Development and Validation of Grade 10 Physics Test in K-12 Science Curriculum

J B Jalon¹ and V M Mistades²

¹ De La Salle Santiago Zobel School, Metro Manila, Philippines
² De La Salle University, Manila, Philippines

E-mail: joel.jalon@dlszobel.edu.ph

Abstract. The purpose of this study was to develop and validate a Physics test that measures the competency level of Grade 10 students in the new Philippine K-12 Science Curriculum. The test was composed of items in electromagnetic spectrum, light, and electricity and magnetism. During the test development, the test was divided into 2 sets and underwent content and face validation from 6 Physics content-experts. It was determined that for both sets, at 95% confidence, the levels of reliability ranged from moderate to good. After revising the items based on the feedback from the experts, Set A and Set B were administered to 569 students and 526 students, respectively, and 520 responses from each set were included in the item analysis. Using the IRT, items that violated the unidimensionality IRT assumptions were removed and item characteristics were computed for item selection. The data generated were most fit in unidimensional 2-PLM with acceptable marginal reliability response pattern scores of 0.70 and 0.67 for Set A and Set B, respectively. The final form of the test contains 35 items covering all the competencies in said topics with medium to hard difficulty level and high discrimination index. The distribution of the items for the final Physics test form is balanced; 50% measures low order thinking skills and the other 50% measures higher order thinking skills. The results also showed that item parameters estimated using IRT were invariant. Based on the hard difficulty level of the test sets and average scores, the students performed fairly-well.

1. Introduction
On May 15, 2013, the then-Philippine President Benigno S. Aquino signed Republic Act No. 10533 also known as Enhanced Basic Education Act of 2013. The Philippine Educational System shifted from the Basic Education Curriculum (BEC) to K-12 BEC. However, the transition to K-12 program already started in 2011. In the new K-12 Science curriculum, starting Grade 3 to Grade 10, the areas to be taught are matter, Earth and space, living things and their environment, and force, motion, and energy. The scope and sequence of the contents are designed in such a way the lessons and skills are reencountered in each grade level with increasing complexity. In the old BEC, science content was not revisited in each grade level and all topics in a science area (i.e. Earth Science) were the focus in each grade level with increasing depth throughout the school year. For the next year level, all lessons are in another science area. Before the implementation of K-12 program, kindergarten was not compulsory, and high school was just 4 years, but in the K-12 BEC there are additional 2 years (Grades 11 and 12).
In order to gauge the effectiveness of the new curriculum, assessment and evaluation play a vital role. Examinations are very important because the results from the test can be used to assess the preconceptions of the students and in return design appropriate activities to address the misconceptions, and to measure how much the students have learned at the end of a lesson. There is an increasing number of studies that responded to this call in the Philippines. Researchers [1,2,3,4] from different fields developed items, such as for Physics, Biology, Music, and Mathematics. There are two widely used theories in validating and developing tests, and these are Classical Test Theory (CTT) and Item Response Theory (IRT). An increasing number of studies showed that IRT is more advantageous than CTT [5,6,7]. Unlike CTT, the item parameters of an item in IRT do not change across different groups of test takers, which makes possible to compare the groups. IRT primary focuses on item level, while CTT focuses on test level. IRT can also compute how well each item can measure a specific latent trait. The IRT ability to compute item characteristics at item level is very useful in selecting items that are appropriate for the envisioned type of test to be administered to a target group. However, even with the increasing number of literatures showing that IRT is advantageous than CTT, there are still item writers who used CTT in developing and validating tests, which was the case of the researchers in the Philippines that were mentioned earlier.

2. Statement of the Objective
The main purpose of this study was to develop and validate a test that measures the competency level of Grade 10 (G10) students in Physics under the new K-12 Science curriculum. The test was composed of items aligned with Physics topics and competencies in G10 for electromagnetic spectrum, light, and electricity and magnetism and G10-related competencies and topics from Grade 9 (G9) and Senior High School (SHS) Physical Science.

3. Methodology

3.1 Developing the G10 Physics Test
The major stages in developing and validating the test items are as follows:

3.1.1. Determining the Competencies. The items in the test were focused on the topics and competencies for electromagnetic spectrum, light, and electricity and magnetism for G9, G10, and SHS Physical Science set by Department of Education (DepEd) under the new K-12 Science curriculum. The topics and competencies for G9 and G10 were included in constructing the test to measure the level (below, on target, or advance) of competency of G10 students.

Unpacking of the curriculum guide was done to uncover the competencies not explicitly written in the said curriculum, but are needed for the students to completely grasp the content to be learned and successfully accomplish the performance task/s described in each grade level and/or prerequisites for the related topics in higher grade level. In unpacking the curriculum, the suggested method by Konrad, et. al [8] was incorporated. The unpacked competencies were then added to the list of competencies found in the K-12 Science curriculum.

There were total of thirty-five (35) identified competencies for the said topics in G9, G10, and SHS, distributed as follows: 7 competencies for electromagnetic spectrum; 15 competencies for light; and 13 competencies for electricity and magnetism.

3.1.2. Deciding the Test Questionnaire. The test items were written in conventional multiple-choice format with four (4) choices because conventional multiple-choice is suitable in measuring different cognitive skills. The responses of multiple-choice test are easier and faster to check, and rubrics and trained judges are not needed. Thus, rater-effect, which is a threat to validity is avoided [9].

3.1.3. Constructing the Table of Specification. The researcher targeted to have one item per competency. Since there are thirty-five identified competencies, there should be thirty-five items for
the final test form. Lorenzana [1] pointed out that more items are needed for each competency since items are being rejected during the experts’ validation and item analysis stages. To have high item survival rate, the researcher constructed three questions measuring each competency and targeted to measure higher cognitive level skills. The target total items for the pilot testing was 105. The following are the projected number of items to be written for each lesson: 21 items for electromagnetic spectrum; 45 items for light; and 39 items for electricity and magnetism. The distribution of the projected items are as follows: 37% Knowledge, 20% Understanding, 31% Applying, 9% Analyzing, and 3% Evaluating.

3.1.4. Expert Examination of the Instruments. Haladyna [9] noted that test type, the way the test is administered, and test anxiety contribute negatively to test validity. To factor out the test anxiety which affects the validity of the test, the 105 items were divided into 2 sets taking in consideration the distribution of items measuring the competencies and test length.

To establish content validity and face validity, three teacher-experts with master’s and doctoral degrees in Physics were asked to rate and comment each test set based on the guidelines adapted from Haladyna [9]. The rating depends on the number of guidelines that were followed in constructing each test item: 5 is the highest rating and 0 is the lowest. Items that were rated 2 in content and/or face/structure, and with mean score of 3 and below were removed. Items with mean score between 3 to 3.5 were revised based on the comments and suggestions of the experts. Items with mean score of at least 3.5 were retained unless the experts have suggested revisions. To determine the agreement of the experts in rating the content and the face/structure, inter-rater reliability was computed using the intraclass correlation coefficient (ICC) at 95% confidence via a statistical analytical tool.

Ten G10 students were also asked to rate and comment on the clarity of each item and each choice using a 4-point Likert scale.

3.1.5. Pilot Testing. Set A was administered to 569 G10 students and Set B was administered to 526 G10 students. The test sets were administered to different schools to ensure that the respondents are heterogeneous in terms of their ability level in Physics.

3.1.6. Item Calibration. It is very important to examine the dimensionality of the test. If the test does not satisfy the unidimensional assumption when unidimensional IRT is implemented, the estimate parameters and standard error are dubious. The standardized LD $\chi^2$ statistic [10] for item pair was used. Values greater than [10] show violations and should be removed. In removing an item, the following additional conditions were followed: An item that appeared more than once in the set of items pairs should be removed; and In an item pair with an item that is not included in another item pairs, the item with less item information function values should be removed.

The validity of the test in IRT depends on the model used for treating the data [11]. The model to be used should represent the data gathered. If this assumption is not met, conclusions from the test and items can be mistaken [12]. The S - $\chi^2$ item-fit statistics as proposed by Orlando and Thissen [13] was used, because it effectively identifies dichotomous items that do not fit the IRT model; items with $p > 0.05$ are considered fit.

Developing a test is costly and removing all items that do not fit the model is not ideal. An option is to remove one item which can make the data a better fit to the model [14]. Misfit items that have the least item information values were removed. The process of calibrating the item and reassessing of item level fit was repeated until all the remaining items fit the model. This was followed by removing items that do not give information on the examinees’ ability based on the item information function values for any ability level from -2.8 to 2.8.

Lastly, selection of the remaining items either to be retained, revised, or removed was based on Figure 1. The range of the difficulty level and discrimination level is consistent with Baker [6].
In addition, items with negative value of $a$ were removed. Items measuring G9 and G10 competencies exceeding the maximum suggested difficulty level (3.0) were revised, while items measuring G11/12 competencies exceeding the maximum suggested difficulty level (3.0) were reviewed to either revise or retain. All revised and retained items were automatically part of the item bank.

### 3.1.7. Item Selection for the Final Test Form

For the final test form, the test item measuring the highest possible cognitive level for each competency from the test bank was chosen. Furthermore, among the items measuring G9 and G10 competencies, the items with very high discrimination level and medium difficulty and items close to this type when the preferred item type was unavailable were chosen, since the G10 Physics Test was intended to measure if the examinees are at least at the satisfactory proficiency level in the said competencies. From among the items measuring SHS competencies, the items with very high discrimination level and hard difficulty level and items close to this type when the preferred type was unavailable were chosen, since it should be expected that the item should be at least hard for G10 students. For the competencies with no corresponding items, items that were initially removed were reviewed and revised based on the experts’ ratings and following the conditions above to complete the Grade 10 Physics test.

### 3.2 Investigation of the invariant property of IRT

The estimated item parameters of the total respondents for each set and the item parameters estimated using each group were analyzed. In computing the estimated parameters for the 2 groups in each set same number of responses was used: the lower number of responses between the groups in each set was the basis for the number of responses used. For Set A group, 253, and for Set B group, 249 responses were used. The excess responses that were deleted were the responses with many blank items. To ensure the validity of the comparison among the estimated item parameters across the 3 groups in each set, the items that were compared were the items that survived the item calibration. In addition, only the items that were present among the 3 groups in each set were compared. While, for CTT all items were compared. The estimated item parameters were then analyzed using the repeated-measures of ANOVA because as stated by Adedoyin et al. [15] the Pearson Product-Moment Correlation Coefficient is not enough to test for the invariant property of IRT.

### 3.3 Interpreting the G10 Students’ Proficiency Level in Physics

In interpreting the proficiency level of the students, it is not enough to base only on the number of correct answers the students got because the difficulty level of the test and each item should also be taken into consideration. One cannot judge a person’s ability just based on the number of correct items obtained, rather, the item attribute should also be taken into account [16]. To ensure the validity of the estimated item parameters, only items that fit the IRT model and with 0 to positive item discrimination indices were used. Items with -$a_i$ were not included because further investigation to these items should be done.

### Table 1

**Decision Table of Difficulty and Discrimination Indices**

| Difficulty Level | Discrimination Level | Decision |
|------------------|----------------------|---------|
| Easy (-3 to <=0.5) | Very Low/Low (0.1 to < 0.65) | Revise |
|                  | Moderate (0.65 ≤ 1.35)   | Retain  |
|                  | High/Very High (≥ 1.35)  | Retain  |
| Average (-0.5 to <0.3) | Very Low/Low (0.1 to < 0.65) | Revise |
|                  | Moderate (0.65 ≤ 1.35)   | Retain  |
|                  | High/Very High (≥ 1.35)  | Retain  |
| Difficult (0.5 to 3.0) | Very Low/Low (0.1 to < 0.65) | Revise |
|                  | Moderate (0.65 ≤ 1.35)   | Retain  |
|                  | High/Very High (≥ 1.35)  | Retain  |

*Adapted from Cento and Cavero (2012) as cited by Elejo and Esanmonu (2018)*
4. Results and Discussion

4.1 Development of the G10 Physics Test

Three questions per competency was the target, but 2 competencies have less than 3 questions due to the limited variation for the item and limited lesson. A total of 102 items were constructed instead. Four items in Set A was rated 2 by one validator, and one item in Set B was rated 2 by one validator. These items were removed, while other items were revised or replaced based on the experts’ comments. For the pilot testing, Set A contained 47 items and Set B contained 50 items.

For Set A, the content validity and face/structure validity, ICC are 0.785 and 0.749, respectively. These suggested good reliabilities, their 95% confidence interval ranged from 0.649 to 0.873, and 0.594 to 0.851, respectively. Thus, it appropriate to conclude that the level of reliabilities of the two areas were moderate to good.

For Set B, the content validity and face/structure validity, ICC are 0.691 and 0.709, respectively. Similar to Set A, these likewise suggested moderate reliabilities, their 95% confidence interval ranged from 0.506 to 0.814, and 0.536 to 0.825, respectively. We likewise say that the level of reliabilities of the two areas for Set B were moderate to good.

Among the three general unidimensional models of IRT, 2PLM was the best model for the data for both sets because most of the items fitted in this model; 41 items from Set A and 47 items from Set B. In 3PLM, only 38 items from Set A and 36 items from Set B fitted; while items in 1PLM were no longer counted because at the start of the item analysis, the results already showed that almost all items violated the assumptions of unidimensionality and local independence.

The marginal reliability for response pattern scores of Set A was 0.70 and of Set B was 0.67 and were greater than the acceptable value: acceptable values are 0.623 and above as noted by Zakria et. al. [17]. This indicated that both test sets were reliable.

![Figure 2](image-url) Sample Item that was Revised from Set B Based on the Item Parameters

After removing, revising, and retaining items based on Figure 1, 36 items remained; 18 items from each set. These items measure only 22 out of 35 competencies. These items were then added to the item bank. An example of revised item based on the item parameters is shown in Figure 2.

To complete the G10 Physics test, items that were initially removed were revised based on the suggestions and comments from the experts. The final G10 Physics Test form contains questions classified as: 34% Remembering, 17% Understanding, 34% Applying, 11% Analyzing, and 3% Evaluating. The distribution of the items for the final Physics test form was balanced: 50% measures low order thinking skills and 50% measures higher order thinking skills.
4.2 Investigation on the Invariant Property of IRT and CTT

After item calibration under IRT, 41 items survived for Set A-All, 37 items for Set A-Group 1, and 41 items for Set A-Group 2. However, only 29 items were similar among the 3 groups. For the Set A groups, the $a_i$ data met the assumption of sphericity. The value of F was 0.624, which reaches significance with a $p=0.540$ (greater than the 0.05 alpha level). This indicated that there was no significant difference among the means of the $a_i$ indices (0.230 vs 0.225 vs 0.264) across the 3 groups in Set A, and the item discrimination indices across the group was invariant. The $b_i$ data also met the assumption of sphericity. The value of F was 1.023, which reaches significance with a $p=0.366$. This indicated that there was no significant difference among the means of the $b_i$ indices (-8.974 vs -1.839 vs -0.861) across the 3 groups in Set A, and the $b_i$ indices across the group was invariant.

After item calibration under IRT, 42 items survived for Set B-All, 42 items for Set B-Group 1, and 41 items for Set B-Group 2. However, only 35 items were similar among the 3 groups. For the Set B groups, the $a_i$ data met the assumption of sphericity. The value of F was 0.504, which reaches significance with a $p=0.607$. This indicated that there was no significant difference among the means of the $a_i$ indices (0.194 vs 0.093 vs 0.128) across the 3 groups in Set B, and the $a_i$ indices across the group was invariant. The $b_i$ data also met the assumption of sphericity. The value of F was 1.770, which reaches significance with a $p=0.178$. This indicated that there was no significant difference among the means of the $b_i$ indices (-21.974 vs 16.149 vs 2.503) across the 3 groups in Set B, and the $b_i$ indices across the group was invariant.

Now for Set A groups under CTT, the $a_i$ data met the assumption of sphericity. The value of F is 0.236, which reaches significance with a $p=0.790$. This indicated that there was no significant difference among the means of the $a_i$ indices (0.186 vs 0.181 vs 0.189) across the 3 groups (all A responses, AG1, and AG2) in Set A, and the $a_i$ indices across the group was invariant. The $b_i$ data also met the assumption of sphericity. The value of F was 7.231, which reaches significance with a $p=0.001$. This indicated there was a significant difference among the means of the $b_i$ indices of the 3 groups in Set A. A post hoc pairwise comparison using the Bonferroni correction, showed that the differences in the means of the $b_i$ indices between using all A responses and using only AG1 responses (0.295 vs 0.283, $p=0.013$) were significant. Similarly, the differences in the means of the $b_i$ indices between using AG1 responses and AG2 responses (0.283 vs 0.304, $p=0.029$) were significant. But, the differences in the means of the $b_i$ indices between using all A responses and using only AG2 responses (0.295 vs 0.304, $p=0.077$) did not reach significance. Therefore, it can be concluded that the $b_i$ indices across the group lacked the invariant property.

For the Set B groups under CTT, the $a_i$ data met the assumption of sphericity. The value of F was 14.896, which reaches significance with a $p=0.000$. This indicated that there was a significant difference among the means of the $a_i$ indices of the 3 groups (All B responses, BG1, and BG2). A post hoc pairwise comparison using the Bonferroni correction, showed that the differences in the means of the $a_i$ indices between using all B responses and using only BG1 responses (0.022 vs 0.0134, $p=0.109$) are not significant. But, the differences in the means of the $a_i$ indices between using all B responses and using only BG2 responses (0.0220 vs -0.0032, $p=0.000$), and between using only BG1 responses and using only BG2 responses (0.0138 vs -0.0032, $p=0.001$) were statistically significant. Thus, it can be concluded that the $a_i$ indices across the different groups failed to show the invariant property. The $b_i$ data also met the assumption of sphericity. The value of F was 15.981, which reaches significance with a $p=0.000$. This indicated that there was a significant difference among the means of the $b_i$ indices of the 3 groups in Set B. A post hoc pairwise comparison using the Bonferroni correction, showed that the differences in the means of $b_i$ indices between using all responses and using only BG1 (0.283 vs 0.295, $p=0.022$), between using all responses and using only BG2 (0.283 vs 0.266, $p=0.000$), and between using BG1 and BG2 responses (0.295 vs 0.266, $p=0.000$) was statistically significant. Thus, it can be concluded that the $b_i$ indices across the different groups were variant.
4.3 Grade 10 Proficiency Level in Physics

For the interpretation of the scores of the examinees who took the Set A test in the pilot testing, only 18 items were considered. The 8 items with -α were not considered because further investigation of these items should be done. The 18 items measure only 16 out of 35 competencies. Based on the test characteristic curve of the 18 items of the Set A Physics Test as shown in Figure 3, an examinee with a satisfactory ability level (θ=0) has the probability to get a score of only 7, while an examinee with a very high ability level (θ=3) has the probability to get a score of 12. This means that Set A test was a difficult test for the G10 students. This was evident in the difficulty level; 14 items were hard and very hard. The average score of Set A was 6.67 (=7) and 248 students (47.69%) got at least 7. This indicated that, in general, the examinees did almost fairly-well in the test given the difficulty level of the test.

![Figure 3. Set A Test Characteristics Curve](image)

![Figure 4. Set B Test Characteristics Curve](image)
For the interpretation of the scores of the examinees who took the Set B test in the pilot testing, only 18 items were also considered, and 10 items with \(-a\) indices were not considered. The 18 items measure only 16 out of 35 competencies. Based on the test characteristic curve of the 18 items of the Set B Physics Test as shown in Figure 4, an examinee with a satisfactory ability level (\(\theta=0\)) has the probability to get a score of only 6. Meanwhile, an examinee with a very high ability level (\(\theta=3\)) has the probability to get a score of only 10. This means that Set B test was also a hard test for the G10 students. This was evident in the difficulty level; 13 were hard and very hard. However, the average score of Set B was 6.26 (\(\approx6\)) and 300 students (57.69\%) got at least the minimum probable score for an examinee with satisfactory ability level can get. This indicated that, in general, the examinees did fairly well in the test given the difficulty level of the test.

5. Conclusions
The inter-rater reliability for both test sets ranged from moderate to good. 2PLM was the best model for the data of both test sets. Both test sets were valid and reliable. The results also showed that the item parameters of both test sets under the 2-PLM IRT were invariant across the different groups and different groups with varying sizes, but not under the CTT. The invariant property of IRT is one of the main reasons item developers choose IRT over CTT.

In general, both test sets were difficult for the G10 students. However, given the difficulty level of the test sets, the G10 students still performed fairly in the test. The final G10 Physics test form contained 35 items: 1 item per competency. The items selected for the final form measure the highest possible cognitive level. The items selected for measuring G9 and G10 competencies have very high discrimination indices and with medium difficulty level and items with this close characteristic type. While items selected for measuring G11/12 competencies has also high discrimination indices but with hard difficulty level and items with this close characteristic type. The distribution of the items was balanced; about 50\% of the items measures low order thinking skills and the other 50\% measures higher order thinking skills.

Removing all the items that violated the assumptions of IRT and misfit items in each run is not ideal because developing a test is costly. The item information function is helpful in deciding to remove this kind of items. This method helps to salvage more test items with acceptable item parameters.

In developing a test, IRT is recommended because of its invariant property and primary focuses on item level. In addition, the results of the student’s proficiency level in the subject area that is being assessed are interpreted better because the difficulty level of the test and each item are taken into consideration.

6. References
[1] Lorenzo A A 2016 Development and Validation of Grade 8 Physics Test in K12 Science Curriculum Unpublished Master Thesis De La Salle University
[2] Berame J et. al. 2018 Intl. J. of Biotechnol. and Recent Adv. 1 30-34
[3] Marasigan N V 2019 Intl. J. of Recent Innov. in Acad. Res. 3 262-268
[4] Tabuena A C 2020 Intl J. of Innovative Res. and Devel. 9 16-25
[5] An X and Yung Y 2014 Item Response Theory: What it is and How You Can Use the IRT Procedure to Apply It https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf November 16, 2015
[6] Baker F B 2001 The Basic of Item Response Theory (USA: ERIC Clearinghouse on Assessment and Evaluation)
[7] Gatchalian C M 2008 Item Response Theory: Application and Comparison with Classical Test Theory Using the College Scholastic Aptitude Test Unpublished Master Thesis De La Salle University
[8] Konrad et. al. 2014 Int. in Sch. and Clin. 50 75-85
[9] Haladyna, T M 2004 Developing and Validating Multiple-Choice Test Items (New York: Lawrence Erlbaum Associates)
[10] Chen W H and Thissen D 1997 J. of Educ. and Beh. Stat. 22 265-289
[11] Eleje L I and Esonomonu N P M 2018 Asian J. of Educ and Training 4 18-28
[12] Essen C B, Idaka I E and Metibemu M A 2017 Global J. of Educ. Res. 16 87-94
[13] Orlando M and Thissen D 2000 Appl. Psych. Meas. 24 50–64
[14] Maydeu-Olivares A and Liu Y 2015 Psych. Methods 20 276–292
[15] Adedoyin O O, Nenty H J, and Chilisa B 2008 Educ. Res. and Rev. 3 83-93
[16] Yu C H, Osborn-Popp S, DiGangi S and Jannasch-Pennel A 2007 Practical Assessment, Res., and Eval. 12 1-19
[17] Zakria, N M, Tengku Ismail, T A, Wan Mansor, W N A, and Sulaiman, Z 2019 Intl. J. of Env. Res. and Pub. H. 16 2147