International Journal of Applied Research and Sustainable Sciences (IJARSS)
Vol. 4 No. 3 , 2026: 245 - 254

Characteristics of the History of Indonesian National Struggle test

Suwarto Suwarto[1]*, Farida Nugrahani[2], Singgih Subiyantoro[3]
Universitas Veteran Bangun Nusantara
Corresponding Author: Suwarto Suwarto suwartowarto@yahoo.com

A R T I C L E I N F O

Keywords: Rasch Model, Quest Program, Item Difficulty, Person Ability

Received : 25 December 2025
Revised : 25 February 2026
Accepted: 25 March 2026

©2026. Suwarto, Nugrahani, Subiyantoro: This is an open-access article distributed under the terms of the Creative Commons Atribusi 4.0 Internasional.

A B S T R A C T

This descriptive–exploratory quantitative research analyzes the psychometric characteristics of the Indonesian National History test using the Rasch model. Involving 200 respondents from urban-rural transition areas, data were analyzed using Quest to test reliability and validity via Infit MNSQ criteria. Results show high reliability and consistent logical validity. However, Wright Map visualization reveals a significant gap: average person ability (1.47 logits) far exceeds item difficulty (0.00 logits). The presence of misfit items indicates unexpected response patterns due to differing educational backgrounds. Overall, the test material is categorized as easy, necessitating more challenging instruments to measure the spectrum of participant ability more proportionally and objectively.

DOI: https://doi.org/10.59890/ijarss.v4i3.221
E – ISSN : 3026-0205
https://jpnmultitechpublisher.my.id/index.php/ijarss/index

INTRODUCTION

Measuring the younger generation's understanding of national history is a logical step to ensure the sustainability of national identity amidst the increasingly blurring cultural boundaries of globalization (Ricklefs, 1993). Phenomenologically, the History of the Indonesian Nation's Struggle test is often viewed merely as an administrative formality in the competency selection process, thus the essence of patriotism within it is often overlooked in technical analysis (Suwarto et al., 2023). This study contributes to the enrichment of knowledge by conducting an in-depth evaluation of the instrument's psychometric characteristics, an area rarely explored compared to studies of purely historical content (Azwar, 2016).

The novelty of this study lies in the use of a sample of selection participants from an urban-rural transition region, which provides a unique insight into the distribution of item difficulty levels across specific demographic groups. Through the Rasch Model approach operated by the Quest program, this study enriches the theory of educational evaluation by proving the empirical validity of the instrument through item fit analysis and ensuring that the test not only tests factual memorization but also the depth of historical understanding (Adam & Khoo, 1996; Bond, 2015). Therefore, this study aims to analyze the parameters of item difficulty level and participant ability level on the same scale to map the effectiveness of the instrument in measuring respondents' historical competencies and to test the internal reliability of the test in providing objective measurement results for the development of evaluation instruments in the future.

LITERATURE REVIEW

Quality test characteristics are an absolute prerequisite for ensuring the accuracy of educational evaluations, particularly in the affective-cognitive domain, such as the history of the nation's struggle. According to Mukhlis (Mukhlis, 2021), historical understanding is not merely a mastery of chronological facts, but rather a strategic instrument in shaping national character and identity in an era of disruption. Therefore, the instruments used must have a high level of validity and reliability to distinguish between participants with in-depth understanding and those who rely solely on memorization (Azizah, 2023). Field observations indicate that many historical assessment instruments have not undergone rigorous item characteristic testing, thus their effectiveness in measuring student competency is often questioned (Mukhlis, 2021).

In modern psychometric studies, the analysis of test characteristics has evolved from Classical Test Theory to an objective measurement model using the Rasch Model, which is operated with the Quest program. Adam & Khoo, (1996) explained that the Quest program allows the analysis of item difficulty and participant ability parameters on the same logit scale. The use of the Rasch Model in evaluating history tests provides the advantage of identifying item fit through Infit and Outfit Mean Square statistics, thus ensuring that each item consistently measures the desired dimensions (Bond, 2015). Through variable mapping (Variable Map), researchers can accurately see the match between the level of

difficulty of history questions and the spectrum of respondents' abilities. The integration of the nation's struggle values with the Rasch Model-based evaluation methodology is crucial for producing credible and dignified national assessment standards.

METHODOLOGY

This study employs a quantitative approach with a descriptive-exploratory design to analyze the psychometric characteristics of the History of the Indonesian Nation's Struggle test instrument. The population in this study includes 1,500 first-semester students of Universitas Veteran Bangun Nusantara originating from various regions, both urban and rural. This group was selected as they represent a diverse range of historical information access and unique educational backgrounds within an urban-rural transition zone (*niche sample*).

The sampling technique was conducted through *purposive sampling*, resulting in 200 respondents. This technique was applied to ensure the representation of respondents with balanced demographic characteristics between urban and rural areas. Primary data were collected by documenting participants' responses to a multiple-choice test instrument covering national history, from the national awakening movement to independence.

Data analysis was performed using the Rasch model approach assisted by the Quest program. The Rasch model was chosen for its ability to equate item difficulty parameters with person ability levels on the same logit scale(Adam & Khoo, 1996). The analysis stages include: (1) Parameter Estimation: Calculating item difficulty levels and participant abilities. (2) Item Quality Evaluation: Assessing item-model fit based on the Infit Mean Square (Infit MNSQ) statistical criteria, with an acceptable range between 0.77 and 1.30 (Bond, 2015). (3) Reliability Testing: Measuring instrument reliability through item reliability and person reliability values. (4) Data Visualization: Mapping analysis results through a *Variable Map* (Wright Map) to review the extent to which the distribution of item difficulty levels proportionally measures the ability spectrum of the 200 participants.

RESEARCH RESULT

1. Statistical Description of the Rasch Model

Reliability estimation is a crucial aspect of test analysis, as it demonstrates the instrument's consistency and stability in measuring the intended construct. In the context of educational assessment, a reliable instrument ensures that the results obtained are reliable and can be interpreted with confidence. Within the Rasch measurement framework, reliability is examined from two perspectives: individual reliability and item reliability, which reflect the consistency of respondent performance and the quality of the test items, respectively. To interpret the reliability values obtained from the analysis, specific criteria are required. Therefore, this section refers to the reliability criteria proposed by Sumintono, (2014) as presented in Table 1, which outlines the standards for evaluating individual reliability and item reliability.

Table 1. Person reliability criteria and item reliability

Reliability score	Category
<0.667	Weak
0.67-0.80	Average
0.81-0.90	Good
0.91-0.94	Very good
>0.94	Excellent

An analysis of 200 respondents showed that the History of the Indonesian Nation's Struggle test instrument had stable psychometric qualities. Based on the Quest program output, a person reliability score of 0.92 and an item reliability score of 0.87 were obtained. This indicates excellent consistency of participant responses and a good quality of item difficulty sequencing within the instrument.

Table 2. Summary of Reliability and Separation Statistics (n=200)

Indicator	Statistical Value	declaration
Person Reliability	0.92	Very good
Item Reliability	0.87	Good
Mean Person Ability	0.47 logit 0.00 logit	Average ability is slightly above average on questions
Mean Item Difficulty		Default value of difficulty level parameter

Table 2 shows the overall quality of the instrument in measuring participant abilities.

2. Evaluation of Model Fit (Item Suitability)

Item fit evaluation is a crucial step in Rasch model analysis to determine whether each test item functions as intended in measuring the intended construct. Item fit analysis helps identify items that are consistent with the model as well as items that may distort the measurement due to unexpected response patterns. One commonly used statistic to assess item fit is the Mean Infit Square (MNSQ) value, which is sensitive to response patterns from respondents whose ability level is close to the item's difficulty. By examining the MNSQ Infit value, researchers can assess the extent to which each item fits the Rasch model's assumptions. To determine the fit of each item, the obtained Mean Infit Square value is compared with predetermined criteria. Therefore, this study refers to the Mean Infit Square criteria proposed by Setyawarno, (2017) as presented in Table 3, which provides guidelines for evaluating item fit within the Rasch measurement framework.

Table 3. Infit mean square criteria.

Infit MNSQ	Interpretation
>1.33	Misfit

0.77-1.33	Fit
<0.77	Misfit

The results of the item analysis are summarized in Table 4. Based on the analysis using the Mean Squared Infit (MNSQ) value, it is clear that 12 items (items 4-6, 11, 12, 14-16, 19, 23, 29, and 30) are within the acceptable fit range of 0.77 to 1.33. There are 18 items (items 1-3, 7-10, 13, 17, 18, 20-22, and 24-28) that require attention. Therefore, 18 items need improvement.

Table 4. Infit mean square estimation results

Items	Infit MNSQ	Criterion Infit MNSQ	Items	Infit MNSQ	Criterion Infit MNSQ	Items	Infit MNSQ	Criterion Infit MNSQ
1	2.20	Misfit	11	1.13	Fit	21	0.75	Misfit
2	1.70	Misfit	12	1.01	Fit	22	1.40	Misfit
3	1.44	Misfit	13	0.60	Misfit	23	0.90	Fit
4	1.24	Fit	14	0.92	Fit	24	0.65	Misfit
5	1.00	Fit	15	0.90	Fit	25	0.63	Misfit
6	0.84	Fit	16	1.11	Fit	26	0.69	Misfit
7	0.64	Misfit	17	0.51	Misfit	27	0.61	Misfit
8	0.59	Misfit	18	0.40	Misfit	28	1.46	Misfit
9	0.39	Misfit	19	1.25	Fit	29	1.10	Fit
10	1.52	Misfit	20	0.64	Misfit	30	1.11	Fit

This analysis can be more clearly understood through a visual representation, as presented in Figure 1, which was generated using the Quest program. The figure provides a comprehensive illustration of the distribution of item difficulty and respondent ability, allowing for a clearer interpretation of how well the test items align with the ability levels of the students. By examining this figure, readers can identify patterns related to item fit, the spread of item difficulty, and the consistency of respondents' performance across the measurement scale. Consequently, Figure 1 serves as an important supporting component of the Rasch analysis, as it helps to strengthen the interpretation of the statistical results obtained from the Quest software (Adam & Khoo, 1996).

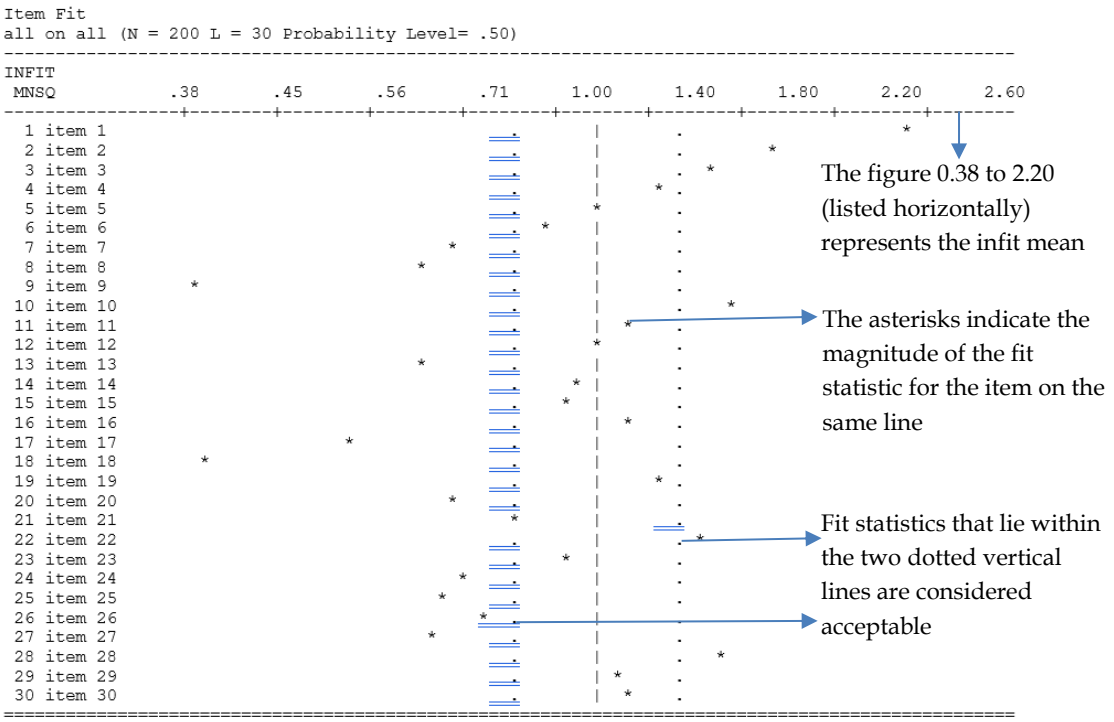


Figure 1. Item fit map for the History of Indonesian National Struggle test

3. Item Difficulty Level and Participant Ability

The estimation results show the distribution of parameters on a logit scale: (a) Difficulty Level: Questions regarding the value of struggle tend to have high logit values (difficult), while questions related to national movement figures have low logit values (easy). (b) Participant Ability: The average participant ability is slightly above the average value of the level of difficulty of the questions (mean person logit > mean item logit). This indicates that this test is generally considered quite easy for samples in the transition area.

4. Wright Map Analysis (Variable Map)

Wright's map reveals a gap in information across the ability spectrum. There is a distribution of questions concentrated at high difficulty levels, but there are fewer items that can measure participants with moderate or very low history abilities. However, demographic representation indicates no significant bias between urban and rural respondents in responding to core items. In the Quest program, these results are visualized using the Variable Map below (Adam & Khoo, 1996).

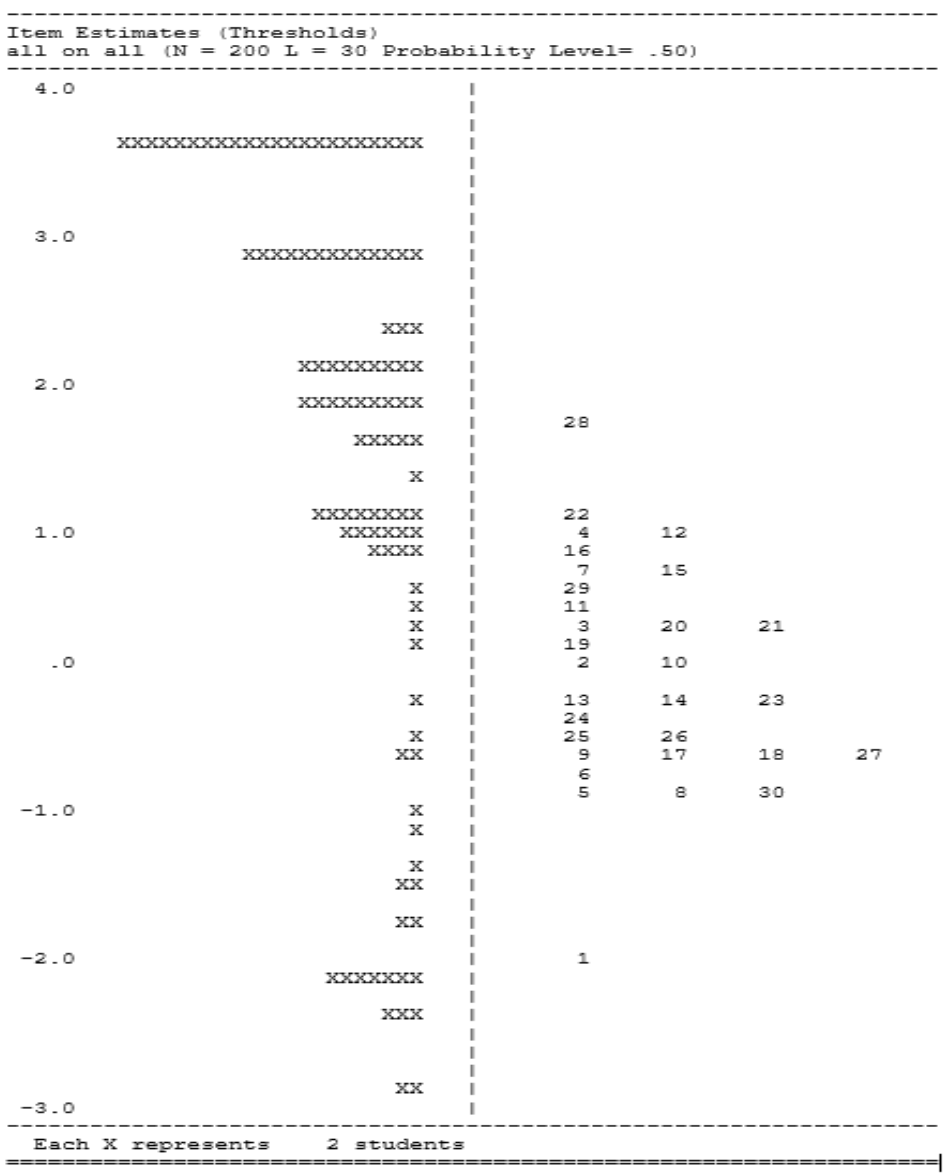


Figure 2. Variable map

The following is a description of the distribution: (a) Left Side (Person): Shows 200 participants spread from -3.0 logit to +4.0 logit. The largest concentration is in the transition region (near 3.0 logit). (b) Right Side (Item): Shows question 28 as the most difficult and 1 as the easiest.

DISCUSSION

1. Validity and Reliability of Instruments in the Rasch Model

The analysis results show that the Person Reliability (0.92) and Item Reliability (0.87) values are in the excellent category. This indicates that the History of National Struggle instrument has strong internal consistency when tested on samples in urban-rural transition areas. According to Bond, (2015), a reliability value above 0.70 indicates that the instrument is sensitive enough to differentiate respondents' ability levels and provides a stable order of item difficulty.

2. Item Fit and Sample Characteristics

Based on the Infit Mean Square (Infit MNSQ) criteria between 0.77 and 1.30, a small portion of the items were deemed fit. However, the presence of misfit items, such as items 1-3, 7-10, 13, 17, 18, 20-22, and 24-28, indicates the presence of unexpected response patterns (noise). From a Rasch Model perspective, misfit items indicate that the item fails to measure the same ability dimension as other items or contains ambiguity for certain respondent groups (Sumintono, 2014).

In the context of transitional areas, differences in information access between urban and rural subsamples can trigger Differential Item Functioning (DIF), where participants with the same ability level respond differently due to cultural backgrounds or heterogeneity in local curricula (Boone et al., 2014).

3. Distribution of Abilities on the Wright Map

Visualization using the Variable Map (Wright Map) shows that the average participant ability (1.47 logit) is above the average item difficulty level (0.00 logit). This gap indicates that national history material, particularly that related to struggle values, is still considered difficult by participants in the transition region. Equating ability and difficulty parameters on the same logit scale is a key advantage of the Rasch Model, allowing researchers to precisely identify at what level of difficulty participants begin to experience failure in answering (Adam & Khoo, 1996).

CONCLUSIONS AND RECOMMENDATIONS

Based on the results of the analysis using the Rasch Model, it can be concluded that: (a) Instrument Quality: The Indonesian National Struggle History Test has good reliability (Item Reliability 0.87 and Person Reliability 0.92), indicating that this instrument is stable and consistent for use in samples in urban-rural transition areas. (b) Model Suitability: 12 items (40%) meet the Infit MNSQ criteria (0.77–1.30), which means that these items have sufficient logical validity in measuring participants' historical abilities. (c) Ability Level: There is a gap between participants' abilities and the level of difficulty of the questions. The average participant ability (1.47 logit) is above the average difficulty of the questions (0.00 logit), which indicates that national history material is still relatively easy for respondents in the transition area. (d) Special Findings: Misfit items were found (such as material on the national movement, youth pledge, national figures, proclamation, struggle figures, Pancasila, national resistance, regional resistance, women figures, physical struggle, diplomacy, values of struggle), which showed unexpected answer patterns. This is thought to be due to the influence of educational backgrounds or different local information sources between urban and rural sub-samples.

Research result recommendations: (a) Instrument Improvement: It is necessary to revise or eliminate the test items that have MNSQ Infit values outside the standard range (1-3, 7-10, 13, 17, 18, 20-22, 24-28) so that the accuracy of future measurements increases. (b) Material Development: Educators in urban-rural transition areas are advised to further strengthen their understanding of materials that have a high level of difficulty (struggle value) through a more contextual approach. (c) Further Research: Further researchers

are advised to conduct a more in-depth Differential Item Functioning (DIF) analysis to identify which test items provide advantages or disadvantages for one of the groups (urban versus rural).

ADVANCED RESEARCH

Based on the findings regarding the characteristics of the sample in transitional areas (urban-rural) and the use of the Rasch Model, the following are several strategic opportunities for further research: (a) Differential Item Functioning (DIF) Analysis: Conducting an in-depth investigation to determine whether certain test items significantly benefit urban groups over rural groups (or vice versa) at equivalent ability levels. This is important to ensure the fairness of the national test instrument. (b) Qualitative Study of Misfit Phenomena: Conducting interviews or Focus Group Discussions (FGDs) with participants who provide unusual answer patterns (the cause of misfit). This research aims to explore whether the incorrect answers are caused by historical misconceptions or differences in historical narratives accepted in rural versus urban areas. (c) Adaptive Question Bank Development: Developing more test items within a specific logit range that is still empty (based on the Wright Map). The goal is to create an instrument with high precision for participants with very low or very high ability in transitional areas. (d) The Effect of Digital Literacy on History Scores: Examining the relationship between access to digital information (as a characteristic of urban areas) and historical reasoning ability (historical thinking) compared to conventional learning in rural areas. (e) Longitudinal Study of Historical Ability: Measuring whether the ability gap between urban and rural participants has narrowed as internet penetration in transition areas has increased over time.

REFERENCES

Adam, R. J., & Khoo, S.-T. (1996). Acer Quest: The Interactive Test Analysis System. In *Australian Council for Educational Research* (pp. 1–96).

Azizah, I. (2023). Analisis kualitas butir soal penilaian harian bersama fisika kelas X SMA negeri 1 Patikraja. *Jurnal Pendidikan Fisika, Oktober*, *10*(02), 90–104. https://journal.student.uny.ac.id/ojs/index.php/pfisika/index

Azwar, S. (2016). *Konstruksi Tes Kemampuan Kognitif. Pustaka Pelajar*. Pustaka Pelajar.

Bond, T. G., & F. C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed). Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781410614575/applying-rasch-model-trevor-bond-christine-fox

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (Vol. 10). Springer.

Mukhlis, M. (2021). Pendidikan Sejarah Dalam Pendidikan Karakter Bangsa. *Banjarmasin. Doi*, *10*.

Ricklefs, M. C. (1993). *A History of Modern Indonesia since c. 1200*. MacMillan London.

Setyawarno, D. (2017). *Upaya peningkatan kualitas butir soal dengan analisis aplikasi Quest*. FPMIPA UNY.

Sumintono, B. , & W. W. (2014). *Model Rasch untuk penelitian sosial kuantitatif*. 1–9. http://deceng3.wordpress.com

Suwarto, S., Suyahman, S., Suswandari, M., Zakiyah, Z., & Hidayah, A. (2023). The COVID-19 pandemic and the characteristic comparison of English achievement tests. *Perspektivy Nauki i Obrazovania*, *62*(2), 307–329. https://doi.org/10.32744/pse.2023.2.18