Four Tier Test (FTT) Development in The Form of Virtualization Static Fluid Test (VSFT) using Rasch Model Analysis to Support Learning During the Covid-19 Pandemic

N Anggraini, B H Iswanto and F C Wibowo
Physics Education Master Program, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta 13220, Indonesia
Email: nerai.anggraini@gmail.com, bhi@unj.ac.id, fcwibowo@unj.ac.id

Abstract. This study aims to develop a valid and reliable test instrument. The test instrument developed is a Four Tier Test (FTT) in a Virtual Static Fluid Test (Test) on Static Fluid material using the Rasch model analysis. The method used in this research is the method of research and development with the 4D model. This research produces an educational product in the form of an FTT in the form of VSFT. The items developed in this study amounted to 20 items. The validation results from several experts gave an average score of 88.13% with very good criteria. The CVR value for each item is 1, while the CVI value is 1 with very good criteria. The validation results from several educational practitioners gave an average score of 89.58% with very good criteria. The results of the Rasch Model analysis showed that the test instrument is accepted, the reliability of all the items has good criteria, all the items are valid and divided into three difficulties, also each item is not biased towards gender differences. The results of the analysis of student misconception showed that only 4.68% of the students understood the whole concept, 23.42% of students only understood partially, 58.95% of students have a misconception, and 12.95% of students did not understand the concept at all. The conclusion of this study is that the FTT instrument in the form of VSFT is valid and reliable as an assessment instrument during the Covid-19 period, and also able to analyze student conceptions.

1. Introduction
The new coronavirus pneumonia triggered by Covid-19 is now spreading around the world [1]. Covid-19 is a global health emergency, and the World Health Organization (WHO) declared it a pandemic. The pandemic has led education providers to decide on expanded education virtualization to remotely conduct online exams and master or doctoral seminars [2]. Online learning and virtual teaching are emerging as new avenues of science and engineering education and have accelerated the spread of educational resources globally, especially during pandemics [3].

During the online learning period, a test tool is needed to help students understand the concepts asked in the questions so that it makes it easier for students to answer questions. One of the test tools that support this convenience is a type of virtual test (virtualization test). The use of the Virtualization test allows question makers to use images, graphics, animations, and videos that can visualize test subject statements that contain abstract concepts and are difficult to explain in words to make it easier to understand questions [4]. The use of virtual tests gives a positive response to teachers and students in terms of ease of use of virtual tests, multimedia, benefits and time of testing [5,6].

Static fluid is a fundamental physics material that examines the physical concept with all its reactions to a given force. However, understanding the concept of static fluids is often seen as complex by students because it requires in-depth reasoning. One of the learning difficulties that is often found is a misconception.
Misconceptions can be a barrier to organizing knowledge, so it needs to be overcome [7]. Therefore, it is crucial to develop an evaluation tool that can detect misconceptions experienced by students [8].

The four-tier test format instrument is one of the instruments used to diagnose students’ conception of a physics concept [9]. The four tiers diagnostic test is a four-level multiple-choice test composed of questions with four answer choices, confidence in choosing an answer, the reason for choosing the answer, and the level of confidence in the reason. [10]. The first level is a regular multiple-choice test with a cheat that addresses specific misconceptions. The second level of the test asks for confidence in the first level. The third level of the test asks the reasons for the answer at the first level. The fourth level of the test asks for the confidence of the answers at the third level. However, until now, many test instrument developments have been carried out using classical measurement model analysis.

The Rasch model has been used to analyze various assessment instruments to provide detailed information on several aspects of the quality of the test. The Rasch model is used to analyze instrument items and analyze students [11]. The Rasch model can be used to determine the reliability index to the item analysis for each level, the student's reliability, and dimensionality and detect the bias of the items on the instrument. [12]. Thus, the Rasch model analysis provides several options and ways in which a test instrument is developed.

Using the Rasch model has many advantages. The Rasch Model analysis supports the development and validation of invariance measures in providing empirical evidence and insights for critical psychometric properties of tests and questionnaires. The Rasch model can explain the differences in the difficulty of different levels [13]. The Rasch model can also determine a student's ability, where the likelihood of a student answering an item correctly increases monotonically with the student's abilities. [14].

So that in the context of assessing the Covid-19 pandemic, this study aims to develop a valid and reliable test instrument. The test instrument developed is a Four Tier Test (FTT) in a Virtual Static Fluid Test (Test) on Static Fluid material using the Rasch model analysis.

2. Method
The method used in this research is the method of research and development with the 4D model. 4D development consists of four stages, namely: Defining, Designing, Developing, and Disseminating).

2.1. Defining
The purpose of this stage is to define and define the terms of the virtual test instrument in the four-tier test format, which begins with an objective analysis of the constraints of the material the instrument will develop. This stage includes four main steps: needs analysis, student analysis, concept analysis, and learning objectives.

2.2. Designing
The purpose of this stage is to prepare a Four-Tier Test (FTT) instrument, which will be outlined in the form of a Virtualization Static Fluid Test (VSFT). It consists of three steps, namely designing a VSFT grid in FTT format, compiling VSFT questions in FTT format and compiling a validation instrument.

2.3. Developing
The purpose of this stage is to produce revised learning instruments based on suggestions from experts. This stage includes validation of the tools by experts and revisions.

2.4. Disseminating
The purpose of this stage is to produce tested learning instruments. This stage includes revision and broad implementation to students, data analysis and making conclusions. The data analysis obtained will be analyzed with the Rasch Model measurement model.
3. Result and Discussion

3.1. Four-Tier Test (FTT) in the form of Virtualization Static Fluid Test (VSFT)

This research produces an educational product in the form of a Four Tier Test (FTT) in a Virtualization Static Fluid Test (VSFT). Initially, this educational product had a homepage that contained student identities.

After students fill in their data entirely and start the test, the system will move from the Home page to the next page, which contains many items that students will answer. The items developed in this study amounted to 20 items. Each item is equipped with a stimulus to help students work on it. Each item requires students to give four answers in sequence (four-tier test).

3.2. Expert Validation

The validation results by several experts gave an average value of 88.13% with very good criteria. This informs that the FTT in the form of VSFT has been developed with very good quality and can be used for the test implementation. The value is divided into three aspects: the Physics material aspect of 86.67% with very good criteria, the media aspect of 88.00% with very good criteria, and the assessment aspect of 90.00% with good criteria.
3.3. Content Validity Ratio (CVR) and Content Validity Index (CVI)

Each item that has been developed is then validated to determine the validation of the content of each item. After that, the validation results are used to calculate the CVR value for each item. The CVR value for each item is 1. This shows that all items are valid and fit for use. This is in line with previous research regarding developing an instrument that states that content validity (through CVR calculations) is very important and crucial. [15,16].

The CVI value obtained has a value of 1 with very good criteria. This explains that all items that have been developed when viewed as a whole as a test instrument are declared valid and usable. Validity content ensures that items and test instruments that have been designed and developed can measure adequately [17].

3.4. Educational Practitioner Validation

The validation results by several educational practitioners gave an average score of 89.58% with very good criteria. This informs that the FTT in the form of VSFT has been developed with very good quality and can be used for the test implementation. The value is divided into three aspects: the Physics material aspects of 88.39% with very good criteria, the design aspect of 89.84% with very good criteria, and the aspects of function and purpose of 91.41% criteria.


3.5. Rasch Model Analysis

The items that have been developed and validated theoretically will then be tested on many students to be analyzed with the Rasch Model. This is supported by a statement in a previous study that the Rasch Model analysis provides a solution to the limitations experienced using the classical measurement model [18] so fit items can be gotten [19]. The results of the Rasch Model analysis for each item that has been developed are as follows.

The results of the first Rasch Model analysis showed a unidimensionality value of 22.4%. This explains that the test instrument is accepted and can measure what should be measured. Unidimensionality is the latent property of a measuring instrument [20,21] in verifying the measurement construction of the instrument [22].

The results of the second Rasch Model analysis show a reliability value of 0.92. This explains that the reliability between all the items that have been developed and all the students taking the test has a special criterion. The results of the third Rasch Model analysis show a reliability value of 0.90. This explains that the reliability of all the items that have been developed has good criteria. This is in line with several previous studies which stated that reliability provides more detailed information between the property of the items on the instrument and the individuals who respond to these items [23]. Reliability also informs about consistency, both consistency of items and consistency of person [24].

The fourth Rasch Model analysis results show that each item that has been developed meets at least one of the three item-fit criteria. This explains that all the items that have been developed are valid. This is supported by several studies that state that validity informs the validity of the test content and informs the consequences of using the test score [25]. Using the Rasch model, the analysis results can determine the suitability of the items to the identified constructs [26] against the test instruments in general [27].

The results of the fifth Rasch Model analysis show that all the items that have been developed are divided into three difficulty levels as follows:

| Difficulty Level | Number of Question |
|------------------|--------------------|
| Hard             | 13, 2, 19          |
| Medium           | 10, 3, 9, 17, 8, 4, 14, 1, 12, 20, 5 |
| Easy             | 11, 6, 15, 18, 16  |

This is in line with other studies which state that the varying levels of difficulty of the questions is another criterion of a good measuring instrument [28,29]. The level of difficulty of the items does not have to be sequential in a test instrument [30].

This study also analyzed the possibility of bias against the existence of gender differences (male and female). The sixth Rasch Model analysis results showed that the DIF value of each item was less than 0.05. This explains that each item is not biased towards gender differences (male and female). Previous studies have shown that the DIF value checks the possibility of item bias caused by responses from different groups [31–33]. Biased questions arise because there are two different groups with the same ability, but there are significant differences when answering the questions [34].

The interpretation using the Rasch measurement model provides more detailed information about the quality of items and persons and their appropriateness in measuring what should be measured [35,36]. Apart from analyzing the quality of the items that have been developed, this study also analyzes the quality of the students taking the test. This is following several previous studies that stated that the Rasch Model analysis was able to explain the quality of the items developed and the quality of the students taking the test [37]. The Rasch model provides information about the consistency of test-takers in answering questions so that there is a possibility that the test taker may be careless in answering, guessing, or cheating [38]. This is, of course, very helpful for teachers in providing evaluations by the abilities of students who take the test [39,40]. This also makes it easy for teachers to check/evaluate and...
determine the appropriate feedback for each student [41] so that the evaluation of student test-takers by the teacher becomes easier and more precise [42].

The results of the first Rasch Model analysis show a reliability value of 0.86. This explains that the reliability of all students taking the test has good criteria.

The second Rasch Model analysis results showed that the test takers' ability was divided into three levels, namely high, medium, and low abilities. By changing the scores obtained by the students taking the test, the test takers' ability can be determined [43]. There were 52 students who took the test with high ability, 246 students with moderate ability, and 65 students with low ability. This agrees with previous research, which states that a person measure indicates the ability of a person [44].

The third Rasch Model analysis results show that each student who takes the test meets at least one of the three person-fit criteria. This explains that each student who takes the test has an appropriate response pattern. The results of this analysis are important to obtain because if there is an inappropriate response pattern, the Rasch Model analysis will predict the possibility of students taking the test carelessly working on the questions, being lucky in answering, or the possibility of cheating [45].

3.6. Analysis of Student Misconceptions

The analysis results showed that out of 363 students, only 4.68% of the students understood the whole concept of Static Fluid, and 23.42% of students only understood part of the concept of Static Fluid. Meanwhile, 58.95% of students experienced a misconception of Static Fluid material, and 12.95% of students did not understand the concept of Static Fluid at all. This is in line with the statement in previous research that many students still experience misconceptions about physics material [46]. Wherewith the four-tier test, students' misconceptions can be easily identified [47–50].

4. Conclusion

Based on the results of the research that has been obtained, it can be concluded that the Four Tier Test (FTT) instrument in the form of Virtualization Static Fluid Test (VSFT) is valid as an assessment instrument for learning support during the Covid-19 period. In addition, the Four Tier Test (FTT) instrument in the form of a Virtualization Static Fluid Test (VSFT), which has been developed, is also reliable as a learning support assessment instrument during the Covid-19 period. Then, the Four Tier Test (FTT) instrument in the form of a Virtualization Static Fluid Test (VSFT) that has been developed is also able to analyze student conceptions, making it easier for teachers to evaluate and respond appropriately to each student.

References

[1] Feng G, Liu L, Cui W and Wang F 2020 Electron beam irradiation on novel coronavirus (COVID-19): A Monte-Carlo simulation Chinese Physics B 29
[2] Ray S and Srivastava S 2020 Virtualization of science education: a lesson from the COVID-19 pandemic Journal of Proteins and Proteomics 11
[3] Hunter P 2015 The virtual university: Digital tools for e-learning and remote learning are becoming an increasingly important tool for teaching at universities EMBO Reports 16 146–8
[4] Nahadi N, Firman H and Kurniadi H 2018 Development and Validation of Chemistry Virtual Test Based Multiple Representations Journal of Education and Learning (EduLearn) 12
[5] Stavesand J E, Reglitz S and Himmler A 2017 Optimizing the Benefit of Virtual Testing with a Process-Oriented Approach SAE International Journal of Aerospace 10
[6] Mahfira C, Sanjaya Y and Rusyati L 2018 Students’ Impression towards Science Virtual Test (SVT) on Digestive System Topic Journal of Physics: Conference Series vol 1013
[7] Istighfarin L 2015 Profil Miskonsepsi Siswa Pada Materi Struktur Dan Fungsi Jaringan Tumbuhan BioEdu 4
[8] Fariyani Q, Rusilowati A and Sugianto 2015 Pengembangan Four-Tier Diagnostic Test Untuk Mengungkap Miskonsepsi Fisika Siswa SMA Kelas X Journal of Innovative Science Education 4 41–9
[9] Rawh P, Samsudin A and Nugraha M G 2020 Pengembangan Four-Tier Diagnostic Test untuk Mengidentifikasi Profil Konsepsi Siswa pada Materi Alat-Alat Optik WaPFi (Wahana Pendidikan Fisika) 5
[10] Islami D, Munawaroh F and Hadi W P 2018 Analisis miskonsepsi siswa pada mata pelajaran ipa materi listrik statis menggunakan four tier test Science Education National Conference

[11] Zamri A and Nordin 2015 Modeling a Multiple Choice Mathematics Test with the Rasch Model Indian Journal of Science and Technology 8 1–6

[12] Suryani Y E 2018 Aplikasi Rasch Model dalam Mengevaluasi Intelligenz Structure Test Psikohumaniora: Jurnal Penelitian Psikologi 3 73–100

[13] Fulmer G W, Chu H, Treagust D F and Neumann K 2016 Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model Asia-Pacific Science Education 1 1–16

[14] Xiao Y, Han J, Koenig K, Xiong J and Bao L 2018 Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson’s classroom test of scientific reasoning Physical Review Physics Education Research 14 2014

[15] Alizadeh-Siuki H, Tehrani H, Gholian-Aval M, Ebrahimipour H, Jafari A and Vahedian-Shahroodi M 2020 Psychometric properties of a questionnaire on brucellosis prevention behaviors based on the precede model among rural farmers and their family members Risk Management and Healthcare Policy 13

[16] Waltz C F, Strickland O L and Lenz E R 2016 Measurement in Nursing and Health Research

[17] Al-Taweel D and Awad A 2020 Development and validation of medication assessment tools to evaluate prescribing adherence to evidence-based guidelines for secondary prevention of coronary heart disease in post-acute coronary syndromes patients in Kuwait PLoS ONE 15

[18] Nurhudaya, Taufik A, Yudha E S and Suryana D 2019 The Raven’s advanced progressive matrices in education assessment with a Rasch analysis Universal Journal of Educational Research 7

[19] Azizah A and Wahyuningsih S 2020 Penggunaan Model Rasch Untuk Analisis Instrumen Tes Pada Mata Kuliah Matematika Aktuaria JUPITEK: Jurnal Pendidikan Matematika 3

[20] Al Zoubi F, Mayo N, Rochette A and Thomas A 2018 Applying modern measurement approaches to constructs relevant to evidence-based practice among Canadian physical and occupational therapists Implementation Science 13

[21] Nielsen T and Dammeyer J 2019 Measuring higher education students’ perceived stress: An IRT-based construct validity study of the PSS-10 Studies in Educational Evaluation 63

[22] Chang J, Li Y, Song H, Yong L, Luo L, Zhang Z and Song N 2020 Assessment of Validity of Children’s Movement Skill Quotient (CMSQ) Based on the Physical Education Classroom Environment BioMed Research International 2020

[23] Bodzin A, Hammond T, Fu Q and Farina W 2020 Development of Instruments to Assess Students’ Spatial Learning Attitudes (SLA) and Interest in Science, Technology and Geospatial Technology (STEM-GEO) International Journal of Educational Methodology 6 67–81

[24] Koçak D 2020 Investigation of rater tendencies and reliability in different assessment methods with many facet rasch model International Electronic Journal of Elementary Education 12

[25] Mokshein S E, Ishak H and Ahmad H 2019 The use of rasch measurement model in English testing Cakrawala Pendidikan 38

[26] Md Yunos J, Ibrahim Mukhtar M, Alias M, Lee M F, Tee T K, Rubani S N K, Hamid H, Yunus F A N, Sulaiman J and Sumarwati S 2017 Validity of vocational pedagogy constructs using the rasch measurement model Journal of Technical Education and Training 9

[27] Yasin S N T M, Yunus M F M and Ismail I 2018 The use of rasch measurement model for the validity and reliability Journal of Counseling and Educational Technology 1

[28] Barus G, Sinaga J D and Moron Y 2019 Testing the quality of film-based creative and innovative character education test with rasch model Jurnal Konseling dan Pendidikan 7

[29] Hamdu G, Fuadi F N, Yulianto A and Akhiriani Y S 2020 Items Quality Analysis Using Rasch Model To Measure Elementary Students’ Critical Thinking Skill On Stem Learning JPI (Jurnal Pendidikan Indonesia) 9

[30] Isnani I, Utami W B, Susongko P and Lestiani H T 2019 Estimation of college students’ ability on real analysis course using Rasch model Research and Evaluation in Education 5

[31] Xu R H, Wong E L yi, Lu S Y jun, Zhou L M, Chang J H and Wang D 2020 Validation of the
Toronto Empathy Questionnaire (TEQ) Among Medical Students in China: Analyses Using Three Psychometric Methods *Frontiers in Psychology* **11**

[32] Veas A, Castejón J L, Gilar R and Miñano P 2016 Validation of the Spanish adaptation of the School Attitude Assessment Survey-Revised using multidimensional Rasch analysis *Anales de Psicología* **33**

[33] Rahmani B D 2018 Differential Item Functional Analysis on Pedagogic and Content Knowledge (PCK) Questionnaire for Indonesian Teachers using RASCH Model *Journal of Physics: Conference Series* **11**

[34] Alwi I 2017 Sensitivity Of Mantel Haenszel Model And Rasch Model As Viewed From Sample Size *Journal Evaluasi Pendidikan* **2**

[35] Bichi A A, Talib R, Atan N A, Ibrahim H and Yusof S M 2019 Validation of a developed university placement test using classical test theory and Rasch measurement approach *International Journal of Advanced and Applied Sciences* **6** 22–9

[36] Tornabene R E, Lavington E and Nehm R H 2018 Testing validity inferences for Genetic Drift Inventory scores using Rasch modeling and item order analyses *Evolution: Education and Outreach* **11**

[37] Maulana S, Supriyati Y and Astra I M 2020 Development of Physics Problems for The Final Assessment in High School Level with Rasch Model Analysis *Unnes Science Education Journal* **9** 22–9

[38] Susongko P 2016 Validation of Science Achievement Test with the Rasch Model *Jurnal Pendidikan IPA Indonesia* **5** 268–77

[39] Suranata K, Rangka I B, Ifdil I, Ardi A, Susiani K, Prasetyaningtyas W E, Daharnis D, Alizamar A, Erlinda L and Rahim R 2018 Diagnosis of students zone proximal development on math design instruction: A Rasch analysis *Journal of Physics: Conference Series*

[40] Mursidi A and Soeharto 2016 An Introduction: Evaluation of Quality Assurance for Higher Educational Institutions using Rasch Model *Journal of Education, Teaching and Learning* **1** 1–6

[41] Börkan B 2017 Exploring Variability Sources in Student Evaluation of Teaching via Many-Facet Rasch Model *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*

[42] Erman Aslanoglu A, Karakaya I and Sata M 2020 Evaluation of university students’ rating behaviors in self and peer rating process via many facet rasch model *Eurasian Journal of Educational Research* **2020**

[43] Planinic M, Boone W J, Susac A and Ivanjek L 2019 Rasch analysis in physics education research: Why measurement matters *Physical Review Physics Education Research* **15**

[44] Anselmi P, Colledani D and Robusto E 2019 A Comparison of Classical and Modern Measures of Internal Consistency *Frontiers in Psychology* **10**

[45] Boone W J 2016 Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education* **15** 1–7

[46] Abbas M I H 2019 Identifikasi Miskonsepsi Mahasiswa Tadris Fisika Menggunakan Four Tier Diagnostic Test pada Mata Kuliah Kalkulus II *Jurnal Matematika dan Pendidikan Matematika* **4**

[47] Fenditasari K, Jumadi, Istiyono E and Hendra 2020 Identification of misconceptions on heat and temperature among physics education students using four-tier diagnostic test *Journal of Physics: Conference Series* vol 1470

[48] Tumanggor A M R, Supahar, Kuswanto H and Ringo E S 2020 Using four-tier diagnostic test instruments to detect physics teacher candidates’ misconceptions: Case of mechanical wave concepts *Journal of Physics: Conference Series* vol 1440

[49] Yuberti Y, Suryani Y and Kurniawati I 2020 Four-Tier Diagnostic Test with Certainty of Response Index to Identify Misconception in Physics *Indonesian Journal of Science and Mathematics Education* **3**

[50] Sholihat F N, Samsudin A and Nugraha M G 2017 Identifikasi Miskonsepsi dan Penyebab Miskonsepsi Siswa Menggunakan Four-Tier Diagnostic Test Pada Sub-Materi Fluida Dinamik: Azas Kontinuititas *Jurnal Penelitian & Pengembangan Pendidikan Fisika* **3**