A novel, adaptive, multimedia-based assessment test for Generation Z in senior vocational high schools

Soeryanto, Rachmad S. Hidayatullah & Wahyu D. Kurniawan

State University of Surabaya
Surabaya, Indonesia

ABSTRACT: Evaluation is a way to gather information about learning activity outcomes and achievements. It requires a valid and reliable instrument to collect the required data in measurable values. The research outlined in this article aimed to design a suitable measuring tool for Generation Z (Gen Z) in vocational high schools. The 4-D model was used as the development method. There were 39 items that were validated by experts and tested on 188 students of class XI TBSM in Indonesia. Data were analysed using the Rasch model. The test results show the instrument's reliability (kr-20) is 0.93. The separation item value is 6.88, and the INFIT-MNSQ item is < 1.19. All items are fit, with nine items needing refinement, and the quality of the items in the instrument is classified as special. The developed test using a video is an innovation in the field of educational evaluation. The effective duration of using the video is 0.29 seconds. This video-based test is very suitable for the 5.0 era, where virtual videos support daily activities especially of Gen Z.

INTRODUCTION

Evaluation of learning is an important part of education and is usually carried out through structured examinations [1]. This activity attempts to improve the quality of education by teachers and authorities [2]. Other studies suggest that learning evaluation is an activity to obtain data used to analyse the achievement of indicators of learning objectives [3] to monitor students' progress through assessment and evaluation. Assessment and evaluation can guide students to determine changes in learning outcomes that are more complex [4]. This activity can facilitate the development of teaching materials, media, methods and all learning activities [5].

The evaluation results are influenced by the design of the questions. In compiling questions building the right construction of questions is a challenge. A teacher needs to do an item analysis to provide feedback on the validity value of the questions compiled [6]. Before the questions are given to students, the question maker needs to analyse the items to obtain quality questions with appropriate validity and reliability values [7]. A good test consists of operational items that are of good quality and can be an accurate indicator of students' knowledge, skills and abilities [8].

Well-crafted multiple-choice questions are proven methods to test students' understanding of a topic [9]. Multiple-choice questions can be given to students off-line or on-line. Students can benefit cognitively, metacognitively, emotionally and socially while answering multiple-choice question tests.

Student character data is very influential in the success of a learning evaluation. A test instrument is said to be good if it can provide relevant information on the skills of the students being tested [10]. Vocational high school (in Indonesian abbreviated as SMK) students born in 2010 are currently the majority. They are Generation Z (Gen Z) and Generation Alpha (Gen Alpha). Gen Z was born between 1995 and 2010, and Gen Alpha between 2011 and now. The learning character of Gen Z and Gen Alpha has learning patterns that always dominate their activities with technology [11].

As Gen Z was born between 1995 and 2010, it means that the oldest in this group are in their 20s and comprise the current college population. Gen Z, also known as iGeneration, are smartphone natives who are mobile technology savvy. They love to see animations and watch videos. Previous research has shown Gen Z are *digital natives* who access digital devices from an early age, integrate with technology [9] and are more technically connected in learning and communication styles when compared to Millennials [12].

The Rasch model's choice depends on the instrument's response scale. Rasch analysis is used when a set of questionnaire items (or items from a given scale) are intended to be summed together to give a total score (which may include several total subscales, as well as an overall score) [13]. The Rasch model can be dichotomous or polytomous (including rating scales and partial credit models) [14]. Dichotomous Rasch models are used with two response option instruments, item response theory (IRT) as a rigorous validation tool, and applying it to concept tests for education as it demonstrates the model fit and acceptable items. The discrimination parameter indicates that the instrument can effectively discriminate between students with different ability levels [15]. High-quality items will be maintained, while low-quality ones will be discarded or replaced with better questions.

On-line teaching is expected to grow in the future. It will likely lead to an increase in electronic assessment, which is known to make it easier for students to access examination answers and/or receive assistance without permission from others. Students reported cheating more frequently on-line than during on-site examinations [16].

Most of the current item developments concern higher-order thinking skills (HOTS) questions. Usually, researchers develop HOTS questions using application assistance, such as applying HOTS question items to quizzes. The student HOTS question application averages 96%, meaning that students respond very positively when using the quiz application. The items developed are multiple-choice questions [17]. The design stage begins, after the researchers know the problems from the definition stage, then they can design a HOTS-based multiple-choice test scoring instrument [17].

The development of the question items involves using technology in this process. Some designers have used applications such as, Quisis, Kahoot, Google Forms, CAT, and so on, but the majority of the content in these applications is still in the form of text and images.

The current development outlined in this article is based on previous researchers' work regarding the item development, but the questions presented in the form of text and images, are developed into items that use text and video. Students will work on item questions that contain text and video. The video will appear if the writing uploaded with Google Drive is clicked. The video will appear containing instructions, component videos, and videos for troubleshooting questions or what is usually known about HOTS.

This study aimed to find out the results of a Gen Z test using Google Forms for on-line test media and design a draft test combining text and video suitable for Gen Z vocational students in a motorcycle business engineering programme.

Based on the conditions above, this study proposed a new multimedia-based and adaptive assessment method for Gen Z students in vocational high schools in a motorcycle business engineering programme. The test instrument that the researchers developed is very suitable for high school students because this test instrument contains questions that use video as a medium to trigger students HOTS. The use of video on the item questions will stimulate students to observe the information conveyed in the next video. From these observations, students are invited to think quickly to find answers to these items. The videos in the items are also taken from the results of the documentation of motorcycle components, troubleshooting from motorbikes in motion, in other words, the material used as test material for the items is real and presents everyday cases in actual life. Hence, it is expected that by using live recordings of motorcycles, students can connect the knowledge in their memory directly and quickly with the real conditions in the field.

RESEARCH METHODOLOGY

Questions were developed using the 4-D model proposed by Thiagarajan, D. Semmel and M. Semmel and used in other works [4]. The define stage includes determining: the test objectives, competencies to be tested, test material and test distribution techniques. The design stage includes: preparing test grids, writing items, making test videos and designing test distribution. The development stage includes: validating items, repairing items and assembling items. The dissemination stage includes determining the test subject and conducting the trial.

Table 1: Descriptive data of respondents.

Variable	Information
Number of subjects	188 students of class XI TBSM
Gender	All male
Age	17-18 years old
Birth year	2005-2008
Location	Bangkalan Regency

This study used three SMK test sites in Bangkalan Regency with 188 XI TBSM class students. The instrument was validated by three validators, one language expert validator, one social expert validator and one content expert validator. The test material aims to measure basic competence in maintaining periodic gasoline injection systems of motorbikes with performance indicators, including students' ability to:

- explain the working principle of the control correctly without looking at the reference according to the specified time; explain the working principle of the actuator correctly according to the specified time;
- explain the working principle of sensors correctly according to the specified time;

- translate the blinking of the malfunction indicator lamp light into numbers correctly according to the specified time;
- explain the stages of carrying out maintenance on injection system components correctly according to the standard operating procedure within the specified time;
- describe equipment used in carrying out maintenance of the injection system correctly according to the specified time.

The research instruments in this study were: one set of multiple-choice tests with HOTS question types complete with answer keys. Instruments in the form of questions given to students using Google Forms have gone through the stages of validation analysis, difficulty level reliability and discriminating power tests. In the instrument testing stage, the researchers used Winstep software assistance and conducted analysis using the Rasch model [18]. Rasch analysis has several advantages, such as fulfilling the fundamental measurement requirements of converting raw data into linear interval scales (logits), enabling researchers to investigate student performance and item difficulty using person-item maps, and being a psychometric technique developed to increase the accuracy of measurements where researchers can build instruments and monitor instrument quality [19]. Testing the reliability of measuring instruments in this study was carried out with an internal consistency approach using the Cronbach's alpha formula. A summary description of the research procedure is presented in Figure 1.

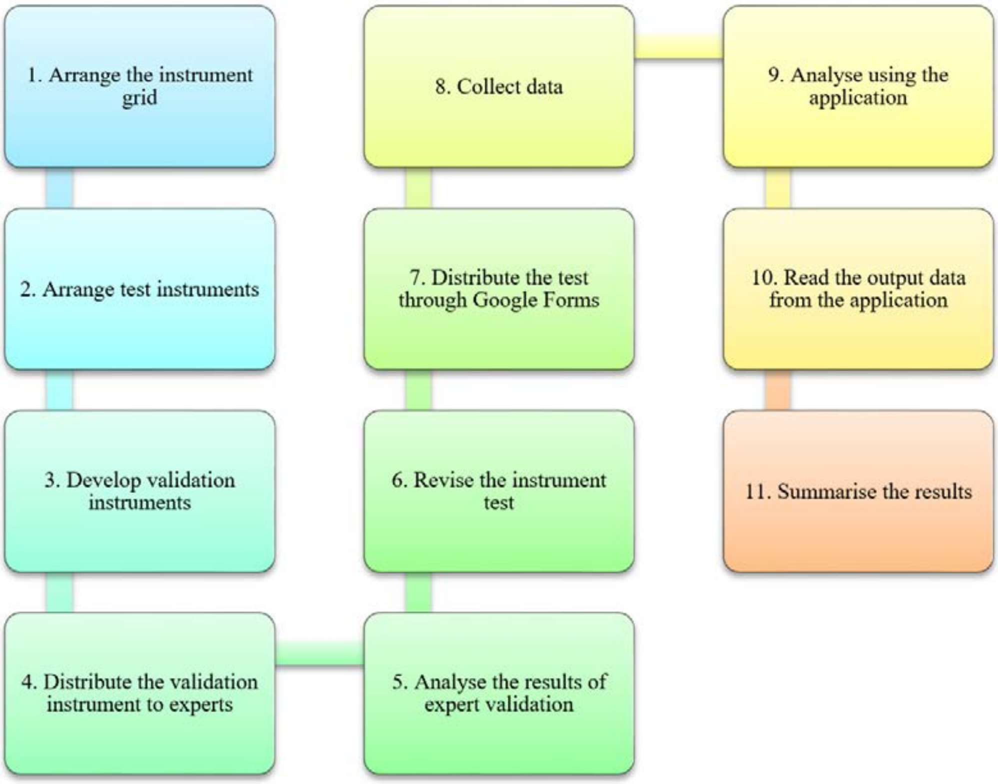

1. Arrange the instrument grid
2. Arrange test instruments
3. Develop validation instruments
4. Distribute the validation instrument to experts
5. Analyse the results of expert validation
6. Revise the instrument test
7. Distribute the test through Google Forms
8. Collect data
9. Analyse using the application
10. Read the output data from the application
11. Summarise the results

Figure 1: Research procedure.

RESULTS

Table 1 is a summary of 188 measured persons, and shows that the logit person value is 2.03, which means the average score for all students when working on examination questions.

Table 1: Summary of 188 measured (extreme and non-extreme) persons.

	Total	Count	Measure	Model	Infit		Outfit			
	Score			Error	MNSQ	ZSTD	MNSQ	ZSTD		
Mean	28.1	39	2.03	0.66						
SD	8	0	2.2	0.44						
Max.	39	39	6.24	1.84						
Min.	11	39	-1.61	0.41	0.36	-0.1	0.24	-2.5		
Real	RMSE	0.81	True	SD	2.04	Separation	2.51	Person	Reliability	0.86
Model	RMSE	0.8	True	SD	2.05	Separation	0.58	Person	Reliability	0.87
SE	OF	Person	Mean	-	0.16					

Person raw score-to-measure correlation = 0.96
Cronbach's alpha (kr-20) person's raw score *test* reliability = 0.93

Table 2 is a summary of 39 measured items.

Table 2: Summary of 39 measured (non-extreme) items.

	Total	Model	Measure	Error	Infit		Outfit			
	Score	Count			MNSQ	ZSTD	MNSQ	ZSTD		
Mean	135.4	188	0	0.25	0.96	-0.9	1.66	0.1		
SD	40	0	1.82	0.05	0.45	3.4	1.97	3.2		
Max.	180	188	3.32	0.37	3.12	9.9	9.87	9.9		
Min.	61	188	-2.46	0.2	0.36	-7.2	0.27	-5.5		
Real	RMSE	0.26	True	SD	1.8	Separation	6.88	Item	Reliability	0.98
Model	RMSE	0.25	True	SD	1.8	Separation	7.2	Item	Reliability	0.98
SE	OF	Item	Mean	-	0.29					

Table 2 shows the item logit value of 0.00, and the logit person value is greater than the logit item value, meaning that the tendency for students' abilities is higher than the difficulty level of the question. The mean outfit MNSQ value is 1.66, ZSTD value is 0.1, standard square outfit (MNSQ) 0.5 < MNSQ < 1.5, z-standard outfit (ZSTD) -2.0 < ZSTD < +2.0, when compared to the standard value, the value of square outfit (MNSQ) and z-standard outfit according to the standard, that means the average question is said to be fit.

Table 1 and Table 2 inform that the number of respondents is 188, and the number of items is 39. A person's responses can be interpreted as a person reliability of 0.86 and an item reliability of 0.98. It shows that the consistency of student answers is good, and the quality of the items in the instrument is classified as special. The Cronbach alpha (kr-20) value of 0.93 is included in the very good criteria, which means that the developed instrument has a very good reliability coefficient [18].

Table 3 is the item measure table including the logit values arranged from the highest (3.32) to the lowest (-2.46).

Table 3: Measure order.

Entry	Total	Total		Model	Infit		Outfit		Pt-measure		Exact match		
Number	Score	Count	Measure	S.e.	MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	Obs%	Exp%	Item
8	61	188	3.32	0.22	1.71	5.7	3.02	4.4	0.47	0.7	64.5	80.9	s8
36	73	188	2.75	0.21	0.49	-5.5	0.32	-4	0.84	0.7	94.6	81.3	s36
39	74	188	2.71	0.21	0.6	-4.1	0.46	-3	0.81	0.7	92.2	81.4	s39
35	75	188	2.66	0.21	0.44	-6.2	0.3	-4.4	0.85	0.7	96.4	81.5	s35
32	81	188	2.4	0.21	0.38	-7	0.28	-5.3	0.87	0.7	98.8	81.6	s32
33	81	188	2.4	0.21	0.44	-6	0.35	-4.5	0.85	0.7	97.6	81.6	s33
38	82	188	2.35	0.21	0.36	-7.2	0.27	-5.5	0.87	0.7	99.4	81.6	s38
31	85	188	2.22	0.21	0.44	-6	0.35	-4.8	0.85	0.69	97.6	81.4	s31
34	87	188	2.14	0.21	0.51	-5	0.5	-3.4	0.82	0.69	96.4	81.2	s34
37	96	188	1.76	0.2	0.81	-1.6	0.98	0	0.72	0.68	81.9	80.5	s37
2	97	188	1.72	0.2	0.97	-0.2	0.89	-0.7	0.69	0.67	83.7	80.5	s2
1	99	188	1.63	0.2	1.07	0.6	1.17	1.1	0.64	0.67	81.3	80.4	s1
16	100	188	1.59	0.2	3.12	9.9	4.09	9.9	0.01	0.67	38.6	80.3	s16
9	114	188	1.03	0.2	0.98	-0.2	1.19	1.1	0.62	0.62	80.7	78.3	s9
10	116	188	0.95	0.2	0.98	-0.1	0.98	-0.1	0.62	0.62	78.3	77.9	s10
7	123	188	0.68	0.2	0.64	-4.4	0.46	-3.4	0.71	0.59	86.1	76.8	s7
26	142	188	-0.07	0.2	1.23	2.6	2.83	4.4	0.38	0.5	71.1	77.2	s26
29	149	188	-0.37	0.21	1	0	1.42	1.2	0.45	0.46	79.5	78.7	s29
14	150	188	-0.41	0.21	0.83	-1.9	0.91	-0.1	0.5	0.45	84.3	79	s14
11	152	188	-0.5	0.21	0.99	-0.1	1.33	1	0.42	0.44	83.1	79.7	s11
27	154	188	-0.6	0.22	1.38	3.6	3.24	3.9	0.23	0.43	78.3	80.5	s27
13	162	188	-1	0.23	0.83	-1.5	1.11	0.4	0.42	0.37	86.1	84.5	s13
17	162	188	-1	0.23	0.93	-0.5	0.87	-0.1	0.39	0.37	86.1	84.5	s17
30	163	188	-1.06	0.24	1.19	1.5	1.13	0.4	0.3	0.37	84.3	85.1	s30
21	166	188	-1.23	0.25	0.92	-0.5	2.21	1.9	0.33	0.34	87.3	86.8	s21
18	168	188	-1.36	0.26	0.94	-0.4	1.18	0.5	0.34	0.33	88.6	88	s18
15	169	188	-1.43	0.26	0.85	-1	0.45	-1	0.38	0.32	89.2	88.6	s15
20	170	188	-1.5	0.27	0.86	-0.8	0.43	-1	0.37	0.31	89.8	89.2	s20
28	171	188	-1.57	0.27	1.29	1.6	5.8	4.1	0.08	0.3	89.2	89.7	s28
22	172	188	-1.65	0.28	0.94	-0.3	0.47	-0.8	0.33	0.29	90.4	90.3	s22
24	172	188	-1.65	0.28	1	0	0.58	-0.6	0.31	0.29	90.4	90.3	s24
3	175	188	-1.9	0.3	1.15	0.7	6.86	4.1	0.08	0.26	92.2	92.1	s3
5	175	188	-1.9	0.3	1.05	0.3	1.45	0.8	0.23	0.26	92.2	92.1	s5

23	175	188	-1.9	0.3	1.03	0.2	1.88	1.2	0.23	0.26	92.2	92.1	s23
19	176	188	-2	0.31	1.05	0.3	1.87	1.2	0.22	0.25	92.8	92.8	s19
6	178	188	-2.21	0.34	1.05	0.3	9.87	4.9	0.13	0.23	94	94	s6
12	178	188	-2.21	0.34	0.85	-0.5	0.35	-0.9	0.3	0.23	94	94	s12
4	179	188	-2.33	0.35	1.01	0.1	2.22	1.5	0.2	0.22	94.6	94.6	s4
25	180	188	-2.46	0.37	1.04	0.2	0.63	-0.3	0.2	0.21	95.2	95.2	s25
Mean	135.4	188	0	0.25	0.96	-0.9	1.66	0.1			87	84.8	
SD	40	0	1.82	0.05	0.45	3.4	1.97	3.2			10.9	5.6	

The item response theory (IRT) is a mathematical model considering the respondent giving the correct answer for each item [20]. The score obtained at the end of this test is not a test score, but an estimate of ability known as theta (θ). One of the scientists who developed IRT was Rasch, a mathematician from Denmark. Rasch argues that the opportunity to solve a problem correctly depends on the comparison between a person's ability and the level of difficulty of the problem [13][14][21]. Respondents with low abilities should not be *de facto* able to answer questions with a high level of difficulty [22]. Probability in the Rasch measurement is determined based on the difficulty level of the problem and the person's ability simultaneously.

The possibility of answering questions is differentiated based on the level of difficulty of the questions and individual abilities [23]. The far-right column describes the identity of the item. The items in the item measure table are sorted by the logit value from the highest to the lowest [21]. S8 obtained the highest logit value, and the lowest was S25. The logit value indicates the difficulty level of the problem. The higher the logit value, the more difficult the item is. Logits are interval data that range from a certain value from negative infinity to positive infinity [24]. The logit value can be seen in the measure column. The logit value in the measure column correlates with the value in the total score column. Item S8 obtained the lowest total score; namely 61, while item S25 - 180. It means that item S8, could answer correctly only 61 out of 188 students, whereas item S25 was answered correctly by 180 students.

The logit average value is always 0.00 with a standard deviation of 1.82. Table 3 also provides information that out of 39 items, there are 13 questions in the very difficult category, eight questions in the difficult category, 13 questions in the easy category, and five questions in the very easy category - this information is summarised in Table 4. The determination of item categories refers to +> 1SD very difficult, +1SD difficult, - 1SD easy, < - 1SD very easy [18].

Table 4: Fit order value.

Result	Percentage	Criteria
13	33.33%	Very difficult
8	20.51%	Difficult
13	33.33%	Easy
5	12.82%	Very easy

The fit order item assessment is used to assess the suitability of the items (validity) to explain whether the items are functional in carrying out measurements. The standard values used are mean square outfit (MNSQ) 0.5 < MNSQ < 1.5, z-standard outfit (ZSTD) -2.0 < ZSTD < +2.0, and point mean correlation (Pt Mean Corr) 0.4 < Pt Measure Corr < 0.85 (see Table 3) [25].

Determining whether the items are fit or not, there are several alternatives; for example, if the mean square value is appropriate, the z standard outfit value is appropriate, but the point mean correlation value does not meet the standard, the item can still be categorised as valid or fit provided that one can improve the item with other terms and *vice versa*. From the three standard criteria, the item does not have to have a value that satisfies all three categories, one or two criteria are enough, but it is better if the item meets the three standard values that have been determined. Table 5 shows the percentage of item fitting the chosen criteria.

Table 5: Item fit percentage.

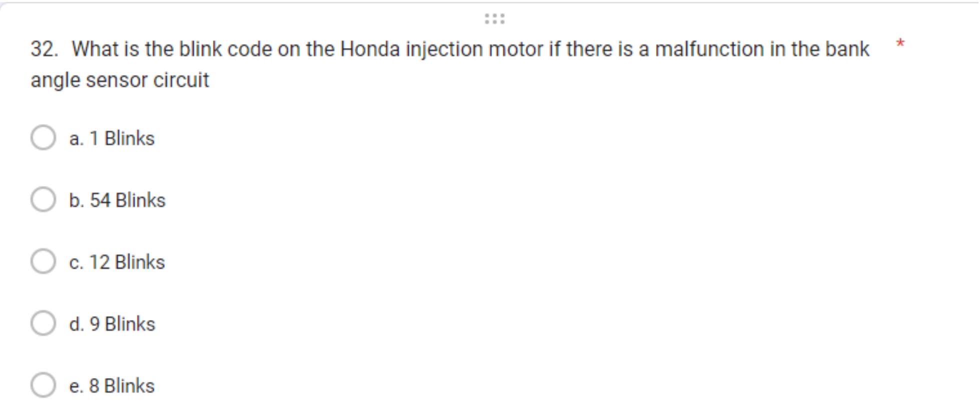
Result	Percentage	Criteria
9	23.08%	Not fit
15	38.46%	Fit 1 criteria
6	15.38%	Fit 2 criteria
9	23.08%	Fit 3 criteria

The analysis results in Table 5 show that nine (23.08%) items do not fall into the three specified value criteria. That means the nine items are invalid or not fit. Fifteen items (38.46%) are valid or fit criteria with one value included in the standard; six items (15.38%) are valid or fit criteria with two values according to the standard; and nine items (23.08%) are valid or fit criteria with three values according to the standard.

DISCUSSION

Learning evaluation also assists teachers in one of their tasks that is to help students recall the material they have learned. Learning and evaluation activities can also be carried out on-line [26]. Gen Z members who have grown up with technology from birth are predicted to earn higher college degrees, and they are now moving into the next phase, where they will make up the majority of the incoming workforce [27]. Student motivation is an important element needed for high-quality education [28]. One way to foster this motivation can come from student enjoyment when using cellular devices [29]. In designing assessment tests, there should be questions in the form of videos to be answered by students [30]. Compering static images and videos, there are more variations in the video form. Videos also contain much more information than static images [31].

As mentioned earlier, the findings from this study indicate that 15 items are included in the valid or fit criteria with one value according to the standard; six items are included in the valid or fit criteria with two values according to the standard; nine items are included in the valid or fit criteria with three values according to the standard. Questions that do not fit into the fit criteria include: 32, 38, 16, 26, 27, 28, 3, 6 and 12. The nine questions that do not fit can be removed from the evaluation sheet [32] so that the total number of items can be used for evaluation is 30 items. Here are some pictures of the items that do not fit.



32. What is the blink code on the Honda injection motor if there is a malfunction in the bank angle sensor circuit *

○ a. 1 Blinks

○ b. 54 Blinks

○ c. 12 Blinks

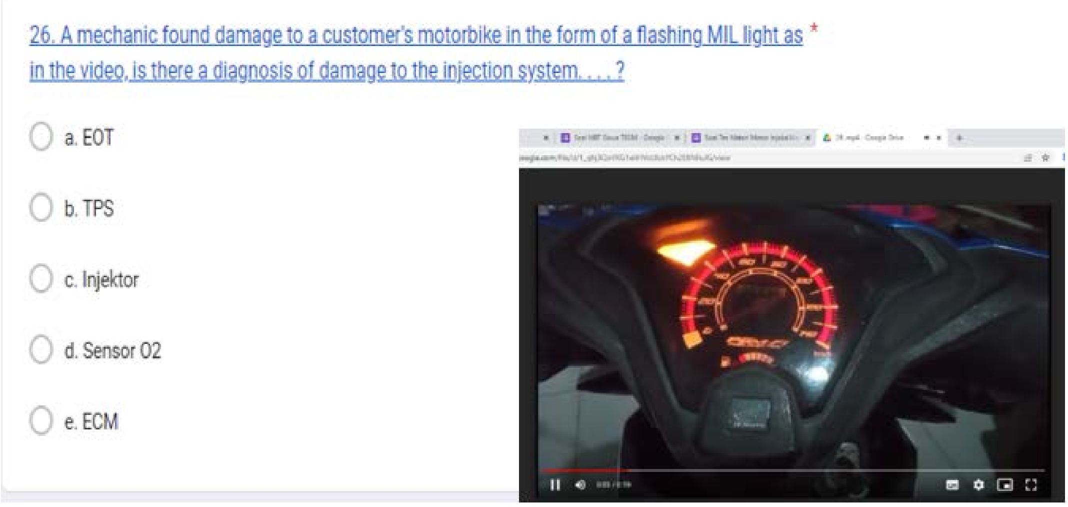
○ d. 9 Blinks

○ e. 8 Blinks

Figure 4: Item 32 does not fit the standard.

In item number 32, representing text items, the value of the MNSQ outfit is 0.28, ZSTD -5.3, PT CORR 0.87, the number of students who answered it correctly was 81 and the logit value of 2.4 put the item in the difficult category. This fit status exceeds the standard threshold value, meaning that item number 32 must be eliminated because it cannot take measurements properly. There is a student's misconception of item 32 that if the item does not meet the three criteria, it means that the item is not good enough to be corrected or replaced [23].

Analysis of the causes of item 32 not being fit showed that the question presented in the test is rarely included in the learning materials related to injection systems on motorbikes and that not all injection bikes use angel bank sensors, so the term angel bank is less familiar to students. Another analysis of the causes of item 32 not being fit points to the disproportionate variation of answers. The correct answer is *blink 54*; namely *b*, but the difference between the correct answer and other answers is too wide, if *blink 54* was the distractor factor, the question maker could use variations of the distractor's answers 52, 53, 55, 56 [33].



26. A mechanic found damage to a customer's motorbike in the form of a flashing MIL light as in the video, is there a diagnosis of damage to the injection system. . . . ? *

○ a. EOT

○ b. TPS

○ c. Injektor

○ d. Sensor O2

○ e. ECM

Figure 5: Item 26 does not fit the standard.

Item 26 represents the item using a video that is not fit as a question that students must answer. To answer the question, students first read the instructions in the next question. To display the video, students must click on the blue text that will take them to display a troubleshooting video on a motorcycle injection in the form of a blinking mil light. Students are asked to watch the next video. After understanding the troubleshooting, students close the video and provide analysis through answers to questions. The outfit value MNSQ 2.8, ZSTD 4.4, PT CORR 0.38, the number of students who answered the item correctly 142, and the logit value -0.07 determined this item as difficult.

The cause of the video questions in item 26 not being fit is that item 26 is the first troubleshooting item in this test. It is assumed that students are not used to working on troubleshooting questions with this video, as evidenced by the data for troubleshooting in item number 27. Item 28 is also not fit, as understanding the relationship between the video depiction and the real world requires some representational insight [34]. Unlike question number 28 that has been included in the unfit category, item 30 which is similar to 28 has been included in the easy item category and it fits with 163 students providing the correct answer [35].

If one compares the duration of the video, item 28 runs for 0.15 seconds, item 26 for 0.19 seconds, item 27 for 0.20 and item 30 for 0.29 seconds. Of the four items, question number 30 has a longer video duration than the other three items. The video duration factor influences the fit category of the items [36]. Also, the duration of the video affects the level of detail conveyed by the video. Lack of detail can make memories difficult to retrieve, and a slight discrepancy between memory and subsequent real-world experience can make transfer to the next situation difficult [37]. Video recordings should provide sufficient information about the programme and focus on the potential benefits for education [38]. The ability to answer video questions is influenced by students' ability to understand the symbolic nature of video images and to draw conclusions between the images and the objects they represent [39].

Although some items are not fit due to the variations of answers to the questions, the material and the duration of the video, overall, the design of the Gen Z items has a person reliability of 0.86 and an item reliability of 0.98, indicating that the consistency of the student's answers is good. The quality of the items in the instrument is classified as special. The Cronbach's alpha (kr-20) value of 0.93 is included in the very good criteria, meaning that the 39 items are suitable for a question design for Gen Z. However, the nine items mentioned earlier: 32, 38, 16, 26, 27, 28, 3, 6 and 12 would require some modifications (variations in the answer options for the items, quality of the item material and the addition of longer video duration than 0.29 second) to be included in the test.

CONCLUSIONS

The test using video is an innovation in the field of educational evaluation. Using video can provide a reality-like experience for test takers. Test participants gain experience working on the items by listening to video recordings and clicking the links provided. The videos displayed in this test is the result of documenting everyday cases of motorbikes. Then, the videos have gone through an editing process to adapt them to the theme and the level of difficulty of the items and make them very clear to observe. The development of video tests can be used to measure higher-order thinking skills. This type of video media is very suitable for the 5.0 era, where virtual videos support daily activities, especially of Gen Z.

Based on the findings and the discussion of the questions designed for Gen Z, the consistency of the student's answers is good, and the quality of the items in the instrument is classified as special. The very good criteria include the Cronbach's alpha (kr-20) value of 0.93. When preparing video items for Gen Z the duration of each item requires special attention. The videos in the test, especially for questions designed for troubleshooting, have a minimum duration of 0.29 seconds.

REFERENCES

1. Nguyen, T., Bui, T., Fujita, H., Hong, T.P., Loc, H.D., Snasel, V. and Vo, B., Multiple-objective optimization applied in extracting multiple-choice tests. *Engng. Applications of Artif. Intellig.*, 105, 104439 (2021).
2. Khonbi, Z.A. and Sadeghi, K., The effect of assessment type (self vs. peer) on Iranian university EFL students' course achievement. *Procedia-Social and Behavioral Sciences*, 70, 1552-1564 (2013).
3. Widoyoko, S.E.P., Teknik Penyusunan Instrumen Penelitian. Yogyakarta: Pustaka Pelajar (2012) (in Indonesian).
4. Istiyono, E., Dwandaru, W.B. and Rahayu, F., The developing of creative thinking skills test based on modern test theory in physics of senior high schools. *J. Cakrawala Pendidikan*, 37, **2** (2018).
5. Keinänen, M., Ursin, J. and Nissinen, K. How to measure students' innovation competences in higher education: evaluation of an assessment tool in authentic learning environments. *Stud. Educ. Evalue.* **58**, 30-36 (2018).
6. Shigli, K., Nayak, S.S., Gali, S., Sankeshwari, B., Fulari, D., Kishore, K.S. and Jirge, V., Are multiple choice questions for post graduate dental entrance examinations spot on? - Item analysis of MCQs in prosthodontics in India. *J. of the National Medical Assoc.*, 110, **5**, 455-458 (2018).
7. Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G. and Srinivas, V., Item analysis of multiple choice questions: a quality assurance test for an assessment tool. *Medical J. Armed Forces India*, **77**, S85-S89 (2021).
8. Mokshein, S.E., Ishak, H. and Ahmad, H., The use of Rasch measurement model in English testing. *J. Cakrawala Pendidikan*, **38**, 16-32 (2019) (in Indonesian).
9. Catanzano, T., Jordan, S.G. and Lewis, P.J., Great question! The art and science of crafting high-quality multiple-choice questions. *J. of the American College of Radiology*, 19, **6**, 687-692 (2022).

10. Azizah, A. and Wahyuningsih, S., Penggunaan model Rasch untuk analisis instrumen tes pada mata kuliah matematika aktuaria. *JUPITEK J. Pendidikan Mat.*, **3**, 45-50 (2020).
11. Tafonao, T., Saputra, S. and Suryaningwidi, R., Learning media and technology: generation Z and Alpha. *Indonesian J. of Instructional Media and Model*, 2, **2**,89-100 (2020).
12. Zain, N.H.M., Johari, S.N., Aziz, S.R.A., Teo, N.H.I., Ishak, N.H. and Othman, Z., Winning the needs of the Gen Z: gamified health awareness campaign in defeating COVID-19 pandemic. *Procedia Computer Science*, 179, 974-981 (2021).
13. Tennant, A. and Conaghan, P.G., The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. *Arthritis Care & Research*, 57, **8**, 1358-1362 (2007).
14. Lerdal, A., Kottorp, A., Gay, C., Aouizerat, B.E., Lee, K.A. and Miaskowski, C., A Rasch analysis of assessments of morning and evening fatigue in oncology patients using the Lee Fatigue Scale. *J. of Pain And Symptom Manage.*, 51, **6**, 1002-1012 (2016).
15. Kong, S.C. and Lai, M., Validating a computational thinking concepts test for primary education using item response theory: an analysis of students' responses. *Computers & Educ*., **187**, 104562 (2022).
16. Janke, S., Rudert, S.C., Petersen, Ä., Fritz, T.M. and Daumiller, M., Cheating in the wake of COVID-19: how dangerous is ad-hoc online testing for academic integrity? *Computers and Educ. Open*, **2**, 100055 (2021).
17. Nisa, S. and Pahlevi, T., Pengembangan instrument penilaian hots berbantuan quizizz pada mata pelajaran kearsipan SMK. *EDUKATIF J. ILMU Pendidik.* **3**, 2146-2159 (2021) (in Indonesian).
18. Sumintono, B. and Widhiarso, W., Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi). *Trim Komunikata Publishing House* (2014).
19. Kleppang, A.L., Steigen, A.M. and Finbråten, H.S., Using Rasch measurement theory to assess the psychometric properties of a depressive symptoms scale in Norwegian adolescents. *Health and Quality of Life Outcomes*, 18, **1**, 1-8 (2020).
20. Huang, Y.M., Lin, Y.T. and Cheng, S.C., An adaptive testing system for supporting versatile educational assessment. *Computers & Educ*., 52, **1**, 53-67 (2009).
21. Sumintono, B. and Widhiarso, W., Aplikasi pemodelan Rasch pada assessment pendidikan. *Trim komunikata* (2015) (in Indonesian).
22. Guillemin, F., Leplège, A., Briançon, S., Spitz, E. and Coste, J. (Eds), *Perceived Health and Adaptation in Chronic Disease*. Routledge (2017).
23. Boone, W.J., Staver, J.R. and Yale, M.S., *Rasch Analysis in the Human Sciences*. Netherlands: Springer (2014).
24. Linacre, J.M., Detecting multidimensionality: which residual data-type works best? *J. of Outcome Measure*., **2**, 266-283 (1998).
25. Maryati, M., Prasetyo, Z.K., Wilujeng, I. and Sumintono, B., Measuring teachers' pedagogical content knowledge using many-facet Rasch model. *Jurnal Cakrawala Pendidikan*, 38, **3**, 452-464 (2019).
26. Perry, M., Bates, M.S., Cimpian, J.R., Beilstein, S.O. and Moran, C., Impacting teachers' reflection on elementary mathematics classroom videos in online asynchronous professional learning contexts. *Teaching and Teacher Educ.: Leadership and Professional Develop.*, 1, 100003 (2022).
27. Pichler, S., Kohli, C. and Granitz, N., DITTO for Gen Z: a framework for leveraging the uniqueness of the new generation. *Business Horizons*, 64, **5**, 599-610 (2021).
28. Al-Azawei, A. and Alowayr, A., Predicting the intention to use and hedonic motivation for mobile learning: a comparative study in two Middle Eastern countries. *Technol. in Society*, **62**, 101325 (2020).
29. Nikou, S.A. and Economides, A.A., Mobile-based assessment: investigating the factors that influence behavioral intention to use. *Computers & Educ.*, **109**, 56-73 (2017).
30. Szymkowiak, A., Melović, B., Dabić, M., Jeganathan, K. and Kundi, G.S., Information technology and Gen Z: the role of teachers, the internet, and technology in the education of young people. *Technol. in Society*, **65**, 101565 (2021).
31. Ali, U., Lee, I.H. and Mahmood, M.T., Robust regularization for single image dehazing. *J. of King Saud University-Computer and Infor. Sciences*, 34, **9**, 7168-7173 (2022).
32. Gupta, V., Williams, E.R. and Wadhwa, R., Multiple-choice tests: A-Z in best writing practices. *Psychiatric Clinics*, 44, **2**, 249-261 (2021).
33. Dali, S.N., Pengantar teori sekor pada pengukuran pendidikan. *Jakarta: Penerbit Gunadarma* (1992) (in Indonesian).
34. DeLoache, J.S. and Marzolf, D.P., When a picture is not worth a thousand words: young children's understanding of pictures and models. *Cognitive Develop.*, 7, **3**, 317-329 (1992).
35. Im, E.O., Theory and research. *Nursing Research*, 61, **2**, 77 (2012).
36. Lynch, J.A., Idleburg, M.J., Kovacic, M.B., Childers-Buschle, K.E., Dufendach, K.R., Lipstein, E.A., McGowan, M.L., Myers, M.F. and Prows, C.A., Developing video education materials for the return of genomic test results to parents and adolescents. *PEC Innovation*, **1**, 100051 (2022).
37. Barr, R., Transfer of learning between 2D and 3D sources during infancy: informing theory and practice. *Developmental Review*, 30, **2**, 128-154 (2010).
38. Hill, M.E., Aliaga, S.R. and Foglia, E.E., Learning with digital recording and video review of delivery room resuscitation. *Proc. Semin. in Fetal and Neonatal Medic*., 27, **5**, 101396 (2022).
39. DeLoache, J.S. and Burns, N.M., Early understanding of the representational function of pictures. *Cognition*, 52, **2**, 83-110 (1994).