Available online at: http://journal.uny.ac.id/index.php/jipi

Jurnal Inovasi Pendidikan IPA, 11 (2), 2025, 469-485

Psychometric Validating a Video-Based Rubric for Physics Teacher Assessment Using the Rasch Model

Dian Artha Kusumaningtyas [1] *, Yuhanis Mhd Bakri [2], Muhammad Syahriandi Adhantoro[3], Ishafit Ishafit [1], Efi Kurniasari [4]

[1] Universitas Ahmad Dahlan. Yogyakarta, Indonesia.

[2] Universiti Pendidikan Sultan Idris. Malaysia.

[3]Universitas Muhammadiyah Surakarta, Indonesia.

[4]Universitas Muhammadiyah Sukabumi, Indonesia.

* Corresponding Author. E-mail: dian.artha@pfis.uad.ac.id

Received:6 January 2025; Revised: 22 May 2025; Accepted: 11 August 2025

Abstract: This study aims to validate the psychometric properties of a video-based assessment rubric designed to evaluate the pedagogical and professional competencies of physics prospective teachers. The validation process employed the Rasch model to examine item reliability, person reliability, item fit, category functioning, and the separation index. Data were collected from 61 participants using a 25-item rubric that covered various aspects, including Pancasila values, instructional structure, student-centered learning, professional ethics, and evaluation of learning. The results showed that the item reliability was high (0.83), while person reliability was moderate (0.75), with a person separation index of 1.75. Some items exhibited misfit with the Rasch model, indicating the need for revision or refinement. Analysis of category functioning revealed overlapping probabilities between adjacent categories, suggesting that the response scale could benefit from more precise distinctions. Overall, the rubric demonstrated acceptable psychometric quality for assessing the competencies of physics prospective teachers, although several items and scale structures require improvement to enhance measurement precision.

Keywords: physics prospective teachers, professional competence, Rasch model, video-based assessment.

How to Cite: Kusumaningtyas, D. A., Bakri, Y. M., Adhantoro, M.S., Ishafit., Kurniasari, E. (2025). Psychometric Validating a Video-Based Rubric for Physics Teacher Assessment Using the Rasch Model. *Jurnal Inovasi Pendidikan IPA, 11*(2), 469-485. doi:https://doi.org/10.21831/jipi.v11i2.82060

INTRODUCTION

Improving teacher competence remains a central focus in Indonesia's efforts to enhance the quality of education. The government has implemented various programs, such as Pendidikan Profesi Guru (Teacher Professional Education), aimed at professionalizing teachers through structured training and certification (Fitriansyah et al., 2020; SA et al., 2021). These initiatives not only recognize teaching competence formally but also offer incentives to encourage continuous improvement. Technological advancements and curriculum modernization have also supported teachers in adapting to the evolving needs of students (Adhantoro et al., 2024). Nevertheless, systemic challenges persist. Unequal access to professional development opportunities between urban and rural areas remains a significant issue (Sudirman, 2019). Additionally, excessive administrative burdens often limit the time and energy teachers can dedicate to instructional improvement (Bai et al., 2021; Li et al., 2020). To address these issues, inclusive policy reforms, digitalization of bureaucratic processes, and collaborative support from multiple stakeholders have been proposed (Hursen, 2021; Larson et al., 2020; Ettekal & Shi, 2020). A crucial aspect of teacher professionalism is pedagogical content knowledge (PCK), a concept that blends subject matter expertise with pedagogical strategies. PCK enables teachers to deliver complex concepts effectively by considering students' backgrounds, cognitive development, and everyday learning

obstacles (Scherer et al., 2018; Akyuz, 2018). For physics prospective teachers, strong PCK is essential due to the abstract and technical nature of many physics topics.

Previous studies have demonstrated that well-developed PCK enables teachers to identify misconceptions, apply appropriate instructional methods, and enhance student engagement and achievement (Metz, 2021; Nurulsari et al., 2020; Hsu & Chen, 2019). However, the effectiveness of PCK in practice depends on how accurately it is measured during teacher preparation. This highlights the need for reliable and valid assessment tools that accurately evaluate the pedagogical and professional competencies of teacher candidates.

To that end, rubric-based video assessments have emerged as a promising method. These tools allow evaluators to observe and rate teaching performance in real classroom scenarios, capturing both pedagogical delivery and professional behaviors. However, to ensure fair and practical use, such instruments must undergo rigorous validation using psychometric models. This study aims to validate the psychometric properties of a video-based rubric instrument designed to assess the pedagogical and professional competence of physics prospective teachers in Indonesia. By applying the Rasch model, this study examines the instrument's reliability, item fit, category functioning, and separation index. The validation process ensures that the rubric not only reflects the intended constructs but also functions consistently across diverse respondents.

The novelty of this study lies in its contextualized and empirical approach. While prior research, such as Nurulsari et al. (2020), has highlighted the importance of PCK in practice, few studies have focused on the instrumentation aspect, especially in the Indonesian context. Furthermore, this study provides a deeper understanding of how item performance and scoring categories behave when used to assess authentic classroom video performances. Through this research, we aim to contribute a validated and reliable instrument that supports ongoing efforts to improve teacher education in Indonesia, particularly in physics education, where pedagogical precision and professional integrity are paramount.

METHOD

1. Research Design and Approach

This study employed a quantitative exploratory descriptive design to examine the psychometric properties of a video-based assessment rubric. The exploratory approach was selected to explore item-level behavior, rating scale functioning, and the instrument's overall measurement quality, particularly because the rubric had not undergone formal psychometric validation. The Rasch model served as the primary analytical framework for assessing item fit, reliability, and category functioning, ensuring that the instrument could consistently and accurately measure the pedagogical and professional competencies of physics prospective teachers.

2. Participants and Data Sources

The participants were 61 physics prospective teachers enrolled in the Teacher Professional Education (PPG) Program at a teacher education institution in Indonesia. Each participant submitted a classroom teaching video during their practicum, which served as the primary source of data. These videos were evaluated using a rubric specifically designed to assess pedagogical and professional competencies, including aspects such as instructional planning, student-centered learning, classroom ethics, and assessment strategies.

3. Research Instrument

The primary instrument used in this study is a video-based performance assessment rubric comprising **25 items.**

Table 1. Description of Video Assessment Rubric Items

Competency Components	Information	Code
Upholding Pancasila Values	Piety-1	KT1
	Piety-2	KT2
	Humanity	KM
	Association	PS
	Democratic	DM

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Competency Components	Information	Code
	Justice	THE
Professional Ethics	Emotional Maturity	KE
	Professionalism	PRO
Entrepreneurial Spirit	Entrepreneurial Spirit	BW
Material Structure Analysis	Applying Teaching Material Structure Analysis in Learning	SMA
	Applying Teaching Material Flow Analysis in Learning	OR
Structured and Continuous Learning	Teachers carry out learning in a structured and continuous manner.	TRS
	Teachers carry out core phase learning in a structured and continuous manner.	BE
	The teacher carries out the closing phase of learning in a structured and continuous manner.	FP
Learner-Centered Learning	Teachers implement student-centered learning according to the model syntax or method/strategy chosen to develop faith, piety, noble character, and independence through the values contained in the content of the field of study.	SCLI
	Teachers carry out student-centered learning according to the model syntax or method/strategy chosen to build knowledge and solve problems in the field of study.	SCLP
	Teachers implement student-centered learning according to the model syntax or methods/strategies chosen to develop skills and solve problems within the content area of study.	SCLK
Safe and Comfortable Learning Environment	Teachers carry out learning by providing a safe and comfortable learning environment.	LB
Accommodative, adaptive, and progressive to the development of the times	Teachers use learning models/strategies that are in accordance with the latest developments in science, technology, and the arts (science and technology).	MSI
	Teachers utilize learning media that are aligned with the latest developments in science, technology, and the arts.	MSI2
Evaluate learning based on student development.	Teachers evaluate learning input according to student development.	EVA
	Teachers evaluate the learning process based on student development.	EVAP
	Teachers evaluate learning outcomes based on student development.	EVAH
Evaluating Curriculum-Based Learning	Teachers carry out evaluations based on the curriculum.	I
Evaluating Learning Based on the Learning Environment	Teachers carry out evaluations based on the learning environment.	HE

4. Research Procedure

a. Rubric Construction and Expert Validation

Following the initial development of the rubric, a validation process was conducted, involving reviews from three subject-matter experts, two university lecturers in educational assessment, and one curriculum specialist, to confirm the content's relevance, clarity, and observable performance indicators.

b. Video Data Collection

Participants recorded one full teaching session during the practicum. Standardized procedures were applied to ensure consistency in recording format, duration, and submission. Videos were securely stored and anonymized.

c. Scoring Process by Trained Raters

Each video was independently evaluated by two trained raters, both of whom were university lecturers with substantial experience in teacher evaluation. Prior to scoring, calibration sessions were conducted to align rating interpretations and ensure inter-rater consistency. In cases where discrepancies exceeded one scale point, a moderation process involving a third expert was implemented to reach consensus. While these procedures were designed to enhance scoring reliability, it is acknowledged that potential rater bias and challenges inherent in assessing teaching practice videos, such as limited camera angles,

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

variations in audio quality, and the absence of specific contextual cues, might still influence the objectivity of the results.

d. Data Compilation and Cleaning

Rubric scores were compiled into a matrix (person × item) with ordinal values from 1 to 5. Missing or inconsistent data were cleaned before analysis.

5. Psychometric Analysis Using the Rasch Model

The Rasch model was applied to examine both item-level and scale-level psychometric characteristics using **Winsteps** software. The analysis encompassed several key components. First, **item fit statistics** were assessed using **Infit and Outfit Mean Squared Error (MSE)** values to detect inconsistencies and potential noise in participants' responses. In addition, **Z-standardized scores (ZSTD)** were calculated to evaluate whether each item performed as expected under the model. Items with MNSQ values outside the ideal range of **0.5 to 1.5** or ZSTD values beyond **-2 to +2** were flagged as misfitting and considered for further review.

Second, the analysis examined **reliability and separation indices**. **Item reliability** reflected the stability and replicability of item difficulty across different samples, while **person reliability** measured the internal consistency of responses across participants. The **separation index** was used to determine the instrument's ability to distinguish between multiple strata of participant competence, indicating the precision of the measurement scale.

Third, a category functioning analysis was conducted to ensure the integrity of the five-point Likert scale. This involved evaluating Andrich thresholds to verify that response categories were ordered logically and distinctly. Category probability curves were generated to visualize how each rating category functioned across different levels of participant ability. Coherence checks were also performed to assess the consistency between the measures and the expected category choices; categories with overlapping thresholds or disorderly patterns were marked for possible revision.

Finally, the analysis included the calculation of point-measure correlations (PtMea Corr) for each item to assess its alignment with the underlying construct being measured. Items with higher PtMea Corr values, ideally approaching +1, indicated a stronger and more consistent contribution to the overall construct.

6. Interpretation and Instrument Refinement

Based on the results of the Rasch analysis, several actions were taken to enhance the instrument's quality. Items that demonstrated poor fit or low point-measure correlation were identified for revision or considered for potential removal to ensure measurement precision and accuracy. Additionally, items with disordered or overlapping response categories were revised to enhance the clarity and discriminative power of the rating scale. Descriptors within the rubric were also refined to better align with the intended constructs, ensuring that each item effectively captured the targeted pedagogical or professional competency. The improved rubric is intended for broader application in teacher education programs, particularly for evaluating video-based teaching performance during the practicum phase.

RESULT AND DISCUSSION

1. Result

Pedagogical content knowledge (PCK) competency for physics prospective teachers is an essential aspect in establishing effective teaching quality. PCK includes the ability to integrate in-depth knowledge of subject matter with appropriate teaching strategies to facilitate student understanding. The results of a study using Rasch analysis with the Many-Facet Rasch Measurement (MFRM) approach to evaluate the PCK abilities of physics prospective teachers showed significant results. These results are

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

illustrated in Figure 1. evaluate the PCK abilities of physics prospective teachers showed significant results. These results are illustrated in Figure 1.

PERSON	61 INPUT	61 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	95.3	25.0	-.41	.32	1.01	-.1	1.01	-.1
P.SD	7.0	.0	.61	.04	.35	1.6	.36	1.5
REAL RMSE	.32	TRUE SD	.52	SEPARATION	1.61	PERSON RELIABILITY		.72
ITEM	25 INPUT	25 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	232.6	61.0	.00	.20	1.00	.0	1.01	.1
P.SD	12.5	.0	.47	.02	.22	1.4	.19	1.2
REAL RMSE	.20	TRUE SD	.42	SEPARATION	2.09	ITEM RELIABILITY		.81

Figure 1. Rasch Analysis to Measure the Reliability and Validity of Psychometric Measuring Instruments

The data analysis presented in Figure 1 utilized the Rasch model to evaluate responses from 61 participants across 25 rubric items. In Rasch analysis, the term "measure" refers to the logit score, which reflects either the estimated ability level of a participant or the difficulty level of an item on a standard interval scale. The results indicated that the mean measure for participants was -0.41 logits, with a standard deviation of 0.61. This suggests that, on average, participants' abilities were slightly below the average item difficulty, which is set at 0.00 logits by default in the Rasch model. In practical terms, this means that the test items were relatively challenging for the group of participants, and their ability level can be interpreted as slightly lower than the overall difficulty of the assessment.

The Real Root Mean Square Error (RMSE) for participants was 0.32, which reflects a relatively low measurement error, indicating precision in estimating participants' ability levels. The person reliability was 0.72, which, according to Wright & Masters (1982), is considered moderate reliability, suggesting that the instrument has a fair level of consistency in distinguishing between participants with different ability levels. In comparison, the item reliability was 0.81, which falls within the "high reliability" category (Bond & Fox, 2015), indicating that the ordering of item difficulty is stable and would be reproducible with a different sample of similar ability.

Furthermore, the mean infit and outfit MNSQ values for participants were 1.01, with corresponding Z-standardized (ZSTD) values of approximately -0.1, indicating a good fit between the observed responses and the Rasch model expectations. Similarly, the item-level infit and outfit mean values were 1.00 and 1.01, with ZSTD values around 0.1, further confirming that the data adequately fit the Rasch model. These results suggest that both the participants' responses and the item characteristics are consistent with the assumptions of the Rasch model, reinforcing the validity of the analysis.

a. Test Item Scale Analysis

To improve the quality of physics education in Indonesia, it is important to understand and develop PCK in prospective teachers. PCK refers to the ability to integrate knowledge of teaching materials with effective teaching strategies, thereby facilitating better student understanding (Hubbard, 2018). Research that uses Rasch analysis to evaluate the category structure in an assessment scale shows the importance of valid and reliable measuring tools in assessing the PCK of physics prospective teachers (Putra & Narulita, 2023).

The Rasch analysis employed in this study offers in-depth insight into how prospective teachers utilize rating scales and how effectively the categories in these scales measure the desired attributes. The analysis results showed that the categories in the assessment scale, especially Category 4, exhibited the best performance, with infit and outfit values close to one, indicating appropriate thresholds and good coherence. However, Category 5 showed a relatively high discrepancy between the data and the model, indicating there is room for improvement in the development of the measuring tool.

The development of PCK in physics prospective teachers through the use of valid and reliable measurement tools has a significant impact on teaching effectiveness and student understanding. By identifying and understanding the strengths and weaknesses in the use of assessment scales, educators can design more targeted learning strategies that improve the

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

pedagogical abilities of prospective teachers. This, in turn, will result in more effective teaching, where students can better understand and grasp physics concepts in greater depth.

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
3	3	595	39	-.88	-.79	.87	.89	NONE	(-1.88)	3
4	4	619	41	-.16	-.33	.91	.96	-.61	.00	4
5	5	311	20	.00	.17	1.16	1.14	.61	(1.88)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

CATEGORY LABEL	JMLE STRUCTURE MEASURE	S.E.	SCORE-TO-MEASURE AT CAT.	----ZONE----	50% CUM. PROBABLTY	COHERENCE M->C	C->M	RMSR	ESTIM DISCR	
3	NONE		(-1.88)	-INF -1.06		79%	43%	.6527		3
4	-.61	.06	.00	-1.06 1.06	-.82	46%	86%	.3395	1.42	4
5	.61	.07	(1.88)	1.06 +INF	.82	44%	8%	1.0541	.70	5

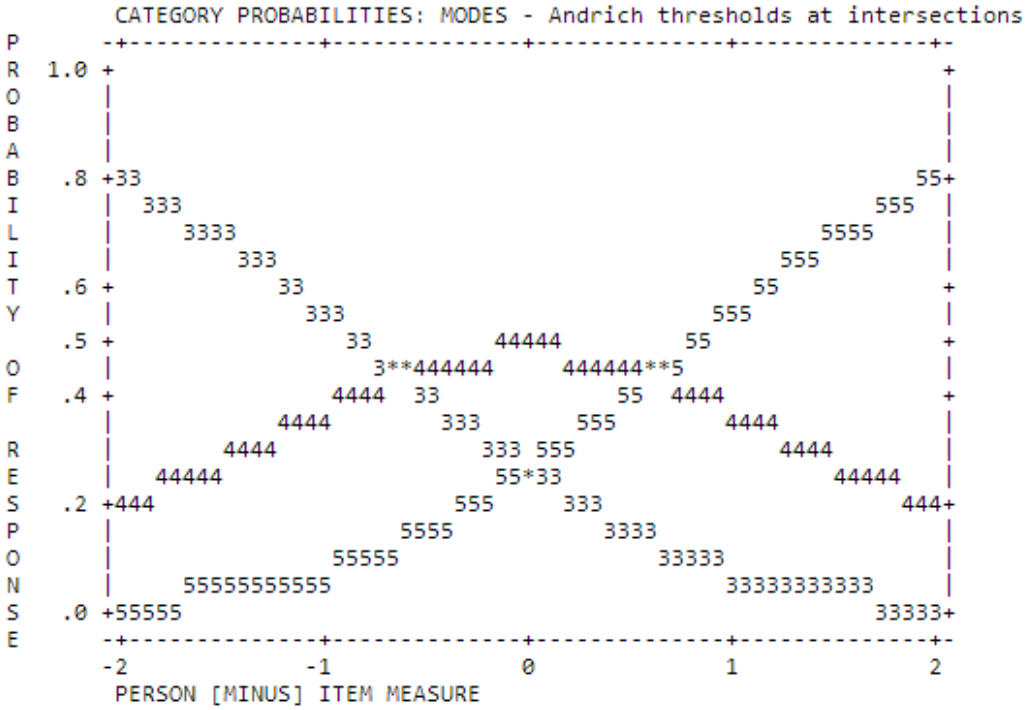

Figure 2. Rasch analysis for category structure

The data illustrated in Figure 2 comes from a Rasch analysis that focuses on evaluating the category structure in a rating scale. This analysis is crucial for understanding how respondents utilize rating scales and how effective the categories are in measuring the targeted attributes. Rasch analysis helps ensure that the rating scale is functioning optimally, identifies potential improvements in rating categories, and provides insight into how the data aligns with the model's expectations.

In this analysis, the average observed measure for Category 3 is -0.88, while the expected average is -0.79. Infit and outfit mean-square (MNSQ) were 0.87 and 0.89, respectively, indicating that the data fit the Rasch model fairly well. However, no Andrich threshold was detected for this category, indicating that the transition between this category and the other categories is not clear. The category measure for this category was -1.88, the measure-to-category coherence (M->C) is 79%, the category-to-measure coherence (C->M) is 43%, and the root mean square residual (RMSR) was 0.6527.

For Category 4, the average observed measure is -0.16 with an expected average of -0.33. The MNSQ infit and outfit were 0.91 and 0.96, respectively, indicating good agreement with the Rasch model. The Andrich threshold for this category is -0.61, and the category measure is 0.00.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Coherence M->C of 46% and C->M of 86% indicates good consistency between measures and categories. An RMSR of 0.3395 indicates a relatively small mismatch between the data and the model.

Category 5 has an average observed measure of 0.00, with an expected average of 0.17. The MNSQ infit and outfit were 1.16 and 1.14, respectively, slightly above 1, indicating some inconsistency with the Rasch model. The Andrich threshold for this category was 0.61, and the category measure was 1.88. Coherence M•C of 44% and C•M of 8% indicated low coherence, and an RMSR of 1.0541 indicated a relatively high mismatch between the data and the model.

The observed average measure is the average value observed in each category, indicating how respondents tend to assign values to that category. The expected average is the average value predicted by the Rasch model, and the difference between the observed and expected averages indicates how well the data fit the model. MNSQ infit and outfit are statistics that measure how well respondents and items fit the Rasch model, with values close to 1 indicating a good fit. Andrich thresholds are the points at which the probability of selecting a particular category is greater than that of other categories, indicating that the category is functioning well on the rating scale.

Coherence M->C and C->M indicate the percentage of consistency between measures and categories, where a high percentage indicates good coherence. RMSR is a measure of the discrepancy between the data and the model, with lower values indicating minor discrepancies. Discrimination measures how well the category differentiates between respondents with different abilities, where a high discrimination value indicates that the category functions well in differentiating respondents.

Overall, the category structure in this scale indicates that Category 4 exhibits the best performance, with infit and outfit values close to 1, indicating appropriate thresholds and good coherence. Categories 3 and 5 also demonstrate adequate performance, but Category 5 exhibits a higher RMSR and lower coherence, suggesting that there is room for improvement in this category. Improvements in Category 5 can be made by clarifying the threshold and enhancing coherence and discrimination, so that the rating scale can function more effectively in measuring the desired attributes.

This research confirms the importance of Rasch analysis in evaluating and optimizing rating scales. By understanding how categories function and how well they measure targeted attributes, researchers can make necessary adjustments to increase the validity and reliability of the scale. These results provide valuable guidance for test developers and educators in designing more effective and accurate assessment instruments (Delgado-Rebolledo & Zakaryan, 2020).

b. Quality of Video Assessment Rubric Items

To improve the quality of physics education in Indonesia, it is essential to have valid and reliable assessment instruments for prospective teachers. One method that can be used is a video assessment rubric, which allows for a comprehensive and objective evaluation of teacher candidates' abilities (Diamah et al., 2022). Based on the analysis results obtained, the instrument used demonstrates adequate performance in measuring item difficulty but requires improvement in distinguishing between participants' abilities. Video assessment rubrics can be an effective tool in this context, as they enable richer and in-depth assessment of various aspects of prospective teachers' abilities, including mastery of the material, pedagogical skills, and communication abilities (Cheah et al., 2019).

The Rasch analysis applied in this research shows that the assessment instrument has good reliability in measuring item difficulty, with a reasonably high separation index. However, to optimize the separation of prospective teachers' abilities, revisions and the addition of more varied items are needed. Video assessment rubrics can be integrated with Rasch models to produce more accurate and comprehensive assessments. By using this rubric, examiners can identify prospective teachers' strengths and weaknesses in more detail, which will assist in the development of a more effective and relevant training curriculum.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Table 2. Reliability and Separation Report of MFRM Analysis

	Logit Mean	Standard Deviation	Separation Index	Reliability	Standard Error
PERSON	-0.47	0.61	1.75	0.75	0.08
ITEM	0.00	0.47	2.2	0.83	0.1

The analysis results presented in Table 2 provide a detailed description of the participants' abilities (PERSON) and item difficulty (ITEM) as measured by the instrument used. For participants, the logit mean was -0.47, indicating that participants' average ability is slightly below the average item difficulty. The standard deviation of 0.61 reflects significant variation in ability among participants. A separation index of 1.75 indicates that this instrument is less effective in differentiating participants' abilities into different levels, with the expected ideal value being more than 2. Reliability of 0.75 indicates that although measurement consistency is quite good, there is still room for improvement. A standard error of 0.08 indicates low uncertainty in participants' mean logit measurements, indicating that the estimation of participants' ability is relatively accurate.

For the items, the logit mean was at 0.00, indicating that the mean difficulty of the items is at the expected midpoint on the logit scale. A standard deviation of 0.47 indicates that there is variation in item difficulty, which is important to ensure that the instrument can measure different levels of participant ability. With a separation index of 2.2, this instrument is highly effective in distinguishing item difficulty levels, exhibiting a clear distinction between easier and more challenging items. A reliability of 0.83 indicates that this instrument is highly consistent in measuring item difficulty, which is a positive indicator of instrument quality. A standard error of 0.1 indicates low uncertainty in the item mean logit measure, indicating that the estimate of item difficulty is relatively accurate.

Overall, this analysis demonstrates that the instrument used is quite effective in measuring item difficulty; however, there is a need for improvement in isolating participant ability. The separation index of 1.75 for participants indicates that this instrument is not capable of effectively differentiating participants with various levels of ability. To achieve more accurate and reliable measurement results, improvements in the separation of participant abilities are necessary.

Some improvement steps that can be taken include evaluating and revising existing items to ensure that each item has varying levels of difficulty and covers a range of participant abilities. Increasing the number of items with more varied difficulties can also help increase participants' separation index. In this way, the instrument will be able to differentiate participants' abilities more effectively. Additionally, further training and calibration for raters can also help ensure consistency of scoring and reduce undesirable variability. Well-trained raters can enhance accuracy and consistency in scoring, which in turn increases the overall reliability of the instrument. Overall, although this instrument has demonstrated promising results in measuring item difficulty, there is a significant opportunity for improvement in terms of the separation and reliability of measuring participant ability. By improving and optimizing this instrument, evaluation of participants' abilities can be carried out more accurately and efficiently, which will ultimately improve the quality of teaching and learning (Valtonen et al., 2019). This improvement is significant in the context of physics education, where precise measurement of participant ability and item difficulty can assist in the development of more effective teaching methods and better learning strategies (Adipat, 2021; Clausen, 2018)

Table 3. Psychometrics Attributes of Items

ENTRY NUMBER	JMLE MEASURE	MODEL S.E	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	Correlation PtMea	ITEM
24	0.96	0.23	1.51	2.42	1.37	1.64	A .47	I
8	-0.25	0.19	1.43	2.82	1.44	2.83	B .10	PRO
7	-0.28	0.19	1.4	2.65	1.42	2.72	C .10	KE
25	-0.45	0.19	1.26	1.83	1.28	1.9	D .14	HE
12	-0.18	0.19	1.21	1.48	1.18	1.23	E .49	TRS
4	-0.03	0.19	1.1	0.73	1.12	0.83	F-.01	PS

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

ENTRY NUMBER	JMLE MEASURE	MODEL S.E	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	Correlation PtMea	ITEM
1	-0.45	0.19	1.06	0.47	1.09	0.68	G .06	KT1
6	0.76	0.22	0.92	-0.39	1.05	0.33	H .25	THE
20	-0.03	0.19	1.05	0.37	1.04	0.3	I .58	MSI3
19	0.42	0.2	1.02	0.21	1.01	0.13	J .61	MSI
10	-0.84	0.19	0.99	0	1.01	0.14	K .09	SMA
14	-0.14	0.19	0.99	-0.01	0.98	-0.12	L .50	FP
15	-0.18	0.19	0.97	-0.21	0.95	-0.33	M .56	SCLI
18	-0.49	0.19	0.95	-0.31	0.97	-0.14	l-.01	LB
13	0.42	0.2	0.93	-0.4	0.91	-0.52	k .69	BE
16	-0.39	0.19	0.92	-0.56	0.9	-0.71	j .56	SCLP
21	0.07	0.19	0.92	-0.5	0.91	-0.57	i .66	EVA
22	-0.49	0.19	0.91	-0.65	0.89	-0.74	h .48	EVAP
2	-0.56	0.19	0.87	-0.93	0.86	-0.98	g .17	KT3
9	0.46	0.2	0.85	-0.95	0.87	-0.75	f .59	BW
17	0.11	0.19	0.86	-0.95	0.86	-0.94	and .64	SCLK
11	0.3	0.2	0.8	-1.37	0.82	-1.13	d .58	OR
23	-0.18	0.19	0.81	-1.42	0.8	-1.48	c .37	EVAH
5	0.76	0.22	0.63	-2.4	0.77	-1.23	b .31	DM
3	0.67	0.21	0.59	-2.82	0.72	-1.64	a .43	KM

In a review of the assessment instruments in Table 3 used to evaluate prospective physics education teachers, several items showed significant variations in performance. Item EK (serial number 24) has a JMLE measure of 0.96 with a standard error model of 0.23. The infit MNSQ value was 1.51, and the outfit MNSQ was 1.37, with ZSTDs of 2.42 and 1.64, respectively, indicating that these items may show excessive fluctuations in participants' responses, thus fitting less well with the Rasch model. The PtMea correlation of 0.47 indicates that this item is quite effective at measuring the desired construct, although further adjustments are needed.

Meanwhile, items PRO (serial number 8) and KE (serial number 7) have JMLE measures of -0.25 and -0.28, respectively, with a standard error model of 0.19. MNSQ infit and outfit values were above 1.4, respectively, with ZSTD above 2.6, indicating a significant mismatch with the Rasch model. The PtMea correlation of 0.10 for these two items indicates that they are less effective in measuring the desired construct, thus requiring in-depth evaluation and revision.

The TRS item (serial number 12) has a JMLE measure of -0.18 with a standard error model of 0.19. The infit MNSQ value is 1.21, and the outfit MNSQ is 1.18, with ZSTD of 1.48 and 1.23, respectively, indicating that although this item is still within acceptable limits, some fluctuations need to be taken into account. The PtMea correlation of 0.49 shows quite good performance in measuring the desired construct. In contrast, items such as KT1 (serial number 1) and MSI3 (serial number 20) show higher stability and consistency. The JMLE measures are -0.45 and -0.03, respectively, with MNSQ infit and outfit values below 1.1, and ZSTD below 1. PtMea correlations of 0.06 and 0.58 indicate that these two items are stable and consistent in their measurement, and are better at measuring the constructs that are desired. Additionally, items KM (serial number 3) and DM (serial number 5) have JMLE measures of 0.67 and 0.76, respectively, with MNSQ infit and outfit values below 0.72 and ZSTD values below -1.6. PtMea correlations of 0.43 and 0.31 indicate high stability and consistency of measurement, although the PtMea correlation still needs a slight improvement. Overall, these data suggest that several items, such as EK, PRO, and KE, require evaluation and revision to improve fit with the Rasch model. Meanwhile, items such as KT1, MSI3, KM, and DM show good and consistent performance in measuring the desired construct. To increase the overall reliability and validity of the instrument, adjustments or deletions of inappropriate items can be made. This effort will ensure that assessment instruments are more accurate and reliable in measuring the abilities of physics prospective teachers, ultimately contributing to the improvement of physics education in Indonesia (DURAN et al., 2021). Continuous evaluation and refinement of instruments are crucial steps in developing practical and effective assessment tools that accurately reflect the actual competence of prospective teachers (Torbeyns et al., 2020).

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

c. Results of Analysis of Teaching Videos for Physics prospective teachers

Teaching video analysis is a crucial tool for assessing the abilities of prospective physics education teachers. Using this method, one can identify and evaluate pedagogical skills and their effectiveness in teaching (B. Li et al., 2018). Based on the data analysis illustrated in Figure 3, it can be seen that item evaluation using the Rasch model provides in-depth insight into the distribution of item difficulty levels and the suitability of the items to the model. The distribution of items on the logit scale, as observed in items with varying logit values, reflects variations in the level of difficulty and the ability of prospective teachers to respond to different teaching situations (Sorge et al., 2019). Items that show high or low MNSQ infit and outfit scores indicate areas where prospective teachers may need further development or adjustments in their teaching methods.

By analyzing teaching videos, it is possible to directly observe how prospective teachers confront various teaching challenges and apply their pedagogical knowledge in practice (Mayer & Girwidz, 2019). This enables the provision of more specific and relevant feedback, as well as identifying areas that require improvement. In addition, this analysis also helps assess the consistency and reliability of prospective teachers' abilities, ensuring they are ready to face diverse teaching situations in real classrooms. The use of the Rasch model in teaching video analysis helps strengthen the validity and reliability of assessments, ensuring that the evaluation of prospective physics education teachers' abilities is carried out objectively and comprehensively (Yusup, 2021).

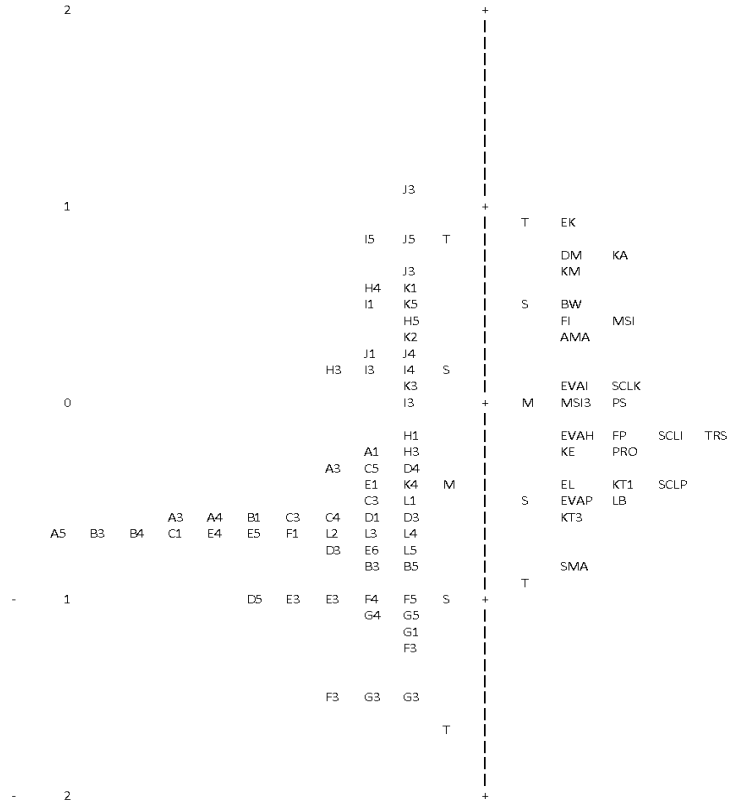

Figure 3. Results of Analysis of Teaching Videos for Physics prospective teachers

The data analysis presented in Figure 3 provides important insights into the distribution of items on the logit scale and the degree to which they fit the Rasch model. This distribution describes how items are evaluated based on their level of difficulty and how well they fit the Rasch model. Items located near the 0 point on the logit scale, such as BW, MSI, AMA, EVAI, SCLK, MSI3, PS, EVAH, FP, SCLI, TRS, KE, and PRO, indicate that they have a balanced average level of difficulty and are easier for examinees to understand. Point-measure correlations (PtMea) for these items were

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

quite variable, indicating that some items may be more effective in measuring the construct of interest than others.

Items such as EK(T), DM, KA, and KM are at higher positions on the logit scale, indicating that they are more difficult than other items. High MNSQ infit and outfit values for some of these items, such as EK having an infit value of 1.51, indicate that they may not fit the Rasch model and require further evaluation. In contrast, items such as BW and MSI are located lower on the logit scale, indicating that they are relatively easier for examinees.

Items such as KT3, EL, KT1, SCLP, EVAP, and LB, which are located around the -1 point on the logit scale, tend to be easier than the other items. The low MNSQ infit and outfit values for these items indicate that they are more stable and consistent in their measurement. This stability and consistency are important indicators that the items function well in measuring the desired construct. Overall, these data indicate significant variation in item difficulty, with some items demonstrating inconsistency with the Rasch model and requiring further evaluation. Items that are more difficult or easier need to be adjusted to ensure that the assessment instrument can consistently and validly measure the desired construct. Items that show MNSQ scores that are too high or too low may need to be revised or deleted to improve the overall reliability and validity of the instrument.

Adapting these items can help ensure that the assessment instrument measures participants' abilities more accurately and reliably. Steps that can be taken include re-evaluating and revising items that do not fit, adding new items with varying levels of difficulty, and further calibration to improve the fit with the Rasch model. Thus, although several items require improvement, overall, this instrument has the potential to provide a valid and reliable measure of examinee abilities. This evaluation and adjustment process is an important part of developing effective and appropriate assessment instruments.

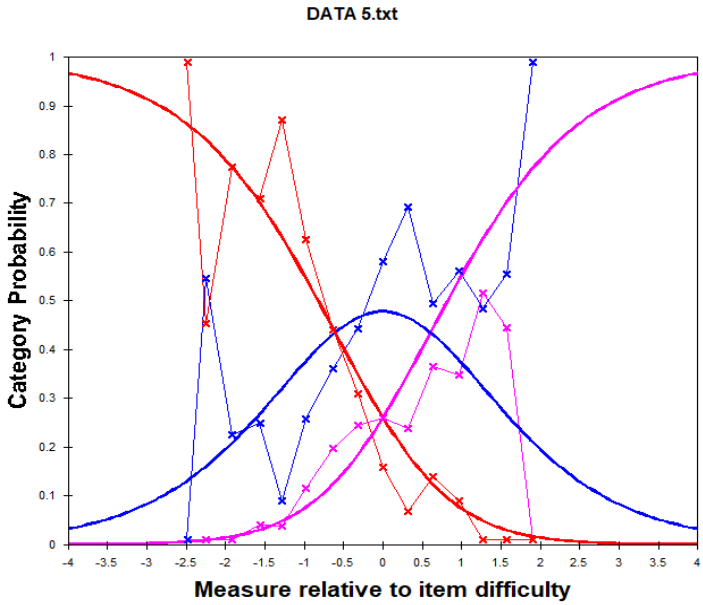

Figure 4. Probability curve for categories and ability levels on the logit scale

Figure 4 presents a graph of category probability curves, which maps the relationship between participants' ability level and item difficulty on a logit scale. The horizontal axis shows "Measure relative to item difficulty" or ability relative to item difficulty, while the vertical axis shows "Category Probability" or category probability. The different curves in this graph illustrate the probability of selecting a particular category based on the participant's ability level.

From the graph, several important points can be identified.

1. Red Category: The red curve represents the category with the lowest level of difficulty (for example, Category 1). This curve shows the highest probability for participants with very low ability (logit around -4 to -1). This shows that participants with low ability tend to choose this category. However, this probability decreases sharply as the participant's ability increases.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

2. Blue Category: The blue curve represents the intermediate categories (e.g., categories 2 and 3). This curve peaks at a logit of -0.5 to 1.5, indicating that participants with moderate ability are more likely to select this category. This probability increases as ability increases and then decreases again after passing the peak.
3. Purple Category: The purple curve depicts the category with the highest level of difficulty (for example, Category 4). This curve shows a significant increase in probability at logit 1.5 and above, indicating that high-ability participants are more likely to choose this category.
4. Category Transitions: The intersection points between curves indicate transition points where the probability of selecting a particular category changes. For example, the intersection point between the red and blue curves indicates the level of ability at which participants have the same probability of choosing the low or medium category.

Overall, this graph indicates that the instrument used has well-defined categories, enabling the items to differentiate effectively between different levels of participant ability. However, the existence of overlapping probabilities in some categories indicates the need for further evaluation and adjustment to ensure that each category is clearly defined. This analysis serves as a crucial tool for evaluating and enhancing the validity and reliability of measurement instruments by refining categories to more accurately reflect differences in participant abilities.

Through analysis of these category probability curves, instrument developers can identify areas that require revision to improve measurement accuracy and reliability. For example, if there are categories that have a significant probability of overlap, item revisions or rating scale adjustments can be made to clarify the definitions of those categories. Thus, this graph not only provides an overview of how participants interact with the items in the instrument but also provides practical guidance for further improvement and development, ensuring that the instrument can measure participants' abilities more effectively and reliably (Syahmani et al., 2021).

2. Discussion

The results of the analysis presented provide in-depth insight into the quality of assessment instruments and their impact on teaching effectiveness and student understanding. The results of the Rasch analysis indicated that the average participant's ability was slightly below the average item difficulty, with the standard deviation indicating significant variation in ability. Although the reliability of participant measurements is at an adequate level, namely 0.75, the separation index of 1.75 indicated the instrument's limitations in effectively differentiating participants based on ability. Increasing the separation index is needed to optimize the instrument's ability to measure differences in participant ability levels more accurately.

On the other hand, analysis of individual items showed variations in fit to the Rasch model. Items such as EK and PRO showed poor fit with the Rasch model, with high MNSQ infit and outfit values, indicating fluctuations in participants' responses that may indicate inconsistencies in measurement. Meanwhile, items with JMLE measures and MNSQ infit/outfit values close to 1, such as KT1 and MSI3, exhibit more stable and consistent performance. Evaluation and revision of inappropriate items, as well as adjustment of the rating scale, can improve the reliability and validity of the instrument, which ultimately affects the effectiveness of measuring PCK of prospective teachers.

Figure 4 shows a category probability curve that illustrates how the probability of category selection changes with variations in participant ability. The curve shows categories with low, medium, and high difficulty, indicating that the instrument has a reasonably good ability to differentiate participants with different levels of ability. However, the overlap in probabilities across some categories suggests the need for further review to ensure categories are clearly defined. These adjustments are crucial for enhancing the instrument's effectiveness in measuring PCK and, simultaneously, supporting the development of pedagogical competencies in physics prospective teachers.

PCK is a key component in determining how prospective teachers translate content knowledge into effective teaching practices (Cetin-Dindar et al., 2018). Assessment instruments that do not fully comply with the Rasch model can harm prospective teachers' understanding of PCK. The instrument's inability to differentiate between participants' ability levels may result in an inaccurate assessment of their pedagogical competence. This can influence prospective teachers' readiness to implement effective teaching strategies and understand the needs and learning styles of their students (C.-J. Wang, 2019).

In the context of physics education, a deep understanding of PCK allows prospective teachers to develop better and more adaptive teaching methods (König et al., 2018). Therefore, improvements to assessment

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

instruments should focus on enhancing the instrument's ability to differentiate between participants' ability levels and clarify assessment categories (Aryan et al., 2023). These steps will ensure that prospective teachers receive more accurate feedback about their pedagogical competencies, which in turn can improve the quality of their teaching (Karousiou et al., 2019). With more reliable assessment instruments, prospective teachers can be more effective in designing and implementing teaching strategies that improve students' understanding and their overall learning outcomes (Rieu et al., 2022).

This research evaluates pedagogical content knowledge (PCK) in prospective physics education teachers in Indonesia through comprehensive quantitative and qualitative analysis using the Rasch model. The analysis results show that the assessment instrument has several significant strengths and weaknesses. In general, the data showed that the average participant ability was slightly below the average item difficulty, with the separation index indicating limitations in effectively differentiating participant ability levels. Although the reliability of participant measurements is at a reasonably good level (0.75), improvements are needed in the separation index to achieve more accurate and reliable results. Analysis of the assessment items revealed variation in their fit to the Rasch model, with some items, such as EK and PRO, exhibiting significant discrepancies that necessitated further evaluation. Other items, such as the KT1 and MSI3, show better consistency, but there is still room for improvement. The category probability curves depicted in Figure 4 illustrate that the categories in the rating scale require adjustment to ensure clear definitions and reduce overlapping probabilities, thereby enhancing the instrument's effectiveness in differentiating between participants with varying ability levels. This study's findings are consistent with various validation efforts of similar instruments in both Indonesian and international contexts. For example, research by Nurulsari et al. (2020) and Kholili et al. (2024) in Indonesia reported comparable item reliability values when using the Rasch model to validate teacher competency rubrics. However, their person separation indices were generally higher, indicating a stronger ability to differentiate participant ability levels. International studies, such as those by Eckes (2019) and Wang et al. (2020), also reported challenges in achieving high person reliability in rater-mediated performance assessments, particularly when employing video-based formats. This convergence suggests that the psychometric limitations identified in the present study, specifically moderate person reliability and a limited separation index, are consistent with the challenges faced in similar validation studies. This reinforces the need for ongoing refinement of rubric descriptors, category structures, and item difficulty ranges (Wati Sukmawati, 2019). These findings highlight the importance of ongoing evaluation and revision of assessment instruments to ensure validity and reliability in measuring prospective teachers' PCK. More reliable instruments would provide more accurate feedback regarding prospective teachers' pedagogical competencies, which could ultimately improve the quality of teaching and student understanding. The suggested adjustments, including item revisions and clarification of assessment categories, are expected to improve measurement effectiveness and support the development of better PCK among prospective physics.

CONCLUSION

This study aims to evaluate the validity and reliability of a video-based assessment rubric designed to measure the pedagogical content knowledge (PCK) of physics prospective teachers in Indonesia, using the Rasch model as the primary analytical framework. The analysis revealed that the instrument demonstrated acceptable measurement quality, particularly in terms of item difficulty calibration. The item reliability value of 0.81 indicates that the items are consistent and capable of distinguishing varying levels of item difficulty.

However, the person reliability score of 0.72 and the separation index of 1.75 suggest that while the instrument can measure participant ability with moderate consistency, its capacity to differentiate among participants across different ability levels remains limited. This finding suggests that certain items may require revision, either by expanding the range of item difficulty or by refining the rubric descriptors, to enhance measurement precision.

Overall, the findings confirm that the Rasch model is a robust method for evaluating the psychometric quality of assessment instruments. The refined rubric is expected to provide more accurate, fair, and reliable evaluations, supporting the professional development of physics prospective teachers within the context of teacher education programs in Indonesia.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

REFERENCE

Adhantoro, M. S., Gunawan, D., Prayitno, H. J., Riyanti, R. F., & Jufriansah, A. (2024). Strategies to Enhance Literacy and Access to Muhammadiyah Information through ChatMu Innovation. *International Journal of Religion*, *5*(11), 2503–2520. https://doi.org/10.61707/7hqaep83

Adipat, S. (2021). Developing Technological Pedagogical Content Knowledge (TPACK) through Technology-Enhanced Content and Language-Integrated Learning (T-CLIL) Instruction. *Education and Information Technologies*, *26*(5), 6461–6477. https://doi.org/10.1007/s10639-021-10648-3

Agricola, B. T., van der Schaaf, M. F., Prins, F. J., & van Tartwijk, J. (2022). The development of research supervisors' pedagogical content knowledge in a lesson study project. *Educational Action Research*, *30*(2), 261–280. https://doi.org/10.1080/09650792.2020.1832551

Akyuz, D. (2018). Measuring technological pedagogical content knowledge (TPACK) through performance assessment. *Computers & Education*, *125*, 212–225. https://doi.org/10.1016/j.compedu.2018.06.012

Aryan, Hegade, P., & Shettar, A. (2023). Effectiveness of Computational Thinking in Problem Based Learning. *Journal of Engineering Education Transformations*, *36*(S2), 179–185. https://doi.org/10.16920/jeet/2023/v36is2/23025

Bai, H., Wang, X., & Zhao, L. (2021). Effects of the Problem-Oriented Learning Model on Middle School Students' Computational Thinking Skills in a Python Course. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.771221

Bayram-Jacobs, D., Henze, I., Evagorou, M., Schwartz, Y., Aschim, E. L., Alcaraz-Dominguez, S., Barajas, M., & Dagan, E. (2019). Science teachers' pedagogical content knowledge development during enactment of socioscientific curriculum materials. *Journal of Research in Science Teaching*, *56*(9), 1207–1233. https://doi.org/10.1002/tea.21550

Cetin-Dindar, A., Boz, Y., Yildiran Sonmez, D., & Demirci Celep, N. (2018). Development of pre-service chemistry teachers' technological pedagogical content knowledge. *Chemistry Education Research and Practice*, *19*(1), 167–183. https://doi.org/10.1039/C7RP00175D

Cheah, Y. H., Chai, C. S., & Toh, Y. (2019). Traversing the context of professional learning communities: development and implementation of Technological Pedagogical Content Knowledge of a primary science teacher. *Research in Science & Technological Education*, *37*(2), 147–167. https://doi.org/10.1080/02635143.2018.1504765

Clausen, S. W. (2018). Exploring the pedagogical content knowledge of Danish geography teachers: teaching weather formation and climate change. *International Research in Geographical and Environmental Education*, *27*(3), 267–280. https://doi.org/10.1080/10382046.2017.1349376

Delgado-Rebolledo, R., & Zakaryan, D. (2020). Relationships Between the Knowledge of Practices in Mathematics and the Pedagogical Content Knowledge of a Mathematics Lecturer. *International Journal of Science and Mathematics Education*, *18*(3), 567–587. https://doi.org/10.1007/s10763-019-09977-0

Diamah, A., Rahmawati, Y., Paristiowati, M., Fitriani, E., Irwanto, I., Dobson, S., & Sevilla, D. (2022). Evaluating the effectiveness of technological pedagogical content knowledge-based training program in enhancing pre-service teachers' perceptions of technological pedagogical content knowledge. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.897447

DURAN, M., USAK, M., HSIEH, M.-Y., & UYGUN, H. (2021). A New Perspective on Pedagogical Content Knowledge: Intellectual and Emotional Characteristics of Science Teachers. *Journal of Research and Social Intervention*, *72*, 9–32. https://doi.org/10.33788/rcis.72.1

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In *Quantitative Data Analysis for Language Assessment Volume I* (pp. 153–175). Routledge. https://doi.org/10.4324/9781315187815-8

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In *Quantitative data analysis for language assessment Volume I* (pp. 153–175). Routledge. https://doi.org/10.4324/9781315187815-8

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Ettekal, I., & Shi, Q. (2020). Developmental trajectories of teacher-student relationships and longitudinal associations with children's conduct problems from Grades 1 to 12. *Journal of School Psychology*, *82*, 17–35. https://doi.org/10.1016/j.jsp.2020.07.004

Fitriansyah, R., Fatinah, L., & Syahril, M. (2020). Critical Review: Professional Development Programs to Face Open Educational Resources in Indonesia. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *2*(2), 109–119. https://doi.org/10.23917/ijolae.v2i2.9662

Hsu, L., & Chen, Y.-J. (2019). Examining teachers' technological pedagogical and content knowledge in the era of cloud pedagogy. *South African Journal of Education*, *39*(S2), 1–13. https://doi.org/10.15700/saje.v39ns2a1572

Hubbard, A. (2018). Pedagogical content knowledge in computing education: a review of the research literature. *Computer Science Education*, *28*(2), 117–135. https://doi.org/10.1080/08993408.2018.1509580

Hursen, C. (2021). The Effect of Problem-Based Learning Method Supported by Web 2.0 Tools on Academic Achievement and Critical Thinking Skills in Teacher Education. *Technology, Knowledge and Learning*, *26*(3), 515–533. https://doi.org/10.1007/s10758-020-09458-2

James, F., & Augustin, D. S. (2018). Improving teachers' pedagogical and instructional practice through action research: potential and problems. *Educational Action Research*, *26*(2), 333–348. https://doi.org/10.1080/09650792.2017.1332655

Karousiou, C., Hajisoteriou, C., & Angelides, P. (2019). Teachers' professional identity in super-diverse school settings: teachers as agents of intercultural education. *Teachers and Teaching*, *25*(2), 240–258. https://doi.org/10.1080/13540602.2018.1544121

Kholili, M. I., Dewantoro, A., Surur, N., & Hapsari, N. T. (2024). The 21st-century skills scales: many facet Rasch measurements. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(3), 1424. https://doi.org/10.11591/ijere.v13i3.26651

Kholili, M. I., Dewantoro, A., Surur, N., & Hapsari, N. T. (2024). The 21st-century skills scales: Many facet Rasch measurements. *International Journal of Evaluation and Research in Education*, *13*(3), 1424. https://doi.org/10.11591/ijere.v13i3.26651

König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., Strauß, S., Bremerich-Vos, A., Flade, I., & Kaiser, G. (2018). Pedagogical knowledge versus didactic knowledge? *Journal of Educational Science*, *21*(3), 1–38. https://doi.org/10.1007/s11618-017-0765-z

Larson, K. E., Hirsch, S. E., McGraw, J. P., & Bradshaw, C. P. (2020). Preparing Preservice Teachers to Manage Behavior Problems in the Classroom: The Feasibility and Acceptability of Using a Mixed-Reality Simulator. *Journal of Special Education Technology*, *35*(2), 63–75. https://doi.org/10.1177/0162643419836415

LEE, J., KIM, J. B., & KIM*, J. B. (2018). Effects of the Experience in Developing Physics Teaching Materials Based on Computational Thinking for Improvement of Science Teachers' and Pre-service Teachers' Technological Pedagogical and Content Knowledge (TPACK). *New Physics: Sae Mulli*, *68*(2), 202–216. https://doi.org/10.3938/NPSM.68.202

Li, B., Zhao, Y., & Zhang, H. (2018). Video Analysis of the Influence of Intelligent Media Application on Teachers' Knowledge Structure: A Case Study of Physics Lesson at Middle School. *2018 International Joint Conference on Information, Media and Engineering (ICIME)*, 249–254. https://doi.org/10.1109/ICIME.2018.00059

Li, J., Shi, Z., & Xue, E. (2020). The problems, needs and strategies of rural teacher development at deep poverty areas in China: Rural schooling stakeholder perspectives. *International Journal of Educational Research*, *99*, 101496. https://doi.org/10.1016/j.ijer.2019.101496

Nurhasanah, M., Suprapto, P. K., & Ardiansyah, R. (2024). The effectiveness of problem-based learning assisted by Articulate Storyline interactive students' critical thinking skills. *Jurnal Inovasi Pendidikan IPA*, *10*(1), 1–12. https://doi.org/10.21831/jipi.v10i1.64847

Mayer, P., & Girwidz, R. (2019). Physics Teachers' Acceptance of Multimedia Applications—Adaptation of the Technology Acceptance Model to Investigate the Influence of TPACK on Physics Teachers' Acceptance Behavior of Multimedia Applications. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00073

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Melo, L., Cañada-Cañada, F., González-Gómez, D., & Jeong, J. S. (2020). Exploring Pedagogical Content Knowledge (PCK) of Physics Teachers in a Colombian Secondary School. *Education Sciences*, *10*(12), 362. https://doi.org/10.3390/educsci10120362

Metz, M. (2021). Pedagogical Content Knowledge for Teaching Critical Language Awareness: The Importance of Valuing Student Knowledge. *Urban Education*, *56*(9), 1456–1484. https://doi.org/10.1177/0042085918756714

Nurulsari, N., Abdurrahman, Maulina, H., Sukamto, I., & Umam, R. (2020). Exploring the Prospective of Pre-Service Physics Teacher's Pedagogical Content Knowledge: A Case Study. *Journal of Physics: Conference Series*, *1467*(1), 012023. https://doi.org/10.1088/1742-6596/1467/1/012023

Nurulsari, N., Abdurrahman, Maulina, H., Sukamto, I., & Umam, R. (2020). Exploring the prospective of pre-service physics teacher's pedagogical content knowledge: A case study. *Journal of Physics: Conference Series*, *1467*(1), 012023. https://doi.org/10.1088/1742-6596/1467/1/012023

Putra, P. D. A., & Narulita, E. (2023). *Teacher professional knowledge: The implementation of STEM pedagogical content knowledge in pandemic era*. 060014. https://doi.org/10.1063/5.0111357

Rieu, A., Leuders, T., & Loibl, K. (2022). Teachers' diagnostic judgments on tasks as information processing – The role of pedagogical content knowledge for task diagnosis. *Teaching and Teacher Education*, *111*, 103621. https://doi.org/10.1016/j.tate.2021.103621

SA, N. H., Suyanto, S., Arifi, A., Putranta, H., & Azizah, A. N. M. (2021). Experiences of Participants in Teacher Professional Education on Obtaining Soft Skills: A Case Study in Indonesia. *European Journal of Educational Research*, *volume-10-2021*(volume-10-issue-1-january-2021), 313–325. https://doi.org/10.12973/eu-jer.10.1.313

Scherer, R., Tondeur, J., Siddiq, F., & Baran, E. (2018). The importance of attitudes toward technology for pre-service teachers' technological, pedagogical, and content knowledge: Comparing structural equation modeling approaches. *Computers in Human Behavior*, *80*, 67–80. https://doi.org/10.1016/j.chb.2017.11.003

Sorge, S., Kröger, J., Petersen, S., & Neumann, K. (2019). Structure and development of physics prospective teachers' professional knowledge. *International Journal of Science Education*, *41*(7), 862–889. https://doi.org/10.1080/09500693.2017.1346326

Splett, J. W., Garzona, M., Gibson, N., Wojtalewicz, D., Raborn, A., & Reinke, W. M. (2019). Teacher Recognition, Concern, and Referral of Children's Internalizing and Externalizing Behavior Problems. *School Mental Health*, *11*(2), 228–239. https://doi.org/10.1007/s12310-018-09303-z

Styck, K. M., Anthony, C. J., Flavin, A., Riddle, D., & LaBelle, B. (2021). Are ratings in the eye of the beholder? A non-technical primer on many facet Rasch measurement to evaluate rater effects on teacher behavior rating scales. *Journal of School Psychology*, *86*, 198–221. https://doi.org/10.1016/j.jsp.2021.01.001

Sudirman, S. (2019). The 21st-Century Teacher: Teacher's Competence Within the Character Education Framework Towards A Cultural-Oriented Development and Promoting Tolerance. *International Education Studies*, *12*(8), 21–25. https://doi.org/10.5539/ies.v12n8p21

Syahmani, S., Hafizah, E., Sauqina, S., Adnan, M. Bin, & Ibrahim, M. H. (2021). STEAM Approach to Improve Environmental Education Innovation and Literacy in Waste Management: Bibliometric Research. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *3*(2), 130–141. https://doi.org/10.23917/ijolae.v3i2.12782

Torbeyns, J., Verbruggen, S., & Depaepe, F. (2020). Pedagogical content knowledge in preservice preschool teachers and its association with opportunities to learn during teacher training. *ZDM*, *52*(2), 269–280. https://doi.org/10.1007/s11858-019-01088-y

Valtonen, T., Sointu, E., Kukkonen, J., Mäkitalo, K., Hoang, N., Häkkinen, P., Järvelä, S., Näykki, P., Virtanen, A., Pöntinen, S., Kostiainen, E., & Tondeur, J. (2019). Examining pre-service teachers' Technological Pedagogical Content Knowledge as evolving knowledge domains: A longitudinal approach. *Journal of Computer Assisted Learning*, *35*(4), 491–502. https://doi.org/10.1111/jcal.12353

Wang, C.-J. (2019). Facilitating the emotional intelligence development of students: Use of technological pedagogical content knowledge (TPACK). *Journal of Hospitality, Leisure, Sport & Tourism Education*, *25*, 100198. https://doi.org/10.1016/j.jhlste.2019.100198

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Wang, P., Coetzee, K., Strachan, A., Monteiro, S., & Cheng, L. (2020). Examining Rater Performance on the CELBAN Speaking: A Many-Facets Rasch Measurement Analysis. *Canadian Journal of Applied Linguistics*, *23*(2), 73–95. https://doi.org/10.37213/cjal.2020.30436

Wang, P., Coetzee, K., Strachan, A., Monteiro, S., & Cheng, L. (2020). Examining rater performance on the CELBAN Speaking: A many-facets Rasch measurement analysis. *Canadian Journal of Applied Linguistics*, *23*(2), 73–95. https://doi.org/10.37213/cjal.2020.30436

Wati, S. (2019). Psychometric validation of a video-based rubric for assessing pedagogical and professional competence of physics teacher candidates using Rasch analysis. *Jurnal Inovasi Pendidikan IPA*, *5*(2), 234–245. https://doi.org/10.21831/jipi.v5i2.12345

Yusup, M. (2021). Using Rasch model for the development and validation of energy literacy assessment instrument for physics prospective teachers. *Journal of Physics: Conference Series*, *1876*(1), 012056. https://doi.org/10.1088/1742-6596/1876/1/012056

Zeller, J., Schiering, D., Kulgemeyer, C., Neumann, K., Riese, J., & Sorge, S. (2024). Cross-project empirical and criteria-oriented analysis of physics prospective teachers' pedagogical content knowledge: What content structures emerge in the context of different models? *Teaching science*. https://doi.org/10.1007/s42010-024-00200-w

ISSN 2406-9205 (print), ISSN 2477-4820 (online)