REiD (RESEARCH AND EVALUATION IN EDUCATION)
Vol. 4, No. 2, December 2018

Indexed in:

Research and Evaluation in Education

Vol. 4, No. 2, December 2018

Research and Evaluation in Education

REiD

Vol. 4, No. 2, December 2018

ISSN 2460-6995

Publisher:
PROGRAM PASCASARJANA
UNIVERSITAS NEGERI YOGYAKARTA

REiD (Research and Evaluation in Education) ISSN 2460-6995

Publisher
Program Pascasarjana Universitas Negeri Yogyakarta

Published biannually, in June and December
REiD disseminates articles written based on the results of research focusing on assessment, measurement, and evaluation in various educational areas

Editorial
Department of Educational Research and Evaluation, Graduate School of Yogyakarta State University
3rd Floor Pascasarjana UNY New Building, Colombo Street No. 1, Karangmalang, Yogyakarta 55281
Telephone: 0274 586168 ext. 229 or 0274 550836, Facsimile: 0274 520326
E-mail: reid.ppsuny@uny.ac.id, reid.ppsuny@gmail.com

REiD (Research and Evaluation in Education), 4(2), 2018

Foreword

We are very pleased that REiD (Research and Evaluation in Education) is releasing its eighth edition. We are also very excited that the journal has been attracting papers from the neighbouring country, Malaysia. The variety of submissions from different countries will help the journal in reaching its aim in becoming a global initiative.

REiD (Research and Evaluation in Education) contains and spreads out the results of research which is not limited to the area of common education, but also comprises the results of research in education in a broader coverage, such as language education, cultural education, physics education, mathematics education, and teacher performance, with focuses on assessment and evaluation.

The editorial board expects comments and suggestions for the betterment of the future editions of the journal. Special gratitude goes to the reviewers of the journal for their hard work, contributors for their trust, patience, and timely revisions, and all staffs of the Graduate School of Universitas Negeri Yogyakarta for their assistance in publishing this journal.

Yogyakarta, December 2018

Editor in Chief

TABLE OF CONTENT

REiD (Research and Evaluation in Education), 4(2), 2018, 144-154
Available online at: http://journal.uny.ac.id/index.php/reid

Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT

*1 2 3

Edi Istiyono; Wipsar Sunu Brams Dwandaru; Revnika Faizah 1
Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia 2,3
Department of Physics Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: edi_istiyono@uny.ac.id

Submitted: 04 December 2018 | Revised: 19 December 2018 | Accepted: 20 December 2018

Abstract

Evaluation using computerized adaptive tests (CAT) is an alternative to paper-based tests (PBT). This study was aimed at mapping physics problem-solving skills using PhysProSS-CAT on the basis of the item response theory (IRT). The study was conducted inSleman Regency, Yogyakarta, involving 156 students of Grade XI of senior high school. Sampling was done using stratified random sampling technique. The results of the study show that the PhysProSS-CAT is able to accurately measure physics problem-solving skills. Students' competences in physics problem solving can be mapped as 6% of the very high category, 4% of the high category, 36% of the medium category, 36% of the low category, and 18% of the very low category. This shows that the majority of the students' competences in physics problem solving lies within the categories of medium and low.

Keywords: *assessment, problem-solving skill, CAT*

Introduction

One of the 21st-century learning and innovation skills is the ability related to critical thinking, problem solving, technology, and information (Daryanto & Karim, 2017). Technology is an integral aspect of the development of a nation. The more advanced the cultures of a nation, the more varied and complicated the technology that is used. Problem solving is a cognitive process directed to the attainment of an objective when there is a solution method to solve a problem (Bueno, 2014). Physics learning highly needs problemsolving skills; it is, therefore, necessary to have an evaluation as one of the efforts in elevating the learners' thinking skills.

Nitko and Brookhart (2011, p. 3) define evaluation as a process to obtain information for making decisions concerning the learners, curriculum, program, school, and educational policy. Evaluation instruments used in learning covers tests and non-tests (Nitko & Brookhart, 2011). Test-type instruments can be further grouped into objective tests and non-objective tests. Objective tests can be in the form of multiple-choice, short answers, matching, and objective essays. Non-objective tests can be open essays, work performance or observation, and portfolios or project tasks (Mundilarto, 2010, p. 52). Multiple-choice test items can be used to assess learning more complex outcomes which are concerned with the aspects of recall, understanding, application, analysis, synthesis, and also evaluation (Arifin, 2016, p. 138). The administering of the test can be done in two modes: paperpencil and computer-based test (CBT). The paper-pencil test is paper-based test (PBT) as has been done for long, while CBT is computer-based (Pakpahan, 2016, p. 24).

PBT is based on the assumption that learners with the same level of age and education have the same level of competences. In

ISSN 2460-6995

reality, there is, however, a significant varia tion (Bagus, 2012, pp. 45–46). The PBT model has many shortcomings especially related to deviating behaviors, such as frauds, discussions, sharing of answer keys, or even teachers or schools giving out answers keys with the intention that the teachers or schools are not regarded as failing in the running of education and learning by the society (Balan, Sudarmin, & Kustiono, 2017, p. 37). Further, Retnawati (2014, p. 190) states that Indonesia is a big archipelago consisting tens of provinces. As such, distribution of test packages from the centre to the regions faces many obstacles including, for example, during the national examination (NE). This causes, among others, test administration to be impartial and tests results not valid in that they do not represent the real competences of the students. These limitations of PBT can be overcome by testing using the computer.

Computer-based testing has some advantages, including: there is no need to wait for weeks for testees to receive their scores; scores can be obtained immediately. CBT also provides the facility for giving each testee test items that are pre-arranged to give the testee the freedom to select the next test item (Miller, Linn, & Gronlund, 2009, p. 12). According to Luecht and Sireci (2011), the CBT model can be categorized into: (1) computerized fixed tests (CFT); (2) linear-on-the-fly tests (LOFT); (3) computerized adaptive tests (CAT); (4) stratified computerized adaptive tests (AS); (5) content-constrained CAT with shadow tests; (6) test-based CAT and multistage computerized mastery tests (combined); and (7) computer-adaptive multistage tests.

Each model has its own advantages and disadvantages. CBT gives more advantages than PBT does in that, among others, its scoring system is automatic and it reduces the burdens on the part of the testees (Riley & Carle, 2012). However, CBT is similar to PBT in that it may not be able to measure the testees' abilities accurately since there is still a potential of fraud in its administration. CBT makes the testees respond to all of the items so that there is inefficiency in the use of time.

There are two theories in assessment that have been empirically and technologically developed. These are classical test theory (CTT) and item response theory (IRT). Both CTT and IRT widely represent two different frames of assessment. In views of the CTT, scoring of a test is done partially, using the steps that need to be taken in answering a test item correctly. Scoring is conducted step by step, each testee's item score is obtained by summing up the score in each step, and achievement is estimated from raw scores. This scoring model may not be appropriate since the difficulty level of each step is not taken into consideration (Istiyono, Mardapi, & Suparno, 2014, p. 4). In the item level, the CTT model is relatively simple; CTT does not demand a complex theoretical model to relate a testee's success in responding to a test item. On the contrary, CTT collectively considers a group of testees for a particular item. IRT has been developed and important to complement CTT in the design, interpretation, and evaluation of a test or examination. IRT has a strong mathematical basis and relies on a complex algorithm more efficiently calculated on the computer (Adedoyin, 2010, p. 108). IRT supports the use of the computer in educational testing. IRT can be used to provide any item saved in the computer independently, so that the computer select a test from item banks, manage the procedure of the item administering, or design a model for a new computer-based item-response test (Masters & Keeves, 1999, p. 139; van der Linden & Glas, 2003). Thus, a test which uses CAT is highly suitable with the item response theory (IRT).

Hambleton, Swaminathan, and Rogers (1991, p. 9) propose three assumptions underlying the item response theory, including: (1) the chance for answering an item is not dependent on that for another item (local independence), (2) an item measures one competence dimension (unidimensional), and (3) the response pattern of each item can be represented in an item characteristic curve. The

weaknesses of the classical theory are tackled up by these three assumptions. Hambleton et al. (1991) identify four limitations of the classical theory. First, item statistics such as difficulty levels and discriminating powers are restricted by specific observed samples that are obtained; i.e. they depend on the group and test. Second, reliability is defined by paralleltest concepts, which are difficult to realize in practice. This is due to the fact that individuals can never be the same in the second test since they may forget, earn new competences, or have different motivation and anxiety levels. Third, standard errors of measurement are assumed to be the same for all subject matters and variabilities in errors are not being considered. Fourth, the classical theory reflects focus on the test-level information to put itemlevel information aside. Test-level information is an additive process, that is, the amount of information across the item, and item-level information is the information only for certain items. These limitations show that the classical theory deals with individual score totals and not each testee's competences in the individual level.

A CAT is based on the item response theory. Hambleton and Swaminathan (1985, p. 48), state that there are three types of scoring systems: dichotomous, polytomous, and continuous. Of the three, dichotomous system is the most used in the educational evaluation. The models that can be used for the dichotomous data are latent linearity, perfect scale, latent distance, Ogive one-two-three normal parameter, one-two-three logistic parameter, and four logistic parameter (Barton & Lord, 1981; Guttman, 1944; Lazarsfeld & Henry, 1968; Lord, 1952). The dichotomous model is only suitable for items with twocategory scores such as true/false. For items with more than two score categories, the polytomous system is used.

The polytomous scoring system has a number of models, such as nominal response, graded response, partial credit model, and others (Bock, 1972; Geoff N. Masters, 1982; Samejima, 1969). The partial credit model (PCM) has been developed in order to analyze the test items which require multiple-step responses, wherein the items follow the partial credit model patterns so that individuals with higher competences will score higher than those who have lower competences (Istiyono, 2017, p. 2). Therefore, it is reasonable that the partial credit model is used for multiplechoice tests.

A CAT is based on the principles that items must be selected by a consideration that they must measure the testees' competences. Generally, an item is selected in that it gives the most information to estimate the testee's competences. Then, based on the true/false response pattern, the competence level is supposed to return and the item is selected on the basis of the newly estimated competence. These processes are then continued up to a certain precision of the obtained testee's competences (Hambleton & Zaal, 1991). Based on the discussion of these facts, a need is felt on the development of a test that will measure the testees' competences in problem solving. The computerized adaptive test (CAT) has been developed as a CBT alternative to examine PBT tests and provide better tests items and shorter tests in accordance with each test. CAT is a testing system which is more advanced than CBT (Hadi, 2013, p. 12). In accordance with Suyoso, Istiyono, and Subroto (2017), computer-based evaluation is needed more and can help teachers in conducting an evaluation in their subject-matter teaching. In the 21st century, more is emphasized on the higher-order thinking cognitive domain such as HOTS Bloomian, HOTS Marzonian, critical thinking, creative thinking and problem solving (Brookhart, 2010; Heong et al., 2011; Schraw & Robinson, 2011). Testees interact directly with the computer containing the test items of the subject matter. They work on answering test items through the computer as they do in PBT through writing. The number of items is the same that in PBT and item characteristics do not function as they do in CAT (Pakpahan, 2016, pp. 26–27).

The use of CAT does not require items in a great number since the computer is able to give the items in accordance with the testees' competence levels. On the contrary, PBT, which is developed by classical theories, needs

items in a great number since it needs to measure the testees' optimum competences repeatedly (Gregory, 2014). According to Weiss (2004, p. 82), CAT is a technology that is viable to have the potentials to give a better assessment, in smaller testing time, for various application in counseling and education. In these two fields, there are needs to measure individuals' changes. There are so many vari eties in the evaluation applications, and one that is able to make use of the superiority of assessment applications which are good and efficient is that which applies the CAT technologies.

Method

The study was conducted in State Senior High School in Sleman Regency, Yogyakarta Province, during the even semester of the 2017-2018 academic year. The subjects of the study were 156 students of the Physics Department selected by a stratified random sampling technique taking the higher, medium, and lower groups into consideration based on the students' scores of the National Examination in Physics. The size of the sample was determined from the population using the 1-PL formula that ended with 150 to 250 students (Linacre, 2006).

Data collection was conducted by a test that was used to map students' competences in problem solving in the field of physics. The research participants were asked to take the PhysProSS-CAT test which was the product of this research development.

The PhysProSS-CAT consists of items that have undergone development in the forms of multiple-choice items with reasons. The mate-rial is related to the balance of solid things, elasticity and Hooke law, static fluid, dynamic fluid, and temperature and calorie. The development of the instrument was based on the Curriculum 2013 which had been revised on the aspects and sub-aspects of problem-solving skills (Ministry of Education and Culture, 2013). The aspects included identification, planning, implementation, and evaluation. The sub-aspects included identifying, differentiating, planning, formulating, sequencing, connecting, applying, checking, and criticizing. The test was developed into four sets of test items, 180 in total with nine anchor items.

The test items had the characteristics that fulfilled the requirements for testing. These requirements were as follows: (a) Based on the results of the content validation by the evaluation experts, the test was content-wise valid with Aiken's V value of 0.97; (b) Based on the empirical evidence, the test had a fit with the Partial Credit Model (PCM) polyatomic data with four categories with a mean score and INFIT MNSQ standard deviation of 1.00±0.25; (c) Based on the Cronbach Alpha reliability estimation values, all items were regarded as reliable at the measure of 0.93; (d) Based on the levels of difficulty, the test was regarded as good with a range of 1.23 to 1.50; and (e) On the information function and SEM, the test was stated to be able to estimate competences on the range between -2 and 1.6.

The scoring of the test used the partial credit model (PCM) technique which was a development of the 1-PL model and was of the Rash family. Meanwhile, the results of the physics problem-solving test used the computerized adaptive test (CAT) categorized in the form of levels adapted from (Azwar, 2010). The categories are shown in Table 1.

Table 1. Intervals of students' problemsolving skills

No	Skill Interval	Level
1	$Mi + 1.5SBi < \theta$	VeryHigh
2	$Mi + 0.5SBi < \theta \leq Mi + 1.5SBi$	High
3	Mi + 0.5$SBi < \theta \leq Mi$ – 0.5SBi	
	Medium 4	Mi – 1.5$SBi < \theta \leq Mi$ – 0.5SBi Low
5	$\theta < Mi - 1.5Sbi$	Very Low

Findings and Discussion

Findings

The level of students' competences in problem solving is directly in comparison with the level of item difficulty. The higher the

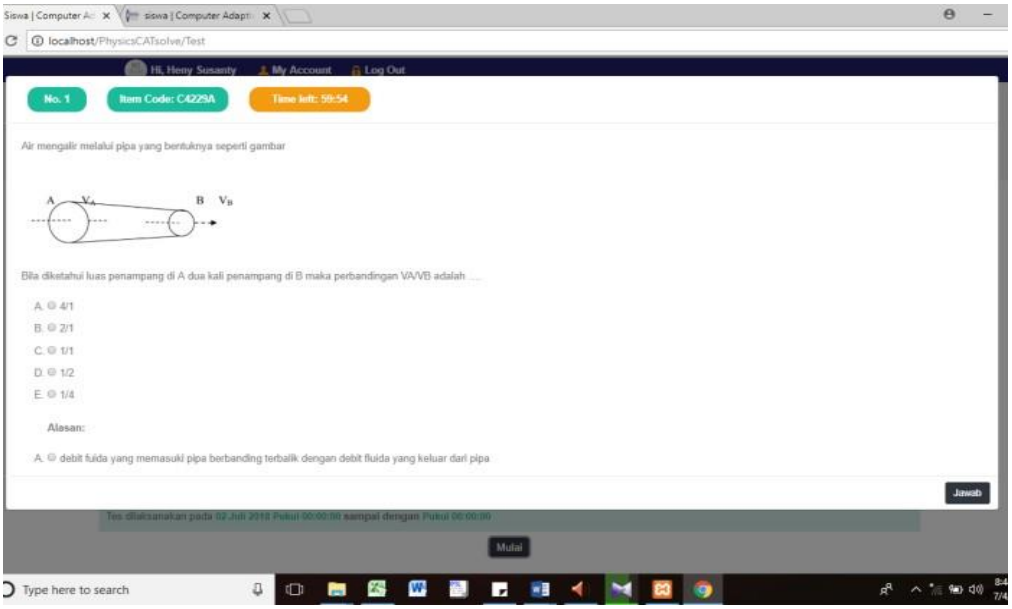


Figure 1. Test item appearance

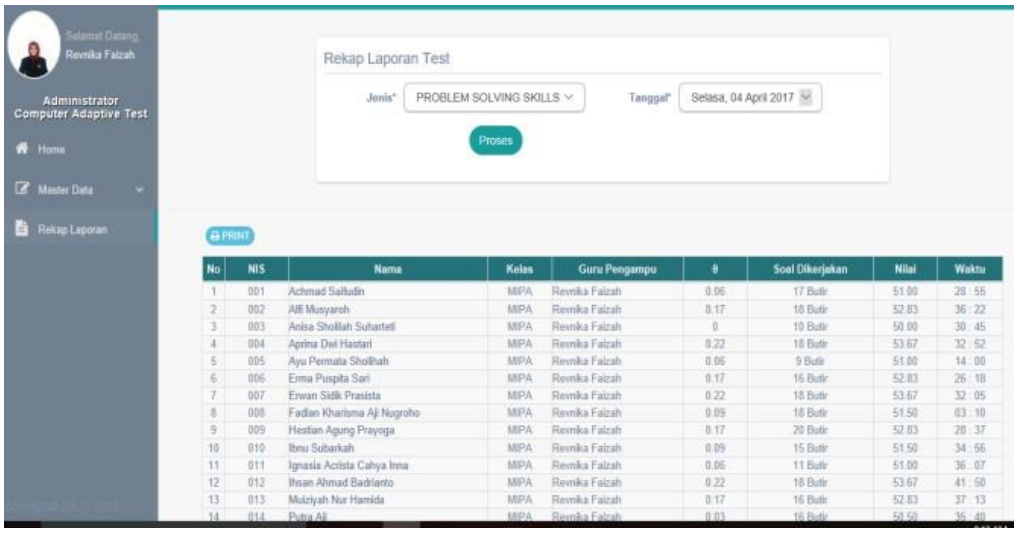


Figure 2. Recapitulation report appearance

students' theta values, the more difficult the items; the lower the theta, the lower the item difficulty. Students respond to an item whose difficulty level is comparable with their competence level. The first item is one with a medium level of difficulty. If the students answer it correctly, the test will give them a more difficult item; and if they get it wrong, the test will give them a less difficult item. The exposed items have been fitted with the problem-solving aspects, namely identification, planning, implementation, and evaluation. The presentation of an item using CAT can be seen in Figure 1.

In Figure 1, a PhysProSS-CAT test item can be seen in the multiple-choice format with reasons. The testees are asked to select the correct answer and give the reasons for selecting it. After a testee completes the test on

the CAT, a recapitulation report from the computer will appear on the screen, as presented in Figure 2.

The recapitulation report can be immediately seen by the administrator, teacher, and student. The administrator can see all the reports of all the test takers. The teacher can see only the reports of his students. The report is in the form of theta scores representing the students' competences. The students' competence level (θ) is categorized into very high, high, medium, low, or very low in a five-level scale (Azwar, 2010, p. 63) as can be seen in Table 2.

Table 2. Problem-solving skill scale conversion

No	Interval	Competence Level
1	$0.27 \leq \theta$	Very High
2	$\theta \leq 0{,}27$	High
3	$\theta \leq 0.21$	Medium
4	$\theta \leq 0.16$	Low
5	$\theta \leq 0.10$	Very Low

In Table 3 and Figure 3, of the 156 stu dents taking the CAT test, ten are in the very high category, six are in the high, 56 are in the medium, 56 are in the low, and 28 are in the very low. In percentages, 6% of the students are in the very high category, 4% in the high, 36 % in the medium, 36% in the low, and 18% in the very low. It means that most students' competence levels are in the medium and low categories.

Table 3. Mapping results of competence levels in three state senior high schools

No	Competence Level	Number of Students	Percentage (%)
1	Very High	10	6.41
2	High	6	3.85
3	Medium	56	35.90
4	Low	56	35.90
5	Very Low	28	17.95

Figure 3. Mapping results of competence levels in three state senior high schools

Mapping is done on the three schools based on the scores which are obtained from the national examination (NE) in Physics, categorized as: high, medium, and low. The results of the mapping are presented in Table 4, Table 5, and Table 6.

Table 4. Mapping of problem-solving competence levels in Senior High School A

No	Competence Category	Number of Students	Percentage (%)
1	Very High	5	7.81
2	High	4	6.25
3	Medium	23	35.94
4	Low	19	29.69
5	Very Low	13	20.31
	Total	64	100.00

Table 5. Mapping of problem-solving competence levels in Senior High School B

No	Competence Category	Number of Students	Percentage (%)
1	Very High	2	3.33
2	High	2	3.33
3	Medium	25	41.67
4	Low	22	36.67
5	Very Low	9	15.00
	Total	Total	100.00

Table 6. Mapping of problem-solving competence levels in Senior High School C

No	Competence Category	Number of Students	Percentage (%)
1	Very High	3	9.38
2	High	2	6.25
3	Medium	9	28.13
4	Low	15	46.88
5	Very Low	3	9.38
	Total	32	100.00

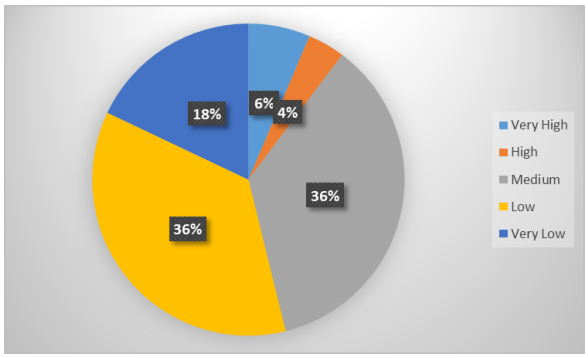


Figure 4. Mapping of problem-solving competence levels in Senior High School A

Shown in Figure 4, in State Senior High School A, of the 64 students, 8% are in the very high category, 6% very high, 36 % medium, 30% low, and 20% very low. It indicates that most students' competence in this school are in the 'medium' category.

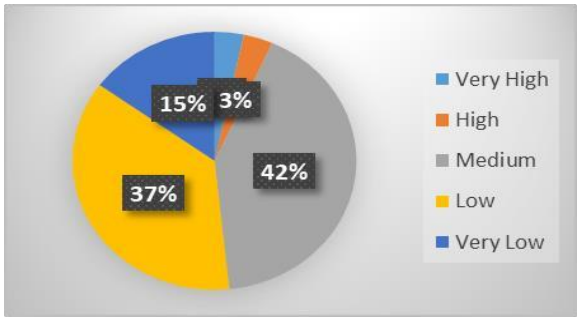


Figure 5. Mapping of problem-solving competence levels in Senior High School B
Based on Figure 5, in State Senior High School B, of the 60 students participating in the study, 3% are in the very high category, 3% very high, 42% medium, 37% low, and 15% very low. It indicates that most students in this school are in the 'medium' category.

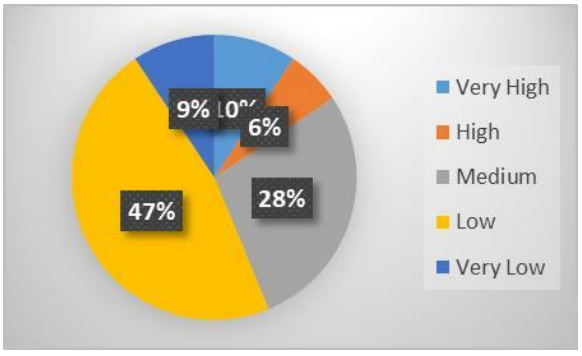
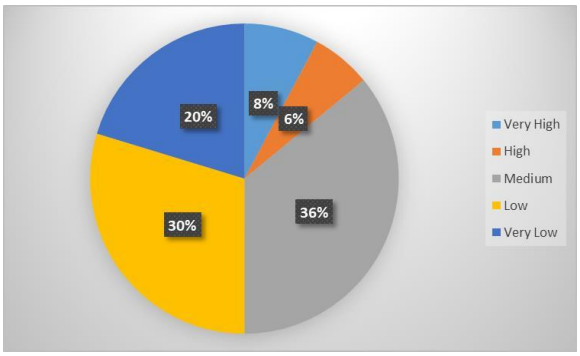


Figure 6. Mapping of problem-solving competence levels in senior high school C



As seen from Figure 6, in State Senior High School C, 32 students participated in the study and 10% of them are in the very high category, 6% very high, 28% medium, 47% low, and 9% very low. It indicates that most students in this school are in the 'low' category.

Discussions

Based on the findings of the research, it is clear that the PhysProSS-CAT test has been quite well and accurately able to map students' competences in Physics problem solving. The CAT-based instrument has been able to select the items in accordance with the students' competence levels. In this case, students of School A who are high in the national examination are dominantly in the medium category, but have the highest score in the problemsolving test. In the B school, which is medium in the national examination, the students are dominantly at the medium and low categories. Meanwhile, in School C, with a low level of national examination results, the students are dominantly low in their problem solving competence. This means that mapping has been done well in matching test items with students' levels of competence.

The results of the overall mapping of the 156 students participating in the study show that many of the students are in the medium category. This can be traced from the factors of students' motivation, instructional processes, and evaluation practices. In this relation, only the evaluation factor will be discussed further. Accurate evaluation will be able to support students to learn using higherorder thinking (Istiyono et al., 2014, p. 2). The learning processes and evaluation are supposed to deal with higher-order thinking, including problem solving, in order that the students' skills in problem solving improve. In time, the need is felt to develop evaluation that will be able to measure these students' skills. Ultimately, this will help in realizing students' learning achievements.

As shown in Figure 7, the PhysProSS CAT test results have categorized students' competences in Physics problem solving into very high, high, medium, low, and very low. Further, the results give accurate information about students' problem-solving skills. The students of State Senior High School A with high national examination scores have the most students with medium problem-solving skills, School B with medium national examination scores has many students with medium problem-solving skills, and School C with low national examination scores has most students with low problem-solving skills.

Meanwhile, Figure 8 presents the recapitulation report of the test results. It consists of scores, test items answered, and time. In the time of the test administration, most students completed 18 to 25 test items, in 35 to 50 minutes out of the total items of 154. The minimum items to be completed were nine items, and the shortest time was 14 minutes. The maximum items completed were 25 and the longest time was 58 minutes. Students did not need to complete all the items but only those within their competences. This is in line with Gregory (2014) stating that CAT testing does not need too many items since, in the computer-based testing, the computer provides test items that are within the range of the testee's competences.

Departing from the weaknesses of the paper-based testing (PBT) mode, in which all testees take all items without considering their skill differences, the computer-based testing (CBT), on the other hand, is designed using the adaptive mode. In this mode, next items are given on the basis of the testee's competence in completing the previous items (Istiyono, 2013). It is, therefore, reasonable to use the computerized adaptive test (CAT) as an alternative technique for testing since it gives a better estimation result and using a shorter test to be adjusted to the testee's competence. Further, testees do not have to answer all

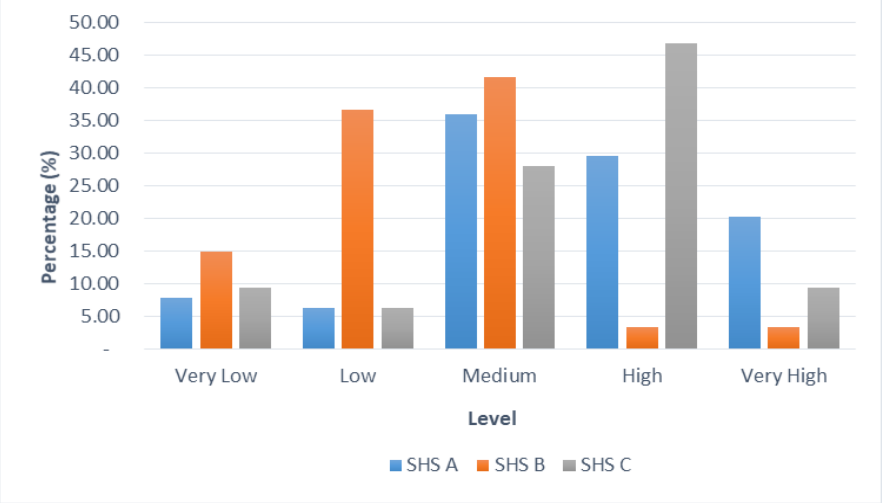


Figure 7. Mapping of the students' problem-solving competences in three schools

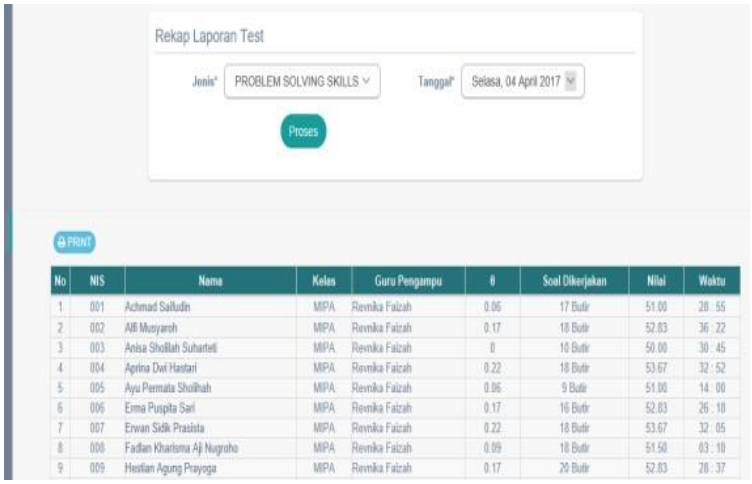
questions, and this saves testing time. In accordance with Huang, Chen, and Wang (2012), the superiority of the CAT over the PBT is that the CAT is able to achieve the same precision with fewer items and shorter time. In CAT, the testee needs only to click on the correct answers until the computer finds and determines his most accurate estimate of his competences to terminate the test and gives his score. CAT is most suitable for such tests for selection and one of a large scale.

The use of PhysProSS-CAT can minimize frauds since testees do different items and have different numbers of items to complete the test; the CAT program gives different items to testees in accordance with their levels of competences. Safety and confidentiality of the items are guarded. On its turn, results of the testing will be reliable. In PBT and CBT testing, chances are abound for frauds to take place for the opposite reasons that testees take the same test with relatively the same items.

The PhysProSS-CAT can do its testing functions safely, fast, and accurately showing the accurate competences of the testees. For this reason, the test helps much in competence mapping for various purposes. The immediate issuance of the test results helps the teacher map the students' competences in a short time. The teacher can also immediately evaluate and plan for further programs.

In line with the opinions proposed by van der Linden and Glas (2003), a number of reasons for switching to the CAT type are: (1) CAT makes it possible for testees to schedule their own testing in accordance with their preferences; (2) Testing is administered in a comfortable atmosphere with fewer people around than there are in conventional paperpencil testing; (3) CAT processes the data and gives out the results fast; and (4) Test items and materials are more varied in levels and sizes.

It is possible for teachers to select a test from a variety of choices but testing must be done in accordance with the needs and situations. In a school with adequate facilities for computers, the CAT type testing is more preferable. For the assessment of higher-thinking skills, more specifically, the CAT model is more appropriate since it measures competences accurately and efficiently and saves energy and time of the administration. This is supported by Jiao, Macready, Liu, and Cho (2012) stating that computerized captive testing achieves higher accuracy of the measurement and provides efficient administering of the assessment. In view of the superiority of PhysProSS-CAT, it is suitable for testing individuals' competences in such testing for selection and the final examination. The test saves time, and energy and minimizes frauds.



No	NIS	Nama	Kelas	Guru Pengampu	θ	Soal Dikerjakan	Nilai	Waktu
1	001	Achmad Saifudin	MIPA	Revnika Faizah	0.06	17 Butir	51.00	28 : 55
2	002	Alfi Musyaroh	MIPA	Revnika Faizah	0.17	16 Butir	52.83	36 : 22
3	003	Anisa Sholilah Suhartati	MIPA	Revnika Faizah	0	10 Butir	50.00	30 : 45
4	004	Aprina Dwi Hastari	MIPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 52
5	005	Ayu Permata Sholihah	MIPA	Revnika Faizah	0.06	9 Butir	51.00	14 : 00
6	006	Erma Puspita Sari	MIPA	Revnika Faizah	0.17	16 Butir	52.83	26 : 18
7	007	Erwan Sidik Prasista	MIPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 05
8	008	Fadlan Kharisma Aji Nugroho	MIPA	Revnika Faizah	0.09	18 Butir	51.50	03 : 10
9	009	Hestian Agung Prayoga	MIPA	Revnika Faizah	0.17	20 Butir	52.83	28 : 37

Figure 8. Recapitulation report of the PhysProSS-CAT test results

Conclusion and Suggestions

Conclusion

Based on the results of the study, it can be shown that the PhysProSS-CAT is able to accurately map the students' competences in problem solving in the physics field. In percentages, students' competences can be rated as very high (6%), high (4%), medium (36%), low (36%), and very low (18%). This means that the majority of the students' competences are within the categories of medium and low. On the average, of the total 154 items provided in the test, students complete between 18 and 25 test items in a time range of 35 to 50 minutes. Meanwhile, the minimum number of items responded is 9 and the time needed is 14 minutes; and the maximum number is 25 and the maximum time 58 minutes. Therefore, PhysProSS-CATis able to map problemsolving competences accurately, efficiently, and saves time and energy.

Suggestions

In the administering of CATs, including PhysProSS-CAT, it is recommended that administrators provide items with difficulty levels that are more normal in distribution. In relation to the technical facilities, it is suggested that administrators use adequate numbers of items to anticipate troubles in the computer webs since testees access the same items in the same time.

References

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Educational Sciences*, *2*(2), 107–113. https://doi.org/10.1080/09751122.2010.11889987

Arifin, Z. (2016). *Evaluasi pembelajaran: Prinsip, teknik, dan prosedur* (8th ed.). Jakarta: Remaja Rosdakarya.

Azwar, S. (2010). *Metode penelitian*. Yogyakarta: Pustaka Pelajar.

Bagus, H. C. (2012). The national exam administration by using computerized adaptive testing (CAT) model. *Jurnal Pendidikan Dan Kebudayaan*, *18*(1), 45–53.

Balan, Y. A., Sudarmin, S., & Kustiono, K. (2017). Pengembangan model computer-based test (CBT) berbasis Adobe Flash untuk sekolah menengah kejuruan. *Innovative Journal of Curriculum and Educational Technology*, *6*(1), 36–44. https://doi.org/10.1186/2229-0443-13-60

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic itemresponse model. ETS Research Report Series* (Vol. 1981). Princeton, NJ: John Wiley & Sons. https://doi.org/10.1002/j.233 3-8504.1981.tb01255.x

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF022 91411

Brookhart, S. M. (2010). *How to assess higherorder thinking skills in your classroom*. Alexandria, VA: ASCD.

Bueno, P. M. (2014). Assessment of achievement in problem-solving skills in a General Chemistry course. *Journal of Technology and Science Education*, *4*(4), 260–269. https://doi.org/10.3926/jotse .100

Daryanto, & Karim, S. (2017). *Pembelajaran abad 21*. Yogyakarta: Gava Media.

Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7th ed.). Wheaton, IL: Pearson.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150. https://doi.org/ 10.2307/2086306

Hadi, H. (2013). *Pengembangan Computerized Adaptive Test berbasis web*. Yogyakarta: Aswaja Pressindo.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K., & Zaal, J. N. (1991). *Advances in educational and psychological testing*. Boston, MA: Kluwer Academic.

Heong, Y. M., Othman, W. B., Yunos, J. B. M., Kiong, T. T., Hassan, R. Bin, & Mohamad, M. M. B. (2011). The level of Marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity*, *1*(2), 121–125.

Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, *36*(8), 689–706. https: //doi.org/10.1177/01466216124 59552

Istiyono, E. (2013). *Pengembangan instrumen untuk mengukur kemampuan berpikir tingkat tinggi dalam mata pelajaran Fisika di SMA*. Yogyakarta: Department of Physics Education, Universitas Negeri Yogyakarta.

Istiyono, E. (2017). The analysis of senior high school students' physics HOTS in Bantul District measured using PhysReMChoTHOTS. In *AIP Conference Proceedings* (Vol. 1868, p. 070008). AIP Publishing LLC. https://doi.org/10.106 3/1.4995184

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PhysTHOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v1 8i1.2120

Jiao, H., Macready, G., Liu, J., & Cho, Y. (2012). A mixture Rasch model–based computerized adaptive test for latent class identification. *Applied Psychological Measurement*, *36*(6), 469–493. https://doi .org/10.1177/0146621612450068

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin.

Linacre, J. M. (2006). *WINSTEP: Rasch-model computer programs*. Chicago, IL: Winstep.com.

Lord, F. (1952). *A theory of test scores*. Richmond, VA: Psychometric Corporation.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing*. New York, NY: The College Board.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Masters, G. N., & Keeves, J. P. (1999). *Advances in measurement in educational research and assessment* (1st ed.). Amsterdam: Pergamon.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). The role of measurement and assessment in teaching. In *Measurement and assessment in teaching* (10th ed., pp. 29–31). Upper Saddle River, NJ: Pearson Education.

Ministry of Education and Culture. (2013). *Pengembangan kurikulum 2013*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Mundilarto. (2010). *Penilaian hasil belajar Fisika*. Yogyakarta: Pusat Pengembangan Instruksional Sains (P2IS) Jurdik Fisika FPMIPA UNY.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson Education.

Pakpahan, R. (2016). Model ujian nasional berbasis komputer: Manfaat dan tantangan. *Jurnal Pendidikan Dan Kebudayaan*, *1*(1), 19–35. https://doi.org/10.24832/jpnk.v1i1.225

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Riley, B. B., & Carle, A. C. (2012). Comparison of two Bayesian methods

to detect mode effects between paperbased and computerized adaptive assessments: A preliminary Monte Carlo study. *BMC Medical Research Methodology*, *12*, 124. https://doi.org/ 10.1186/1471-2288-12-124

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. In *Psychometrika Monograph, No. 17*. Richmond, VA: Psychometric Society.

Schraw, G. J., & Robinson, D. H. (2011). *Assessment of higher order thinking skills: Current perspectives on cognition, learning, and instruction*. Charlotte, NC: Information Age Publishing.

Suyoso, S., Istiyono, E., & Subroto, S. (2017). Pengembangan instrumen asesmen pengetahuan fisika berbasis komputer untuk meningkatkan kesiapan peserta didik dalam menghadapi ujian nasional berbasis komputer. *Jurnal Pendidikan Matematika Dan Sains*, *5*(1), 89–97. https://doi.org/10.21831/jpms.v5i1.12461

van der Linden, W. J., & Glas, C. A. W. (2003). *Computerized adaptive testing: Theory and practice*. London: Kluwer Academic Publisher.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751