in Fig. 3, the ICC of item 2 had the same slope as item 19, but it discriminates at a very high-ability level ($b = 2.82$). The curve of item 2 is very flat at the low-ability levels with $\theta < 0$ and rapidly rises at the high-ability levels. Also the ICCs of items 3 and 12 were similar to that of item 2, but moved further to the left. Roughly, the ICC of item 17 is flatter than that of item 19 owing to the smaller a parameter, which is similar to items 1, 7, 16, and 20. The slope of item 5's curve at its inflection point is close to that of item 17, but the curve is "lifted up" by its greater guessing parameter

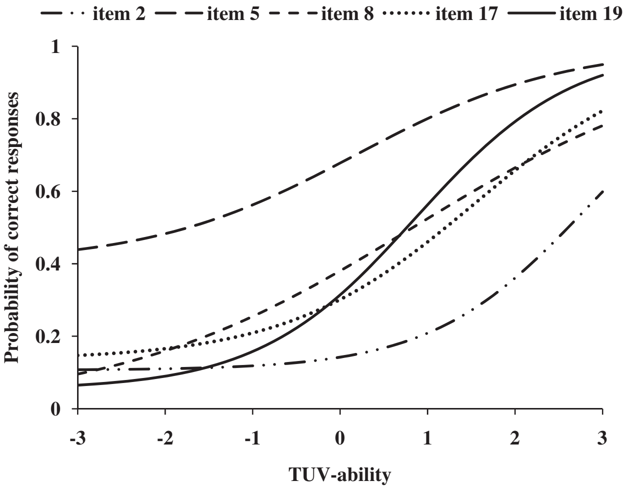

FIG. 3. The ICCs for items 2, 5, 8, 17, and 19 in the TUV, according to data on Thai students.

($c = 0.40$). This is similar to item 4, but with different b values. As shown in Fig. 3, the ICC of item 8 is flatter than those of other items of the TUV and has the lowest a value. TUV items with low discrimination power can still be in the criterion ranges, such as items 15, 10, 14, and 6, and present ICCs similar to that of item 8.

Overall, the results show that each TUV item has proper discrimination power at its specific difficulty (or ability level). Clearly, the steepness of the curve demonstrates the capability of the item to discriminate in the ability domain between examinees who understand the concept targeted by the item and those who do not. In general, a set of test items or questions should cover a range of ability domains in which the test takers are expected to differ. To determine how well the TUV does in testing adoption of the vector concept by the examinees at their various ability levels, the test information function was investigated in IRT framework. The information function for a test at θ, denoted $I(\theta)$, is defined as the sum of the item information functions at θ:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \frac{[P_i(\theta)']^2}{P_i(\theta)Q_i(\theta)}, \tag{2}$$

where $I_i(\theta)$ is the item information function of item i, $P_i(\theta)'$ is the derivative of $P_i(\theta)$ with respect to θ for item i, and $Q_i(\theta)$ is $1 - P_i(\theta)$ [2,7]. In Eq. (2), clearly information of one item is independent of other items in the test. This feature is not available in CTT. For example, the point-biserial coefficient for an item is influenced by all items in the test. A plot of $I(\theta)$ for the 20-item TUV across ability levels is shown in Fig. 4. The information curve peaks sharply to its maximum value 7.5 at $\theta = 1.1$. This indicates that the TUV test provides information about the vector concepts most effectively when the examinees have abilities roughly in the range from 0.1 to 2.1 (medium to high). For cases with abilities less than 0, the TUV test provides very little information that would distinguish their differences, while the results of the 20-item TUV test are highly sensitive to ability differences around ability 1.1. In general, the purpose of the test decides what type of information curve would be desired. For example, a flat curve over the whole range of abilities indicates that the component items are sensitive to variations at different ability levels, so the test information obtained by summing the item information becomes evenly distributed. This is desired if the test serves to assess a wide variety of abilities. In contrast, a sharply peaked curve sensitively reports differences of the test takers only around that peak, elsewhere it acts like a pass or fail threshold. This may be desirable when the purpose of the test is to provide eventual pass or fail labels. Test developers can benefit from these item and overall test information curves, to revise tests with such considerations of purpose in mind.

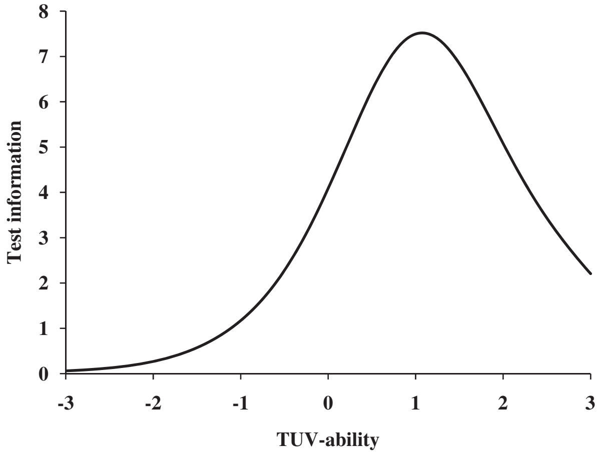

FIG. 4. Test information curve of the 20-item TUV across ability levels of the Thai students tested.

However, we have only analyzed data on test responses by Thai students. The item parameters of the TUV reported in Table II specifically apply to Thai first-year science and engineering students. When the TUV test is administered in another language to a different group of students, the item parameters will likely change from those obtained in the current study. Simply calibrating item parameters using IRT does not automatically make them universal. An equating or scaling procedure is needed to transform the item parameters of a test from one group of examinees to another, or for a given group of examinees the ability may be needed to transform from one test to another. Such equating usually assumes a linear relationship in the transformation of parameters.

There is some arbitrariness to the ability scale, and rescaling it gives an equally valid but different ability scale. We now examine such scaling along with transformations. Using the 3PL model in IRT, the probability of a correct response to item i by a person with ability θ is $P_i(\theta; a_i, b_i, c_i)$, as shown in Eq. (1). On linearly transforming the IRT ability, the probability of a correct response must not change, so $P_i(\theta; a_i, b_i, c_i) = P_i(\theta^*; a_i^*, b_i^*, c_i)$, with the transformed parameters indicated by stars. Notice that the guessing c_i is on the probability unit of measurement, so no transformation is necessary. The transformation equations are $b_i^* = Ab_i + B$, $\theta^* = A\theta + B$, and $a_i^* = a_i/A$, where A and B are the scaling constants of the linear transformation. These transformations do not change $a_i(\theta - b_i)$, which is the invariant property of the item response function [2,9,22]. Researchers can apply the transformation equations to transform the item parameters of the TUV reported in this article to another group of students. Mathematical methods introduced to estimate the A and B scaling constants include regression, the mean and sigma method, the robust mean and sigma method, and the characteristic curve method [2,9,22].

C. Analysis of IRC

To show how well choices of an item function are distributed, we will now assess the item response curves

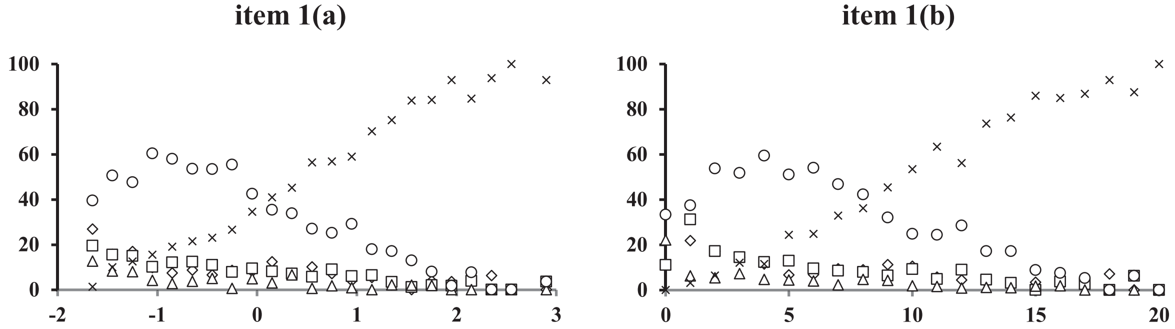

FIG. 5. The IRC of item 1, showing the percentage of respondents by choice, with symbols for the choices being ◇ = A, □ = B, Δ = C, o = D, and × = E. The correct choice was E. The horizontal scale is the ability in plot 1(a), and the total test score in plot 1(b).

(IRCs), in which the vertical axis represents the percentage of respondents, and the horizontal axis represents the ability level as created by the PARSCALE program; the total score is not used as a surrogate for ability, as in Refs. [11–12]. Students with the same overall test score need not share an ability level, because the pattern of their answers still can differ, but the total score and the ability typically have a robust correlation in a well-designed test. To check whether this makes a difference, a linear model was fit to predict the ability θ from the TUV raw score. The model $\theta = 4.84(\text{raw score}) - 6.76$, with coefficient of determination $R^2 = 0.98$, can be used to estimate the ability from the total TUV raw score (within the data on Thai students, not necessarily in general). To display example IRC plots against the ability and the total score, the curves for item 1 are shown in Fig. 5. Each of the five choices in the item gets its own IRC, with symbols ◇ = choice A, □ = choice B, Δ = choice C, o = choice D, and × = choice E. The graphs are very similar, demonstrating the tight relation of the ability and the total score as mentioned before. In this paper, we chose to present the IRCs plotted against the ability, in order to easily compare with results of IRT analysis.

The correct response in item 1 is choice E, which in the IRC plots has a consistently increasing steady trend. In other words, higher ability corresponds to higher probability of the correct answer to item 1, which is desirable. This indicates suitable discrimination power of choice E in item 1, consistent with the results of IRT analysis ($a = 0.91$, in Table II). The ICC shown in Fig. 1 is the logistic curve fit to the IRC for choice E. The most popular distractor in item 1 was D, which also has discrimination

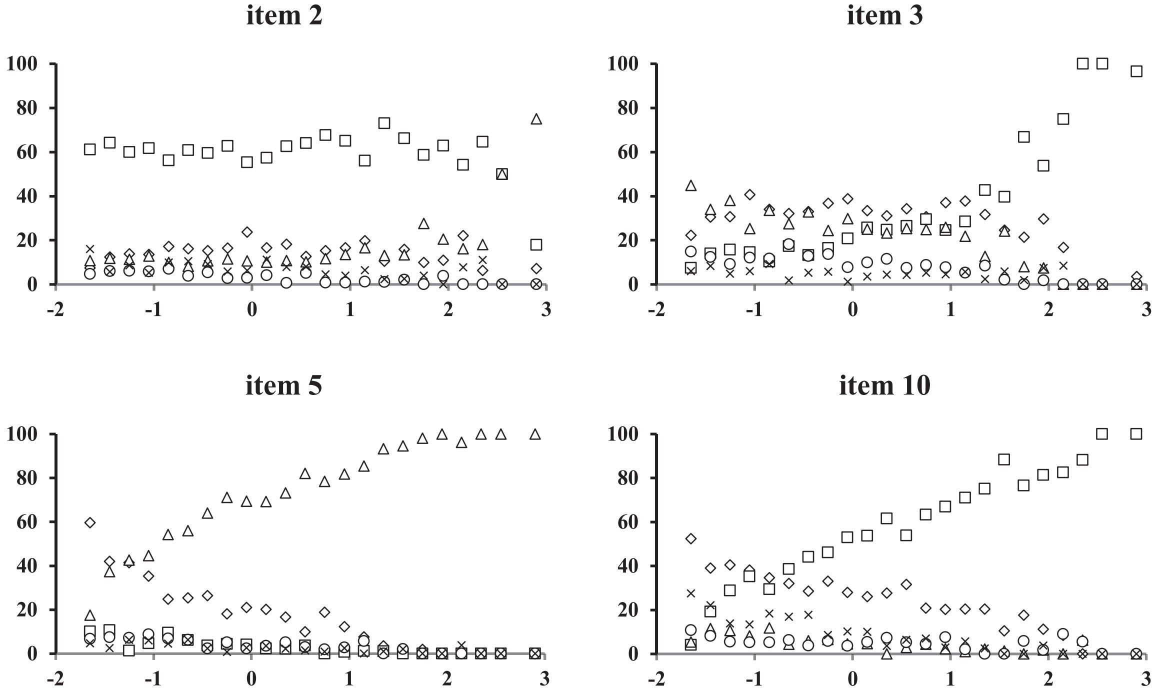

FIG. 6. IRCs for items 2, 3, 5, and 10, representing the fractions of respondents at any given ability that chose option ◇ = A, □ = B, Δ = C, o = D, or × = E.

power: many low-ability students selected D, and the frequency of this choice consistently decreases with ability. The other distractors in item 1 function quite well. This IRC pattern is similar to those of items 8, 11, 13, 15, 17, and 19.

In the IRC for item 2 shown in Fig. 6 the correct choice C has a relatively flat graph at low-ability levels and starts to increase around ability 1.5. This agrees with IRT analysis that only considers the correct option in an item, in showing discrimination power only at high-ability levels. Moreover, the IRC shows that the students were attracted to choose option B in item 2 almost uniformly across the ability levels. This distractor had poor discriminating power, as did the other distractors in item 2. The correct choice C in item 3 had discrimination power at abilities exceeding 1, which agrees with IRT analysis ($a = 1.04$), while its distractors drew about equal attention from the low-ability students (ability < 1). This is similar to item 12. In item 5, the correct option C and the distractor A function very well, but the other distractors do not: they attracted few students from any ability level, with flat response curves. The IRCs of the remaining TUV items are similar to item 10, in which all choices functioned quite well: the distractors were more popular among the low-ability students, and their curves gently decreased with ability, while the trend for the correct choice was increasing with ability.

VI. CONCLUSIONS

Results in this study indicate that the 20-item TUV test, with 5 choices per item, is most useful for testing the vector concepts when the ability of the examinees is from medium to high. It can be applied as a pass or fail threshold instrument at a somewhat high-ability value ($\theta \approx 1.1$). This insight is clearly provided by the test information function. Items 2, 3, and 12 are useful for separating examinees at high-ability levels. There is a very strongly held misconception represented in the distractors of items 8 and 15, shown by the biased preferences of low-ability students for these distractors. Because of this attraction bias of a distractor, the low-ability students have poorer performance in items 8 and 15 than random guessing would give. In contrast, the correct choices of items 4 and 5 are the most attractive responses to the very low-ability students, who have a $>30\%$ chance of answering each item correctly. Moreover, the IRC analysis covering all distractors disclosed that some distractors in the TUV did not function well. For example, choice B of item 2 had a flat response to ability, indicating it discriminates poorly. The students were equally likely to choose distractor B, regardless of their ability level. The distractors B, D, and E of item 5 did not function well either, attracting few students overall. However, as mentioned, the item and ability parameters reported in this study only pertain to the TUV responses by Thai first-year science and engineering students. Rescaling may be required for transfer of the current results to other groups of examinees.

Overall, the approach and findings of the current study may be used to develop and improve testing, and enhance its sensitivity and effectiveness within a given range of abilities. Test developers can analyze item and ability parameters using IRT, and distractors in an item can be assessed with the IRC technique. Moreover, using IRT with the item parameters held constant, the same group of students can be tested before and after instruction to determine the learning gains in ability. Further studies of TUV or its modifications could, in particular, explore the dimensionality of latent traits and implement the multitrait model of item response theory. The current results and approach can directly benefit anyone who uses the TUV, to gain improved accuracy of diagnosis.

ACKNOWLEDGMENTS

The research was supported by Thailand Research Fund and Prince of Songkla University under Grant No. TRG5780099. The authors are grateful to all participating students, Prof. Seppo Karrila, and Prof. Helmut Dürrast for valuable data and suggestions.

[1] P. Barniol and G. Zavala, Test of understanding of vectors: A reliable multiple-choice vector concept test, Phys. Rev. ST Phys. Educ. Res. **10**, 010121 (2014).

[2] R. K. Hambleton, H. Swaminathan, and H. Rogers, *Fundamentals of Item Response Theory* (Sage Publications, Inc., Newbury Park, CA, 1991).

[3] R. K. Hambleton and R. W. Jones, Comparison of classical test theory and item response theory and their applications to test development, Educ. Meas. **12**, 38 (1993).

[4] X. Fan, Item response theory and classical test theory: An empirical comparison of their item/person statistics, Educ. Psychol. Meas. **58**, 357 (1998).

[5] P. MacDonald and S. Paunonen, A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory., Educ. Psychol. Meas. **62**, 921 (2002).

[6] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[7] C. L. Hulin, F. Drasgow, and C. K. Parsons, *Item Response Theory: Application to Psychological Measurement* (Dow Jones-Irwin, Homewood, IL, 1983).

[8] D. Harris, Comparison of 1-, 2-, and 3-parameter IRT models, Educ. Meas. **8**, 35 (1989).

[9] M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking Methods and Practices*, 2nd ed. (Springer Science and Business Media, New York, 2004).

[10] D. T. Mathilda, *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT User Manual* (Scientific Software International, Inc., Lincolnwood, IL, 2003).

[11] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: item response curves and test quality, Am. J. Phys. **74**, 449 (2006).

[12] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the force concept inventory, Am. J. Phys. **80**, 825 (2012).

[13] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[14] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[15] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[16] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[17] G. Kortemeyer, Extending item response theory to online homework, Phys. Rev. ST Phys. Educ. Res. **10**, 010118 (2014).

[18] Y. J. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).

[19] P. Barniol and G. Zavala, Students' difficulties with unit vectors and scalar multiplication of a vector, AIP Conf. Proc. **1413**, 115 (2012).

[20] N. Nguyen and D. E. Meltzer, initial understanding of vector concepts among students in introductory physics courses, Am. J. Phys. **71**, 630 (2003).

[21] A. F. Heckler and T. M. Scaife, Adding and subtracting vectors: The problem with the arrow representation, Phys. Rev. ST Phys. Educ. Res. **11**, 010101 (2015).

[22] M. L. Stocking and F. M. Lord, Developing a common metric in item response theory, Appl. Psychol. Meas. **7**, 201 (1983).