Analysis to Develop Computerized Adaptive Testing with the Force Concept Inventory

J Yasuda¹, N Mae², M M Hull³ and M Taniguchi⁴

¹ Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan
² Faculty of Engineering Science, Kansai University, Suita, Osaka 564-8680, Japan
³ Austrian Educational Competence Centre Physics, University of Vienna, Vienna, 1090, Austria
⁴ Center for Teacher Education, Meijo University, Nagoya, Aichi 468-8502, Japan

Abstract. As a method to shorten the test time of the Force Concept Inventory (FCI), Computerized Adaptive Testing (CAT) is suggested. CAT is a test administered on a computer, where items (i.e. questions) are selected based on the responses of examinees to prior items. As a step of the development, we conducted analyses to find an optimal way for the administration of CAT with the FCI. Specifically, since CAT is based on Item Response Theory (IRT), we examined which IRT model is the most preferable. Using 2812 responses of the FCI of Japanese students, we estimated the item parameters of the One-, Two-, Three-, and Four-parameter logistic model of IRT and then we evaluated the statistics for the goodness of fit and for model selection. Based on the analysis, we suggest that the Two-parameter logistic model is the most preferable for the administration of CAT with the FCI.

1. Introduction
The Force Concept Inventory (FCI) is one of the most widely-used assessment tests in physics education [1-3]. It probes student conceptual understanding of Newtonian mechanics, especially, on issues of force. The test has 30 items with five choices, which students typically take 15 to 30 minutes to complete. The test is designed to “respond properly” when considered from the student’s point of view [4]. For example, the distractors are designed based upon knowledge of students’ common naïve conceptions [5, 6]. Furthermore, the items use everyday speech in order to better elicit what the student personally considers to be correct as opposed to an answer memorized by rote from physics class. The FCI has played an important role in analyzing the effects of newly-developed pedagogy including interactive-engagement methods [7, 8].

When administering the FCI in a classroom, it typically takes about 40 minutes, including the time needed to orient students to the survey. Han et al. [9-11] pointed out that this takes up nearly an entire class and may affect the willingness of instructors to add the FCI to their crowded schedules. In order to shorten the test time, Han et al. divided the FCI into two half-length tests which contain different subsets of the original FCI, but still cover the same set of concepts. They showed, using a large quantitative data set, that the half-length FCIs yield almost identical conversion mean scores compared to the full FCI with an overall uncertainty of less than 3%.

We have the same goal in mind; namely, our objective is developing a method to shorten the test time of the FCI without compromising its precision. In this paper, we suggest another approach based
on Computerized Adaptive Testing (CAT). CAT is a test administered on a computer, which utilizes an algorithm-based approach [12]. In CAT, the items (i.e. questions) are chosen and administered based on the responses of students to prior items. For example, if a student answers an item correctly, the student will next need to answer a more difficult item. On the other hand, if a student answers an item incorrectly, the student next answers an easier item (figure 1). Using this method, high (low) ability students do not need to answer items that are too easy (difficult) for them. In this way, the test time can be significantly shortened. Due to this efficiency, CAT is recently becoming widely used, for example, in PISA 2018 [13], in English assessment tests [14] and also in physics education research [15].

We are investigating the potential use of CAT with the FCI (FCI-CAT). In this paper, as a midterm report, we show the results of our analysis to find an optimal way for the administration of the FCI-CAT.

2. Methodology
We are concerned with the question of how to best implement CAT in the administration of the FCI. To address this research question, since CAT is based on Item Response Theory (IRT), we estimate the item parameters of the FCI based on several IRT models and examine which model is most suitable. In this section, we explain the basics of IRT and then describe the setting of our analysis and survey.

2.1. Item response theory basics
IRT models show the relationship between the latent trait measured by the instrument and an item response [16]. Since a response of the FCI is scored as correct or incorrect (coded as 1 or 0), we focus on the dichotomous models of IRT. Furthermore, we focus on the unidimensional models, where we
assume that the latent trait measured by the FCI is dominated by a single proficiency, namely, student conceptual understanding of Newtonian mechanics.

Among the dichotomous unidimensional models in IRT, the simplest model is the One-Parameter Logistic (1PL) model. An example of the analysis based on the 1PL model is shown in figure 2. In the graph, the horizontal axis, $\theta$, shows the latent trait, which is associated with the total number-correct score of the FCI. (In the case of the 1PL model, $\theta$ depends monotonically on the total number-correct score of the FCI with one-to-one correspondence.) The latent trait distribution in a designated group is often standardized, namely, the estimated mean of $\theta$ is set to 0 and the estimated standard deviation of $\theta$ is set to 1. The vertical axis, $P_i(\theta)$, shows the probability that an examinee given $\theta$ answers the $i$th question of the FCI correctly. In the graph, the white circles show the empirical plot, where examinees were grouped into 10 intervals based on $\theta$, and the proportion in each group who answered the item correctly was calculated. The dashed lines on the white circles show the 95% confidence intervals. The behaviour of the empirical plot in figure 2 is normal, namely, $P_i(\theta)$ is monotonically increasing as the function of the $\theta$.

Based on the 1PL model, the empirical data in figure 2 is fit with the following logistic function,

$$P_i(\theta) = \frac{1}{1 + \exp(-(\theta - b_i))}$$

where $b_i$ is the difficulty parameter of the $i$th question of the FCI. (We follow the notation of the mirt package with the traditional IRT parametrization [17]). The function $P_i(\theta)$ with the estimated $b_i$ is shown as the curve in figure 2. The difficulty parameter corresponds to the $\theta$ value where $P_i(\theta)$ equals to 0.5. Difficult items have higher values of $b_i$.

The Two-Parameter Logistic (2PL) model has the second parameter $a_i$, which is called the discrimination parameter. An example based on the 2PL model is shown in figure 3. The parameter $a_i$ corresponds to the slope at $\theta = b_i$, where the slope of the curve is steepest. The item with steeper slope can better distinguish examinees with different levels of ability. In figure 3, the probability $P_i(\theta)$ is calculated based on the 2PL model using the same data set as figure 2. The empirical plot is fit with the following function,

$$P_i(\theta) = \frac{1}{1 + \exp(-a_i(\theta - b_i))}$$

whose shape is shown as a curve in figure 3.
The Three-Parameter Logistic (3PL) model has the third parameter $g_i$. An example based on the 3PL model is shown in figure 4. The parameter $g_i$ is the asymptotic value of $P_i(\theta)$ when $\theta$ approaches negative infinity. It represents the probability that an examinee would answer an item correctly with guessing, so it is called guessing parameter. Although $g_i$ approaches 0.2 if choosing randomly among five choices, since the distractors of the FCI are designed to be chosen by non-Newtonian thinkers, $g_i$ is expected to be smaller than 0.2. Based on the 3PL model, the empirical plot is fit with the following function,

$$P_i(\theta) = g_i + \frac{1 - g_i}{1 + \exp(-a_i(\theta - b_i))}$$

whose shape is shown as a curve in figure 4.

The Four-Parameter Logistic (4PL) model has the fourth parameter $u_i$, which is called the upper bound parameter. The parameter $u_i$ is the asymptotic value of $P_i(\theta)$ when $\theta$ approaches positive infinity. It represents the probability that a Newtonian thinker does not answer correctly, namely, the response is a false negative. An example based 4PL model is shown in figure 5. The empirical plot is fit with the following function,

$$P_i(\theta) = g_i + \frac{u_i - g_i}{1 + \exp(-a_i(\theta - b_i))}$$

whose shape is shown as a curve in figure 5.

Why is IRT necessary for CAT? Let us consider the situation in which student A takes a test with item set X and gets 60 points and another student B takes another test with item set Y and gets 80 points. With the use of classical test theory, which uses raw scores, we might conclude that student B is more proficient in the trait being measured by the test. Clearly, however, that would be a premature conclusion because item set Y could be, on average, easier than item set X. In using IRT, we estimate and equate the item parameters of X and Y, and that enables us to in turn estimate the latent trait parameters of both students, $\theta_A$ and $\theta_B$. Since these thetas are on the same metric, we can compare the values and evaluate which student has higher ability. In CAT, students answer different questions depending on their responses as described in figure 1. Therefore, in order to compare the results of CAT, it is necessary to calculate student ability based on item response theory.

### 2.2. Data collection

As described above, in order to develop the FCI-CAT, it is necessary to estimate the item parameters of the FCI based on IRT models. In order to do that, we administrated and collected data from the full
FCI in Japan from April 2015 to April 2018. The examinees were students at the beginning of introductory physics courses at one public university and four private universities, which are middle-rank universities in Japan. The total number of survey responses was 2882. From this, we excluded the responses of students who did not answer some of the questions, who wrote a letter which was not one of the choices available for a given question, or who wrote the same or serial letters continuously. In total, the number of valid responses was 2812. Most of the respondents were first-year students. Students were from different departments, mainly the departments of science, technology, and agriculture. The survey was conducted during class, so as to help ensure that students would concentrate on the survey. The respondents were not given any incentive to participate (in the form of money or extra credit).

3. Result

3.1. Estimation of the item parameters
We estimated the item parameters of the FCI based on the 1PL, 2PL, 3PL, and 4PL models. As an example, the result of the estimation for the difficulty parameters based on the 2PL model is shown in figure 6. In the graph, the horizontal axis shows the item numbers, the vertical axis shows the value of the estimated difficulty and the error bars indicate 95% confidence intervals. Since we analyzed with a large sample (2812 responses), the error bars are quite small and the parameters are well estimated. Note that most of the difficulty parameter are between -2 and 2 as expected [18], but the difficulty of question 29 is exceptionally small compared with those of the other questions. This is because the goodness of fit for question 29 is insufficient and the estimation is unstable.

3.2. Examination of the assumptions of IRT
We examined whether the FCI satisfies the typical assumptions of IRT: unidimensionality, local independence and goodness of fit. We tested unidimensionality based on the eigenvalues of the inter-item tetrachoric correlation matrix. Figure 7 is the scree plot which shows the ordered eigenvalues of the correlation matrix calculated using our response data of the FCI. From the graph, we can see the first eigenvalue is about five times larger than the second one (see Figure 7). Scree plot for the tetrachoric correlation matrix using our FCI data suggests a single proficiency is dominant.

![Figure 7. Scree plot for the tetrachoric correlation matrix using our FCI data](image-url)
In order to evaluate the goodness of fit for the whole 30 items of the FCI, we used the Standardized Root Mean Square Residual (SRMSR). The SRMSR is recommended in literature as an index for the goodness of fit and it is suggested that SRMSR ≤ 0.05 be used as a cutoff for well-fitting IRT models [20]. We found that the SRMSR is 0.079 for the 1PL model and 0.043 for the 2PL model, which means the goodness of fit is sufficient for the 2PL model but not for the 1PL model. Since the SRMSR decreases as the number of parameters of a model increases, the goodness of fit for the 3PL model and the 4PL model are also sufficient. This result indicates that it is reasonable to assume unidimensionality for our analysis of the FCI data. According to Wang and Bao [19], if the test data are unidimensional, it automatically satisfies the local independence assumption, therefore, we can also assume local independence for our analysis.

![Figure 8](image)

**Figure 8.** A result of examining the goodness of fit for question 13

As an example of the analyses for the goodness of fit, we show the result for question 13 of the FCI. In figure 8, white circles show the empirical plot and the curves show the logistic function approximating the data. As shown, in the case of the 1PL model, there is a gap between the empirical plot and the curve, therefore the goodness of fit seems to be insufficient. On the other hand, in the case of the 2PL model which has an additional parameter adjusting the slope of the fitting function, the gap between the plot and the curve becomes smaller.

3.3. **Which model is the most suitable for CAT?**

Our primary goal in this paper is to examine the most preferable IRT model for the FCI-CAT. Generally speaking, we prefer the model which fits the response data well, and also which has fewer parameters. The balance of these conditions can be examined using a statistic called Bayesian information criterion (BIC) [21]. BIC is defined by the following equation,

\[
\text{BIC} = d + p \ln l
\]

where \(d\) is the deviance of the response data from the fitted function, \(p\) is the number of parameters, and \(l\) is the number of respondents. The BIC increases if the deviance increases and also if the number of parameters increases, so that the model with the lowest BIC is suggested as the most preferable model.
The BIC for each of the four IRT models is shown in figure 9. We found that the BIC of the 1PL model is larger than that of the other models. On the other hand, the BIC of the 2PL, 3PL, and 4PL model are comparable. From this result, with the analysis on the goodness of fit in Section 3.2, we conclude that the 1PL model is not preferable. For the remaining three models, it is reasonable to think that the simplest model is most preferable from other points of view; for example, we can reduce the time of numerical calculations, the estimation of the item parameters is more stable, and so on. For these reasons, we suggest that the most preferable model is the 2PL model for the administration of the FCI-CAT.

4. Summary
In order to shorten the test time of the FCI, we are investigating the potential use of computerized adaptive testing (CAT). We estimated the item parameters of IRT models and analyzed how best to implement CAT. We found that the 2PL model is the best among the 4 logistic models, as it is the simplest model among the models with comparable small BIC.

In future work, we plan to do the following things. First, we will examine the measurement invariance for the FCI, which means that the item parameters are invariant if linearly transformed. We will compare our results with previous studies, for example [19, 22-23], with the differential item functioning [24].

We will further analyze which number of questions is best for the FCI-CAT. Specifically, we will consider between four-item sequences, which require at least $2^4 - 1 = 15$ FCI items, and five-item sequences, which require at least $2^5 - 1 = 31$ FCI items. As the FCI itself consists of only thirty items, five-item sequences would require the same item appearing multiple times in the tree diagram demonstrated in figure 1. We will examine which sequence length is better using a simulation. This simulation will compare the measurement errors of the respondents’ ability parameters and examine how sequence length affects the estimates. We will also analyze the sizes of the measurement errors themselves and examine the validity of the FCI-CAT.

With these preparations, we will, as a trial survey, administer the FCI-CAT for Japanese students. In the survey, students will take the FCI-CAT using a smart phone application, YU Portal, which was developed at Yamagata University in Japan. By using smartphones, students can take CAT in the classroom without moving to a place where there are computers (computer room or their home, etc.). This allows for greater concentration of students, since teachers can monitor the students during the test. After the trial survey, we will interview the respondents to find any problems of our test, for example, in regards to the interface.
5. References

[1] Hestenes D, Wells M and Swackhamer G 1992 Phys. Teach. 30 141
[2] The latest version of the FCI revised in 1995 and the revised Tables of [1] are available in https://www.physport.org (Retrieved 10/10/2019)
[3] Hestenes D and Halloun I 1995 Phys. Teach. 33 502
[4] Redish E F 2003 Teaching Physics with the Physics Suite (Hoboken: Wiley) p 98
[5] Halloun I and Hestenes D 1985 Am. J. Phys. 53 1043
[6] Halloun I and Hestenes D 1985 Am. J. Phys. 53 1056
[7] Redish E F, Saul J M and Steinberg R N 1997 Am. J. Phys. 65 45
[8] Hake R 1998 Am. J. Phys. 66 64
[9] Han J, Bao L, Chen L, Cai T, Pi Y, Zhou S, Tu Y and Koenig K 2015 Phys. Rev. ST. Phys. Educ. Res. 11 010112
[10] Han J, Koenig K, Cui L, Fritchman J, Li D, Sun W, Fu Z and Bao L 2016 Phys. Rev. Phys. Educ. Res. 12 020122
[11] The half-length FCI s are also available in https://www.physport.org (Retrieved 10/10/2019)
[12] Magis D, Yan D and von Davier A A 2017 Computerized Adaptive and Multistage Testing with R (Cham: Springer)
[13] OECD 2019 PISA 2018 Assessment and Analytical Framework (Paris: OECD Publishing) p 53
[14] Cambridge English language assessment 2015 Computer-adaptive Testing Research Notes 59
[15] Morpew J W, Mestre J P, Kang H-A, Chang H-H and Fabry G 2018 Phys. Rev. Phys. Educ. Res. 14 020110
[16] DeMars C 2010 Item Response Theory (New York: Oxford University Press) p 3
[17] Chalmers R P 2012 J. Stat. Softw. 48 1
[18] DeMars C 2010 Item Response Theory (New York: Oxford University Press) p 21
[19] Wang J and Bao L 2010 Am. J. Phys. 78 1064
[20] M-Olivares A 2013 Meas. Interdiscip. Res. Perspect. 11 71
[21] Schwarz G 1978 Ann. Statist. 6 461
[22] Planinic M, Ivanjek L and Susac A 2010 Phys. Rev. ST. Phys. Educ. Res. 6 010103
[23] Traxler A, Henderson R, Stewart J, Stewart G, Papak A and Lindell R 2018 Phys. Rev. Phys. Educ. Res. 14 010103
[24] Desjardins D C and Bulut O 2018 Handbook of Educational Measurement and Psychometrics Using R (Boca Raton: CRC Press) p 258