





ISSN: 2617-6548

URL: www.ijirss.com



Validation of the teacher education institution's entrance test using the Rasch model

 Manuel O. Maloniso^{1*},  Cherryl C. Maloniso²

¹Aklan State University, Banga Campus Banga, Aklan, Philippines.

²Aklan State University, New Washington Campus New Washington, Aklan, Philippines.

Corresponding author: Manuel O. Maloniso (Email: mmaloniso@asu.edu.ph)

Abstract

This study aimed to evaluate the reliability and validity of the College Entrance Test (CET) used by a teacher education institution to assess the general and specialization knowledge of prospective students in English, Mathematics, Science and Social Studies. The Rasch model was employed to analyze the data collected from the sample of 250 test takers for the general knowledge test, 122 for specialization in English, 74 for Mathematics, 122 for science and 77 for Social Studies. The measurement analysis components including person and item reliability, unidimensionality, person-item map, fit statistics, Point Measure Correlation (PTMEA) and item local dependence were used. The study findings indicated that some test components had poor person reliability and the degree of item difficulty was higher than the students' abilities. There were also concerns about the conformity to the unidimensionality criteria, suggesting an analysis of items that might form another dimension in the constructs of the tests, although each of the items is independent. Furthermore, some items were misfitting or overfitting the Rasch model. In conclusion, the CET needs improvement to ensure its quality as a reliable and valid selection tool for the college. The study's results provide significant insights into the CET's strengths and weaknesses that can guide the test developers in revising and enhancing the CET to effectively measure the general and specialization knowledge of test takers in English, Mathematics, Science and Social Studies.

Keywords: College entrance test, Item response theory, Rasch model, Reliability, Teacher education institution, Validity.

DOI: 10.53894/ijirss.v6i3.1726

Funding: This study received no specific financial support.

History: Received: 24 February 2023/**Revised:** 31 March 2023/**Accepted:** 22 May 2023/**Published:** 2 June 2023

Copyright: © 2023 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Authors' Contributions: Both authors contributed equally to the conception and design of the study.

Competing Interests: The authors declare that they have no competing interests.

Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Ethical Statement: This study followed all ethical practices during writing.

Publisher: Innovative Research Publishing

1. Introduction

In August 2017, the Philippine government started providing universal access to tertiary education through free tuition fees to college students in state universities and colleges in the country (Republic Act 10931).

With this law, every state college and university had to ensure that students entering college were the most qualified for the program offerings since universities are operating on a limited resource [1, 2] and every single centavo of the financial subsidy must be given to the most deserving students. Thus, selecting the most qualified students to benefit from this government policy is crucial.

In our province, Aklan State University is the sole state university that offers degrees in education and every year, a large number of senior high school graduates want to get into the programs. The college must ensure that its selection procedure at admission will select the best available candidates [3] to ensure that its accepted students have a greater probability of succeeding in college life and passing the licensure examination for teachers. Admission and selection to the universities are done through an entrance test or examination [4-6]. The College Entrance Test (CET) has an important role in the admission and selection process and it aims to determine who among the K-12 graduates is fit to enter college [7]. Since a high school diploma may not necessarily reflect adequate preparation "for the intellectual demands of adult life" Porter and Polikoff [8]. Thomas [9] emphasized that the pre-university scenario including inappropriate university program entry is a cause of unsuccessful college completion.

For several years, the College of Teacher Education entrance test has been in use. The tests have undergone validation and reliability testing processes following the Classical Test Theory (CTT) such as face and content validity, pilot testing and reliability testing using Kuder Richardson (KR)-20. However, there is no published evidence or record on the reliability and validity of the test. Evidence of validation is significant in supporting the decision on who is admitted to the college and given privileges from the government. Kane [10] pointed out that the main focus of test validation is making a convincing argument to justify score interpretations and the use of the test.

In traditional practice, face and content validity and a reliability coefficient using Cronbach's alpha are reported. However, there are limitations to these practices. Jafarkarimi [11] pointed out that to validate a test, one may take advantage of more advanced validation steps. Advanced methods of reliability and validity testing can be done using the Rasch model using Winsteps software [12].

2. Review of Related Literature

2.1. Rasch Model

Rasch's measurement model [13] is considered one of the most recent approaches to psychometric measurement [14]. The Rasch model can be used to overcome the limitations associated with classical test theory by assuming the unidimensionality of an instrument's latent traits and employing logistic regression to measure the probability of correct responses on discrete items [15]. Among the logistics models, Rasch is the simplest and generally referred to as the Item Response Theory model [16] which only uses one point and a constant parameter scale of 1 [17].

In this study, the Rasch framework is used to assess the reliability and validity of test scores. It offers procedures for constructing and revising social science measurement instruments, the measurement properties of instruments and enabling critical corrections to raw test scores or survey data [18].

The Rasch model (one of the most widely used item response theory models) serves as the foundation for the assumptions of test reliability and validity. The model is based on the probability of a correct response given by the equation:

$$P(x_{vi} = 1 | \beta_v, D_i) = P_{vi} = \frac{e^{(\beta_v - D_i)}}{1 + e^{(\beta_v - D_i)}}$$

$$P(x_{vi} = 0 | \beta_v, D_i) = 1 - P_{vi}$$

Here β represents the ability (latent trait) of subject v and D represents the difficulty parameter of item i . The probability of a correct response is determined by both the item's difficulty and the subject's ability.

2.2. Test Validation Using the Rasch Model

According to Glynn [19], the Rasch model has been widely applied in educational measurement including the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and other comparative tests. The Rasch model was also used in different countries and universities to validate an entrance test such as the validation of the university entrance English test of Vietnam National University [20], the validation of the university placement test in Nigeria [21] and the validation of the Chemistry national exam in Indonesia [22].

The Rasch model was also used to establish the validity of tests for specific subjects, examine the quality of mathematics test items [23, 24], a multiple-choice English vocabulary test [25], a science achievement test [26], a multiple-choice chemistry test [27] and a reading comprehension test [28].

This study is focused on validating the college entrance test for the teacher education institution using the Rasch model to address the limitations of its previous validation. It aims to determine the characteristics of the test particularly its reliability, item difficulty and validity.

This study was guided by the following objectives:

1. Determine the reliability of the test using item and person reliability.
2. Compare the samples' ability against the test item difficulty using the Person-Item map.
3. Determine the construct validity of the test using the Rasch measurements of unidimensionality, item fit, item polarity and local independence.

3. Methodology

This study used a quantitative research design specifically a descriptive and correlational design to establish significant measurements in test reliability and validity. The descriptive design covers the process of describing the scores of the samples particularly in locating the samples' ability against the item difficulty while correlation is used in establishing the reliability and validity of the test.

3.1. Data Collection Tool

The College Entrance Test (CET) of the College of Teacher Education is a multiple-choice test approved by the college during the admissions process. The test has two main parts: Part 1 has a 60-item General Knowledge Test (GKT) and Part 2 has a 40-item Specialization Knowledge Test (SKT). The GKT is taken by all student applicants to the college while the SKT is given depending on the choice of specialization of the student applicant. Thus, there are SKT-English, SKT-Mathematics, SKT-Science and SKT-Social Studies in line with the areas of specialization offered in the college. The result of the CET is one of the criteria for the admission of the student-applicant in their selected area of specialization.

3.2. Data Collection Process

The responses of the student-applicants in the CET were used in the analysis of reliability and validity using the Rasch model. These responses were collected during the scheduled college entrance test conducted on May 24 and 26, 2022 for batch 1, May 31, 2022 for batch 2 and July 7, 2022 for batch 3. Permission was obtained from the Dean of the College of Teacher Education to use the data for the purpose of this study and confidentiality of the data was observed.

3.3. The Respondents

Table 1 presents the number of respondents in the study. A total of 250 samples were used for GKT, 122 for SKT-English, 74 for SKT-Mathematics, 122 for SKT-Science and 77 for SKT-Social Studies. According to Linacre [29], the number of participants for the Rasch model could range from 30 to 250 with 250 participants for a definitive or high-stakes test.

Table 1.
The number of the test items and number of respondents.

Test	Number of items	f
General knowledge test	60	250
Specialization knowledge test		
English	40	122
Mathematics	40	74
Science	40	122
Social studies	40	77

3.4. Data Analysis

This study used Rasch analysis to establish the reliability and validity of the college entrance test. WINSTEPS and Rasch software were used to analyze the data. The data in this study are the responses of the student-applicants to every item in the test. A correct response was represented as 1 while an incorrect response was represented as 0. Since the Rasch model assumed the unidimensionality of the construct in the test, Rasch measurements such as person and item reliability, unidimensionality (Principal Component Analysis or PCA in raw explained variance, Eigenvalue, and Observed Unexplained Variance), person-item map, infit and outfit statistics (MNSQ), Point Measure Correlation (PTMEA) and item local independence were performed separately in GKT, SKT-English, SKT-Mathematics, SKT-Science and SKT-Social Studies. The values obtained in every Rasch measure were compared to suggested acceptable values to make decisions about the performance of the test in line with the Rasch model.

4. Results

4.1. Unidimensionality

It is necessary to determine a test's unidimensionality to ensure that it assess the expected outcomes. According to Tennant and Pallant [30], a value above 0.9 indicates unidimensionality. Table 2 shows that the Rasch Unidimensionality Coefficients (eigenvalues) for the general knowledge test (2.4) and the specialization knowledge tests in English (2.4), Mathematics (3.7), Science (2.3) and Social Studies (3.2) are all above 0.9 suggesting that the test measures a single construct. However, Linacre [31] has pointed out that an eigenvalue of not >2 in the first contrast of unexplained variance demonstrates that the test is unidimensional. In this study, all test eigenvalues are >2 . Furthermore, the Rasch model's item unidimensionality criterion is also examined based on the raw variance explained by the measure score and the observed unexplained variance. In this study, all scores' observed unexplained variance is $<15\%$ and the raw variance explained by the measures of the general knowledge test, specialization knowledge test in English, and specialization knowledge test in Mathematics is acceptable ($>20\%$). For the specialization knowledge tests in Science and Social Studies, it is $<20\%$.

Table 2.

The raw variance explained by measure and the unexplained variance in the first contrast.

Test	Raw variance explained by measure	Unexplained variance 1st contrast	
		Eigenvalue	Observed
General knowledge test	24.9	2.4	2.9
Specialization knowledge test			
English	22.3	2.4	4.6
Mathematics	22.2	3.7	7.3
Science	19.5	2.3	4.6
Social studies	16.2	3.2	6.7

4.2. Item and Person Reliability

Table 3 presents the summary of the reliability test. The item reliability measure for the test is excellent ranging from 0.88 to 0.98. Specifically, the item reliability of GKT is 0.98 while that of SKT-English is 0.96, SKT-Mathematics is 0.92, SKT-Science is 0.94 and SKT-Social Studies is 0.88. In terms of person reliability, GKT and SKT-Mathematics have acceptable coefficients of 0.61 and 0.67 respectively while SKT-English, SKT-Science and SKT-Social Studies have poor reliability as reflected in the coefficients of 0.44, 0.51 and 0.56, respectively.

Table 3.

Reliability test summary.

Test	Cronbach alpha	Item reliability	Person reliability
General knowledge test	0.63	0.98	0.61
Specialization knowledge test			
English	0.44	0.96	0.44
Mathematics	0.69	0.92	0.67
Science	0.52	0.94	0.51
Social studies	0.57	0.88	0.56

4.3. Person-Item Map

Item 39 is the most difficult item for the general knowledge test while item 60 is the easiest one. Moreover, other items that are considered very difficult (> 1 logit) [32] are items 19, 48, 45, 13, 15, 29, 36, 42 and 34. Further, the mean of the person's ability is -0.5 logit and the mean of item difficulty is 0 logit.

In the specialization knowledge test in English, the most difficult item is item 3 while the easiest item is item 11. There are other very difficult items (> 1 logit), such as items 32, 35, 33 and 6. Moreover, the person's ability mean is -0.34 logit which is lower than the mean of the item difficulty (0 logit).

The most difficult item in the specialization knowledge test in Mathematics is item 2 while the easiest is item 32. Other very difficult items (> 1 logit) recorded include items 28, 1 and 36. The mean of person ability is 0.53 logit while the mean of item difficulty is 0 logit.

In the specialization knowledge test in science, the most difficult item is 2 while the easiest items are 22 and 25. Other very difficult items (> 1 logit) include 32, 15, 27, 28 and 7. The mean of a person's ability in SKT-Science is 0.83 logits lower than the mean of the item's difficulty which is 0 logits.

Moreover, the most difficult item in the specialization knowledge test in social studies is item 37 and the easiest item is item 28. Other very difficult items (> 1 logit) include items 21 and 13. Overall, the person's ability mean is -1.01 logit and the mean of the item difficulty is 0 logit.

Table 4 shows the distribution of the item difficulty level. According to the table, there are 12 GKT items, 7 SKT-English items, 5 SKT-Math items, 6 SKT-Science items and 5 SKT-Social Studies items whose value is > 1 logit and are considered "very difficult". There are 19 GKT items, 12 SKT-English items, 13 SKT-Math items, 12 SKT-Science items and 14 SKT-Social Studies items whose value is between 0 and 1 logit and are considered "difficult." On the other hand, items whose values are between -1 and -0 logit include 16 GKT items, 15 SKT-English items, 15 SKT-Math items, 15 SKT-Science items and 16 SKT-Social Studies items. These items are considered "easy." Moreover, "very easy" items have a value of -1 logit and this includes 13 GKT items, 6 SKT-English items, 7 SKT-Math items, 7 SKT-Science items and 5 SKT-Social Studies items.

4.4. Item Fit and Item Polarity

Mean Square (MNSQ) infit and outfit values are used to determine whether an item is a good fit, overfit or misfit while point measure correlation is used to determine the item's polarity.

According to Table 5, the minimum and maximum infit MNSQ values of the general knowledge test, the specialization knowledge test in English, the specialization knowledge test in Math, the specialization knowledge test in Science and the specialization knowledge test in Social Studies are within the range of 0.70 – 1.30 . However, there are minimum or maximum values in the outfit MNSQ in the general knowledge test, the specialization knowledge test in English, the specialization knowledge test in Math, the specialization knowledge test in Science and Social Studies that are outside the 0.70 – 1.30 range. Thus, there is a need to identify these items that are misfits or overfits in the Rasch measure.

Table 4.
Item difficulty level.

Test	Very difficult	Difficult	Easy	Very easy
General knowledge test	12	19	16	13
Specialization knowledge test				
English	7	12	15	6
Mathematics	5	13	15	7
Science	6	12	15	7
Social studies	5	14	16	5

Table 5.
The overall fit statistics.

Item fit measurements	General knowledge test	Specialization knowledge test			
		English	Mathematics	Science	Social studies
Infit MNSQ					
Minimum	0.88	0.90	0.78	0.86	0.88
Maximum	1.14	1.08	1.27	1.12	1.13
Mean	1.00	1.00	1.00	1.00	0.99
SD	0.05	0.05	0.12	0.06	0.06
Outfit MNSQ					
Minimum	0.63	0.87	0.74	0.72	0.73
Maximum	1.25	1.60	2.40	1.39	1.55
Mean	1.00	1.01	1.06	1.02	1.00
SD	0.09	0.12	0.30	0.13	0.15

The infit and outfit MNSQ and Point Measure Correlation (PTMEA) of the General Knowledge Test ([Appendix B.1](#)) revealed that item 39 is a misfit since it has an outfit MNSQ of 0.63 which is <0.70 . On the other hand, the PTMEA coefficient of item 33 is -0.08. Similarly, the infit and outfit MNSQ and Point Measure Correlation (PTMEA) of the Special Knowledge Test in English ([Appendix B.2](#)) show that the outfit MNSQ value of item 3 is >1.30 and at the same time, it has a -0.19 PTMEA coefficient.

Moreover, items 2, 1, 36, 20 and 40 are overfit items in SKT-Mathematics ([Appendix B.3](#)) because their outfit MNSQ is >1.30 and at the same time, their PTMEA coefficient is negative. In addition, item 3 has a -0.01 PTMEA. In SKT-Science ([Appendix B.4](#)), item 2's outfit MNSQ is 1.39 which is greater than 1.30 and items 27 and 40 have a PTMEA coefficient of -0.02. Further, the MNSQ and PTMEA of the SKT-Social Studies ([Appendix B.5](#)) revealed that items 40 and 33 are considered overfit since their outfit MNSQ values are 1.40 and 1.55 respectively and these are >1.30 .

4.5. Local Independence

The analysis of local independence is another type of item measurement analysis. The standardized residual correlation measures of the tests range from 0.20 to 0.66 which is <0.70 . The highest computed standardized residual correlation measure in GKT is 0.20 for items 3 and 33 and items 9 and 10 while in SKT-English, it is 0.52 for items 1 and 2. Similarly, the highest standardized residual correlation in SKT-Mathematics is 0.66 on items 26 and 35, in SKT-Science, it is 0.31 on items 7 and 32 and in SKT-Social Studies it is 0.50 on items 7 and 18.

5. Discussion

Producing a high-quality test to perform its intended purpose is significant especially if it concerns crucial decision-making in an institution and thus needs substantial proof of its validity and reliability. This study revealed the significant features of the college entrance test particularly its reliability and validity.

The person reliability of college entrance tests particularly for SKT-English, Science and Social Studies suggests that the number of items in these tests is insufficient to discriminate against test taker ability [15]. This is visually presented in the Item-Person map where the majority of the test takers' performance in the entrance test falls below the test's item difficulty indicating that the test takers find it difficult to respond correctly to the test items. On the other hand, the item reliability has a good to very good level of consistency [33] which confirms that there are enough test takers to confirm the difficulty of the items in the test.

Unidimensionality is a key assumption in applying the Rasch model [34]. The test construct as described through the principal component analysis suggests that there is a sign of concern for its unidimensionality since the test eigenvalues are >2 and the raw variance explained by measure for SKT-Science and Social Studies is below the acceptable measure. There is a need to further analyze the tests and if there is no meaningful difference in the items, the other dimension may just happen by chance [30].

With the test displaying signs of having more than one dimension, item fit and item polarity must be performed to identify items forming other dimensions. The analysis of item fit in a questionnaire refers to the Rasch measurement model's fitness for each item [35]. The aim is to ensure that the items measure the same construct [33]. A value between 0.7 and 1.30 is typically considered reasonable for multiple-choice questions indicating a good fit [36]. A value of 1 indicates a perfect fit while values less than 0.70 or greater than 1.30 indicate a misfit or an overfit respectively, and such items may

need to be discarded or repaired [37]. The Point Measure Correlation (PTMEA) can be used to determine the item's polarity. A positive PTMEA correlation indicates that the item measures the desired construct while a negative PTMEA correlation suggests that the item does not measure the intended construct [33]. Therefore, two items in GKT (one item in SKT-English, six items in SKT-Mathematics, three items in SKT-Science and two items in SKT-Social Studies) need to be repaired or discarded since they do not contribute to the measurement of the construct in the test and may represent other dimensions.

To ensure that the items in the test do not overlap, a standardized residual value measurement correlation was determined. The high value in this measurement indicates that items have the same characteristics or could be duplicates and are not independent from one another. If the correlation value of the two items is above 0.7, it shows a high correlation value and only one item has to be maintained while the other items should be dropped [37]. The item to be retained is determined using the MNSQ value which should be close to or equal to 1.0 [12, 37]. Overall, no items in the tests are duplicates.

6. Conclusion

The Rasch model provides robust and empirical evidence on the reliability and validity of the college entrance test of the Teacher Education Institution through its analysis components such as person and item reliability, unidimensionality, person-item map, fit measurements such as infit and outfit MNSQ, PTMEA (item polarity) and item local independence. The analysis showed that there were sufficient samples to determine the construction of the test. However, the specialization knowledge test in English, Science, and Social Studies do not have a sufficient number of items to discriminate between high- and low-performing test takers. The degree of difficulty of the tests is higher than the test takers' abilities which give them a lower probability of responding correctly to some items. The tests show inconsistency in terms of meeting some criteria for unidimensionality which leads to the possibility of the existence of another dimension in the construct. Further, there are items that are misfit or overfit to the Rasch model, these items are either confusing or can be easily predicted by the test takers and they need to be repaired or discarded to improve the overall fit of the tests. Moreover, items in the test are independent from one another.

The college entrance test must be reviewed and revised to address the inconsistencies found using the Rasch model by discarding or improving misfitting or overfitting items to address the concerns about unidimensionality and the overall fit of the test to the Rasch model, adding items to the English, Science and Social Studies Specialization knowledge tests to improve discrimination between high- and low-performing test takers, considering adding items to the general knowledge test and specialization knowledge test in Mathematics to increase their reliability from "acceptable" to "good" or "excellent" and arranging the items according to their degree of difficulty. Moreover, reliability and validity testing confirm whether the tests have improved and conform to the Rasch model measures.

References

- [1] M. I. Conchada and I. G. Zamudio, "The cost efficiency of state universities and colleges in the Philippines," *Philippine Review of Economics*, vol. 50, no. 2, pp. 83-104, 2013.
- [2] R. Alda, H. Boholano, and F. Dayagbil, "Teacher education institutions in the philippines towards education 4.0," *International Journal of Learning, Teaching and Educational Research*, vol. 19, no. 8, pp. 137-154, 2020. <https://doi.org/10.26803/ijlter.19.8.8>
- [3] A. F. Montalbo, Y. P. Evangelista, and M. M. Bernal, "Admission test as predictor of student performance in political science and psychology students of Rizal Technological University," *Asia Pacific Journal of Multidisciplinary Research*, vol. 6, no. 3, pp. 68-73, 2018.
- [4] I. e. a. Testa, *Validation of university entrance tests through rasch analysis*. In: Khine, M. (eds) *rasch measurement*. Singapore: Springer, 2020.
- [5] N. Kuramoto and R. Koizumi, "Current issues in large-scale educational assessment in Japan: Focus on national assessment of academic ability and university entrance examinations," *Assessment in Education: Principles, Policy & Practice*, vol. 25, no. 4, pp. 415-433, 2018. <https://doi.org/10.1080/0969594x.2016.1225667>
- [6] G. Davey, C. De Lian, and L. Higgins, "The university entrance examination system in China," *Journal of Further and Higher Education*, vol. 31, no. 4, pp. 385-396, 2007. <https://doi.org/10.1080/03098770701625761>
- [7] L. Jawad, "Examining college readiness in an early college program that focuses on health careers: Perceptions of graduating students," Dissertation Presented at the University of Michigan-Dearborn, 2017.
- [8] A. C. Porter and M. S. Polikoff, "Measuring academic readiness for college," *Educational Policy*, vol. 26, no. 3, pp. 394-417, 2012. <https://doi.org/10.1177/0895904811400410>
- [9] L. Thomas, "Do pre-entry interventions such as aimhigher 'impact on student retention and success? A review of the literature," *Higher Education Quarterly*, vol. 65, no. 3, pp. 230-250, 2011. <https://doi.org/10.1111/j.1468-2273.2010.00481.x>
- [10] M. T. Kane, "Validating the interpretations and uses of test scores," *Journal of Educational Measurement*, vol. 50, no. 1, pp. 1-73, 2013. <https://doi.org/10.1111/jedm.12000>
- [11] H. Jafarkarimi, "Is there a questionnaire validation protocol I can use to validate a survey?," Retrieved: https://www.researchgate.net/post/Is_there_a_questionnaire_validation_protocol_I_can_use_to_validate_a_survey. 2015.
- [12] T. G. Bond and C. M. Fox, *Applying the rasch model: Fundamental measurement in the human sciences*, 3rd ed. London: Mahwah, New Jersey: Lawrence Erlbaum Associates, 2015.
- [13] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.
- [14] J. F. Pallant and A. Tennant, "An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale (HADS)," *British Journal of Clinical Psychology*, vol. 46, no. 1, pp. 1-18, 2007. <https://doi.org/10.1348/01446506x96931>

- [15] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch analysis in the human sciences*. Dordrecht: Springer, 2014.
- [16] T. Bond, "Validity and assessment: A Rasch measurement perspective," *Behavioral Science Methodology*, vol. 5, no. 2, pp. 179-194, 2004.
- [17] B. D. Wright, "Solving measurement problems with the Rasch model," *Journal of Educational Measurement*, vol. 14, no. 2, pp. 97-116, 1977.
- [18] W. J. Boone, "Rasch analysis for instrument development: Why, when, and how?," Retrieved: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5132390/>. 2016.
- [19] S. M. Glynn, "International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items," *Journal of Research in Science Teaching*, vol. 49, no. 10, pp. 1321-1344, 2012. <https://doi.org/10.1002/tea.21059>
- [20] H. P. Tran, P. Griffin, and C. Nguyen, "Validating the university entrance English test to the Vietnam National University: A conceptual framework and methodology," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 1295-1304, 2010. <https://doi.org/10.1016/j.sbspro.2010.03.190>
- [21] A. A. Bichi, R. Talib, R. Embong, H. B. Mohamed, M. S. Ismail, and A. Ibrahim, "Rasch-based objective standard setting for university placement test," *Eurasian Journal of Educational Research*, vol. 19, no. 84, pp. 57-70, 2019. <https://doi.org/10.14689/ejer.2019.84.3>
- [22] A. Darmana, A. Sutiani, H. A. Nasution, I. Ismanisa, and N. Nurhaswinda, "Analysis of RASCH model for the validation of chemistry national exam instruments," *Journal of Indonesian Science Education*, vol. 9, no. 3, pp. 329-345, 2021. <https://doi.org/10.24815/jpsi.v9i3.19618>
- [23] N. b. A. Razak, A. Z. bin Khairani, and L. M. Thien, "examining quality of mathematics test items using rasch model: Preliminary analysis," *Procedia-Social and Behavioral Sciences*, vol. 69, pp. 2205-2214, 2012. <https://doi.org/10.1016/j.sbspro.2012.12.187>
- [24] A. Z. Bin Khairani and N. bin Abd Razak, "Modeling a multiple choice mathematics test with the Rasch model," *Indian Journal of Science and Technology*, vol. 8, no. 12, pp. 1-6, 2015. <https://doi.org/10.17485/ijst/2015/v8i12/70650>
- [25] P. Baghaei and N. Amrahi, "Validation of a multiple choice English vocabulary test with the Rasch model," *Journal of Language Teaching and Research*, vol. 2, no. 5, pp. 1052-1060, 2011. <https://doi.org/10.4304/jltr.2.5.1052-1060>
- [26] P. Susongko, "Validation of science achievement test with the Rasch model," *Journal of Indonesian Science Education*, vol. 5, no. 2, pp. 268-277, 2016. <https://doi.org/10.15294/jpii.v5i2.7690>
- [27] A. Winarti and A. Mubarak, "Rasch modeling: A multiple choice chemistry test," *Indonesian Journal on Learning and Advanced Education*, vol. 2, no. 1, pp. 1-9, 2020. <https://doi.org/10.23917/ijolae.v2i1.8985>
- [28] P. Baghaei and C. H. Carstensen, "Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types," *Practical Assessment, Research & Evaluation*, vol. 18, no. 5, pp. 1-14, 2013. <https://doi.org/10.7275/n191-pt86>
- [29] J. Linacre, "Sample size and item calibration stability," *Rasch Measurement Transactions*, vol. 7, no. 4, p. 328, 1994.
- [30] A. Tennant and J. F. Pallant, "Unidimensionality matters!(A tale of two Smiths?)," *Rasch Measurement Transactions*, vol. 20, no. 1, pp. 1048-1051, 2006. <https://www.rasch.org/rmt/rmt201c.htm>
- [31] J. M. Linacre, "A user's guide to winsteps rasch model computer programs: Program manual 3.92.0." Beaverton, OR: Winsteps.com, 2011, pp. 601-602.
- [32] B. Sumintono and W. Widhiarso, *Application of rasch modeling in educational assessment*. Jakarta: Communication Trim, Cimahi, 2015.
- [33] T. G. Bond and C. M. Fox, *Applying the rasch model: Fundamental measurement in the human sciences*, 2nd ed. London: Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 2007.
- [34] M. D. Reckase, "Multidimensional item response theory models." New York: Springer, 2009, pp. 79-112.
- [35] S. R. Ariffin, *Innovations in educational measurement and evaluation*. Bangi: National University of Malaysia, 2008.
- [36] J. R. Marchant, "Assessing the validity of multiple-choice questions, using them to undertake comparative analysis on student cohort performance, and evaluating the methodologies used," Retrieved: https://digital.library.adelaide.edu.au/dspace/bitstream/2440/132705/2/Marchant2020_PhD.pdf. 2020.
- [37] J. M. Linacre, "A user's guide and program manual to winstep: Rasch model computer program." Chicago: MESA Press, 2005.

Appendix

Appendix A.1 presents the person and item reliability of GKT.

Appendix A.1.

Summary statistics of GKT.

Summary of 250 measured persons								
	RAW Score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	24.3	60	-0.5	0.3	1	0	1	0
S.D.	5.5	0	0.5	0.01	0.14	1.1	0.25	1
Max.	40	60	0.9	0.38	1.4	2.7	2.03	3.1
Min.	10	60	-2.02	0.29	0.64	-3.4	0.55	-2.4
Real RMSE	0.31	ADJ.SD	0.39	Separation	1.25	Person reliability	0.61	
Model RMSE	0.30	ADJ.SD	0.40	Separation	1.31	Person reliability	0.63	
S.E. of person mean = 0.03								
Person raw score-to-measure correlation = 1.00								
Cronbach alpha (KR-20) Person raw score reliability = 0.63								
Summary of 60 measured items								
	RAW Score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	101.4	250	0	0.16	1	-0.1	1	0
S.D.	56	0	1.25	0.05	0.05	1.2	0.09	1.3
Max.	243	250	2.9	0.38	1.14	2.7	1.25	3
Min.	9	250	-4.16	0.13	0.88	-4.1	0.63	-3.9
Real RMSE	0.17	ADJ.SD	1.24	Separation	7.41	Person reliability	0.98	
Model RMSE	0.17	ADJ.SD	1.24	Separation	7.46	Person reliability	0.98	
Umean=0.000 Uscale=1.000								
Item RAW score-to-measure correlation = -0.98								
15000 Data points. Log-likelihood chi-square: 16165.80 with 14691 d.f. p=.0000								

Appendix A.2 presents the person and item reliability of SKT – English.

Appendix A.2.

Summary statistics of SKT – English.

Summary of 122 measured persons								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	17.3	40	-0.34	0.36	1	0	1.01	0
S.D.	3.8	0	0.5	0.01	0.15	1	0.26	1
Max.	29	40	1.2	0.45	1.54	3.1	2.29	3.9
Min.	7	40	-1.88	0.35	0.73	-2.1	0.65	-1.9
Real RMSE	0.37	ADJ.SD	0.33	Separation	0.88	Person reliability	0.44	
Model RMSE	0.36	ADJ.SD	0.34	Separation	0.93	Person reliability	0.46	
S.E. of person mean = 0.05								
Person RAW score-to-measure correlation = 1.00								
Cronbach alpha (KR-20) Person RAW score reliability = 0.44								
Summary of 40 measured items								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	52.8	122	0	0.21	1	0	1.01	0
S.D.	25.8	0	1.09	0.04	0.05	0.8	0.12	0.9
Max.	109	122	2.42	0.37	1.08	1.8	1.6	1.8
Min.	8	122	-2.56	0.19	0.9	-2.4	0.87	-2.4
Real RMSE	0.22	ADJ.SD	1.06	Separation	4.84	Person reliability	0.96	
Model RMSE	0.22	ADJ.SD	1.06	Separation	4.88	Person reliability	0.96	
Umean=0.000 Uscale=1.000								
Item RAW score-to-measure correlation = -0.99								
4880 Data points. Log-likelihood chi-square: 5493.95 with 4719 d.f. p=.0000								

Appendix A.3 presents the person and item reliability of SKT – Mathematics.

Appendix A.3.

Summary statistics of SKT – mathematics.

Summary of 75 measured persons								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	16.1	40	-0.53	0.37	1	-0.1	1.06	0
S.D.	5	0	0.66	0.02	0.16	1.1	0.4	1.2
Max.	28	40	1.01	0.5	1.45	2.2	2.66	3.5
Min.	5	40	-2.27	0.35	0.65	-3	0.59	-2.4
Real RMSE 0.38	ADJ.SD 0.54		Separation 1.42		Person reliability 0.67			
Model RMSE 0.37	ADJ.SD 0.55		Separation 1.49		Person reliability 0.69			
S.E. of person mean = 0.08								
Person RAW score-to-measure correlation = 1.00								
Cronbach alpha (KR-20) Person raw score reliability = 0.69								
Summary of 40 measured items								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	30.1	75	0	0.27	1	-0.2	1.06	0
S.D.	14	0	0.99	0.05	0.12	1.3	0.3	1.4
Max.	58	75	2.54	0.52	1.27	3.1	2.4	3.4
Min.	4	75	-1.86	0.24	0.78	-3.1	0.74	-3
Real RMSE 0.29	ADJ.SD 0.95		Separation 3.29		Person reliability 0.92			
Model RMSE 0.28	ADJ.SD 0.95		Separation 3.39		Person reliability 0.92			
Umean=0.000 Uscale=1.000								
Item RAW score-to-measure correlation = -0.99								
3000 Data points. Log-likelihood chi-square: 3341.04 with 2886 d.f. p=0.0000								

Appendix A.4 presents the person and item reliability of SKT – Science.

Appendix A.4.

Summary statistics of SKT – science.

Summary of 122 measured persons								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	13.9	40	-0.81	0.37	1	0	1.02	0
S.D.	4	0	0.55	0.03	0.15	1	0.36	1
Max.	31	40	1.47	0.5	1.4	2.4	3.72	3.6
Min.	5	40	-2.27	0.35	0.68	-2.7	0.61	-2.3
Real RMSE 0.38	ADJ.SD 0.39		Separation 1.01		Person reliability 0.51			
Model RMSE 0.37	ADJ.SD 0.40		Separation 1.07		Person reliability 0.53			
S.E. of person mean = 0.05								
Person RAW score-to-measure correlation = 1.00								
Cronbach alpha (KR-20) Person raw score reliability = 0.52								
Summary of 40 measured items								
	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	42.3	122	0	0.22	1	-0.1	1.02	0
S.D.	22.1	0	0.99	0.06	0.06	0.8	0.13	1
Max.	86	122	2.72	0.51	1.12	1.9	1.39	2.1
Min.	4	122	-1.74	0.19	0.86	-1.9	0.72	-2
Real RMSE 0.23	ADJ.SD 0.96		Separation 4.13		Person reliability 0.94			
Model RMSE 0.23	ADJ.SD 0.97		Separation 4.19		Person reliability 0.95			
Umean=0.000 Uscale=1.000								
Item RAW score-to-measure correlation = -0.98								
4880 Data points. Log-likelihood chi-square: 5298.85 with 4719 d.f. p=0.0000								

Appendix A.5 presents the person and item reliability of SKT – social studies.

Appendix A.5.

Summary statistics of SKT – social studies.

Summary of 77 measured persons

	RAW score	COUNT	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	12	40	-1.01	0.38	1	0	1	0
S.D.	4.1	0	0.59	0.04	0.12	0.7	0.28	0.8
Max.	35	40	2.2	0.61	1.4	2	2.55	2.6
Min.	3	40	-2.79	0.34	0.78	-1.4	0.6	-1.4
Real RMSE	0.39	ADJ.SD	0.45	Separation	1.13	Person reliability	0.56	
Model RMSE	0.38	ADJ.SD	0.45	Separation	1.18	Person reliability	0.58	

S.E. of Person mean = 0.07

Person Raw score-to-measure correlation = 1.00

Cronbach alpha (KR-20) Person Raw score reliability = .57

Summary of 40 measured items

	RAW score	Count	Measure	Model error	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	23.1	77	0	0.28	0.99	0.1	1	0.2
S.D.	11.6	0	0.85	0.06	0.06	0.8	0.15	1.2
Max.	51	77	1.82	0.48	1.13	2.2	1.55	4.7
Min.	5	77	-1.74	0.23	0.88	-1.7	0.73	-1.5
Real RMSE	0.29	ADJ.SD	0.79	Separation	2.71	Person reliability	0.88	
Model RMSE	0.29	ADJ.SD	0.79	Separation	2.71	Person reliability	0.88	

Umean=.000 Uscale=1.000

Item raw score-to-measure correlation = -0.98

3080 Data points. Log-likelihood chi-square: 3251.13 with 2964 d.f. p=0.0001

Appendix B.1.

Item difficulty, infit MNSQ, outfit MNSQ, and PTMEA correlation for GKT.

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure Correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
39	0.95	0.63	0.27	12	1.08	1.11	0.06
19	0.99	1.06	0.11	1	0.97	0.97	0.29
48	0.99	1.02	0.15	5	0.95	0.94	0.34
45	0.99	0.98	0.16	41	1.01	1.01	0.22
13	1.02	1.07	0.11	17	1.04	1.05	0.16
29	1.05	1.25	0	55	1.03	1.04	0.17
36	0.99	1.05	0.15	54	0.93	0.92	0.38
42	1.04	1.13	0.04	10	0.96	0.95	0.32
15	1.04	1.18	0.04	58	1.02	1.03	0.19
34	0.99	0.93	0.22	4	0.88	0.88	0.47
18	1.01	1.03	0.16	44	1.09	1.11	0.06
38	0.98	1.01	0.2	25	1.00	0.99	0.24
35	1.01	1.03	0.17	51	1.02	1.02	0.2
3	1.01	1.04	0.15	7	0.93	0.93	0.37
32	0.99	1.00	0.2	6	1.04	1.07	0.14
28	1.06	1.14	0.05	2	0.94	0.94	0.35
37	1.04	1.07	0.11	23	1.06	1.07	0.12
59	0.97	0.96	0.26	26	0.91	0.89	0.41
16	0.99	1.00	0.22	47	1.03	1.05	0.16
14	1.01	1.01	0.19	30	0.95	0.94	0.33
40	1.01	1.05	0.18	43	0.99	1.00	0.25
27	1.01	1.02	0.18	24	0.92	0.9	0.38
33	1.14	1.20	-0.08	46	1.08	1.10	0.07
57	0.96	0.95	0.29	50	0.97	0.96	0.29
20	1.00	1.03	0.2	56	1.03	1.04	0.16
22	0.97	0.98	0.27	9	0.9	0.88	0.41
52	0.9	0.88	0.42	11	0.93	0.89	0.35
21	0.98	0.98	0.26	8	0.97	0.92	0.25
53	1.05	1.04	0.14	49	1.03	1.01	0.12
31	1.06	1.07	0.1	60	1.00	0.93	0.09

Appendix B.2.

Item difficulty, infit MNSQ, outfit MNSQ, and PTMEA correlation for SKT-English.

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
3	1.08	1.60	-0.19	9	1.00	1.00	0.23
32	0.95	0.89	0.26	23	1.08	1.08	0.08
35	1.02	1.20	0.08	25	0.96	0.95	0.31
6	0.99	0.95	0.21	19	0.97	0.96	0.29
33	0.97	0.9	0.26	39	0.91	0.9	0.41
5	1.00	1.05	0.16	4	0.99	0.98	0.25
34	0.94	0.87	0.34	18	0.96	0.96	0.3
21	1.04	1.04	0.12	14	1.00	1.00	0.23
27	1.08	1.16	0.02	29	1.01	0.99	0.23
38	0.98	0.96	0.25	28	0.9	0.88	0.44
17	1.05	1.08	0.1	30	1.08	1.09	0.07
10	0.99	1.01	0.22	22	0.97	0.96	0.28
40	1.02	1.00	0.18	37	1.02	1.05	0.17
15	0.92	0.92	0.37	13	0.99	0.99	0.25
12	1.01	1.07	0.16	7	1.03	1.10	0.12
2	0.99	0.97	0.24	8	1.00	1.03	0.2
24	0.95	0.93	0.33	26	0.97	0.98	0.25
31	0.98	1.00	0.26	16	0.95	0.89	0.29
36	1.08	1.11	0.06	20	1.05	1.13	0.05
1	1.04	1.05	0.14	11	0.99	0.89	0.2

Appendix B.3.

Item difficulty, infit MNSQ, outfit MNSQ, and PTMEA correlation for SKT-mathematics.

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
2	1.11	2.40	-0.24	7	1.01	1.06	0.27
28	0.95	1.08	0.24	10	0.96	0.94	0.36
1	1.19	1.71	-0.19	26	0.81	0.78	0.57
36	1.11	1.48	-0.02	18	0.91	0.9	0.43
23	1.03	1.14	0.17	33	0.86	0.84	0.5
34	0.99	0.9	0.29	31	1.11	1.11	0.14
9	1.05	1.06	0.19	13	0.83	0.82	0.54
20	1.24	1.48	-0.16	27	0.89	0.87	0.47
21	0.92	0.83	0.39	8	1.03	1.05	0.25
22	0.97	1.04	0.28	3	1.18	1.28	-0.01
39	1.08	1.22	0.1	25	0.92	0.89	0.42
6	1.10	1.20	0.1	16	0.95	0.91	0.38
19	0.92	0.95	0.38	24	0.92	0.89	0.42
38	1.12	1.18	0.09	4	1.01	1.02	0.27
15	1.12	1.14	0.1	14	0.9	0.87	0.44
5	1.09	1.07	0.17	29	1.08	1.08	0.17
17	0.97	0.95	0.34	12	0.95	0.92	0.37
30	0.98	0.96	0.33	37	0.84	0.81	0.51
40	1.27	1.36	-0.12	11	0.88	0.8	0.48
35	0.78	0.74	0.63	32	0.88	0.78	0.44

Appendix B.4.

Item difficulty, infit MNSQ, outfit MNSQ, and PTMEA correlation for SKT-science.

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
2	1.02	1.39	0.03	18	1.12	1.17	0.01
32	0.94	0.86	0.28	11	0.95	0.99	0.31
15	0.96	0.9	0.25	37	0.91	0.88	0.42
7	0.9	0.72	0.4	39	1.03	1.03	0.19
27	1.09	1.25	-0.02	26	1.00	1.01	0.25
28	1.03	1.01	0.13	35	1.01	1.00	0.24
24	1.06	1.25	0.02	14	0.92	0.9	0.4
3	0.97	0.93	0.27	21	0.95	0.94	0.34

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
34	1.06	1.15	0.07	38	0.95	0.94	0.35
36	0.93	0.84	0.36	10	0.94	0.92	0.36
4	1.05	1.08	0.11	1	1.00	0.99	0.25
29	1.02	1.02	0.17	16	0.95	0.95	0.34
40	1.10	1.26	-0.02	8	1.03	1.07	0.18
30	1.07	1.17	0.06	33	0.96	0.96	0.31
6	1.10	1.12	0.05	9	1.03	1.10	0.17
31	0.98	0.92	0.29	12	1.02	1.11	0.19
19	1.07	1.15	0.07	13	1.03	1.03	0.18
5	1.00	1.00	0.24	20	0.97	0.95	0.3
17	1.00	1.00	0.23	22	1.00	1.00	0.23
23	0.86	0.82	0.5	25	0.96	0.93	0.31

Appendix B.5.

Item difficulty, infit MNSQ, outfit MNSQ, and PTMEA correlation for SKT-social science.

Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation	Number entry	INFIT MNSQ	OUTFIT MNSQ	PT-measure correlation
37	0.88	0.73	0.4	17	0.97	0.93	0.3
21	0.92	0.93	0.31	12	0.99	1.01	0.26
13	0.9	0.75	0.4	4	0.94	1.02	0.31
11	0.98	1.05	0.23	23	1.01	1.05	0.2
19	0.94	0.88	0.34	39	0.98	0.94	0.29
2	0.99	0.99	0.24	35	0.89	0.85	0.43
7	1.02	1.15	0.15	6	1.02	1.00	0.21
9	0.97	0.96	0.27	15	0.95	0.92	0.33
34	0.95	0.9	0.32	16	1.07	1.28	0.07
18	1.01	1.09	0.18	25	1.03	1.00	0.2
27	0.97	0.92	0.29	24	0.99	0.97	0.26
1	1.03	1.07	0.17	32	0.94	0.92	0.34
3	0.98	0.94	0.28	14	1.12	1.13	0.03
38	1.05	1.02	0.17	26	1.02	1.00	0.2
10	1.03	1.10	0.16	29	1.04	1.05	0.16
31	0.95	0.89	0.34	36	0.92	0.9	0.37
8	0.92	0.86	0.38	40	1.11	1.40	0
20	0.93	0.87	0.37	30	0.95	0.92	0.33
22	1.05	1.07	0.15	33	1.13	1.55	-0.06
5	1.01	0.98	0.24	28	1.08	1.06	0.1