INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN ECONOMICS & MANAGEMENT SCIENCES

OPEN ACCESS JOURNAL

Vol 10, Issue 3, (2021) E-ISSN: 2226-3624

Analysis of Validity and Reliability of Economic Achievement Test Based on Rasch Measurement Model

Noornadiah Md. Sari & Khoo Yin Yin

Faculty of Management and Economics, Sultan Idris Education University

Email: khoo@fpe.upsi.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJAREMS/v10-i3/11441 DOI:10.6007/IJAREMS/v10-i3/11441

Published Online: 27 September 2021

Abstract

Teachers regularly attend assessments on students to identify students' levels of mastery. Achievement tests as a quality measurement tool are quintessential so that the conclusions obtained are reliable and significant. Therefore, high-quality achievement tests need to satisfy specific criteria by going through standard procedures. Nonetheless, time and competency constraints lead teachers to utilise economic test questions that do not reach specific standards. Therefore, this research intended to develop and test the validity and reliability of economic achievement tests. An economic achievement test instrument was developed, consisting of 30 objective questions based on Bloom's taxonomy. The testing of the instrument involved 40 respondents of Form Six economics students. The researchers appointed five experts to evaluate the validity of the content of the achievement test questions. At the same time, the construct validity and instrument reliability test analysis involved item-respondent reliability analysis, item-respondent separation index, Cronbach's alpha, item polarity, item fit, standardised residual item correlation and respondent item-ability difficulty level distribution using Rasch measurement approach through Winsteps software 3.72.3. The data of the tests conducted determined that the achievement test confirmed good content validity and reliability values. The analysis also established that six questions needed to be modified. The development of the economic achievement test offers an alternative measurement design over future performance test testing. The researchers proposed that the implementation of this measurement on other subjects too.

Keywords: Rasch Measurement Model, Validity and Reliability, Economic Achievement Test, Academic Achievement.

Introduction

In 2013, a new examination format was launched at the Malaysian Higher School Certificate level and was administered three times at the end of each semester. The centralised examination is conducted by the Malaysian Examinations Council. Meanwhile, assessment at the school level is handled by teachers. Therefore, students have to go through a series of summative and formative tests throughout the study period for three semesters in Form Six. The assessment process is a systematic process of collecting and processing

information using specific measuring tools to identify the level of mastery of student achievement to improve the teaching and learning process (Amua-Sekyi, 2016; Black & Wiliam, 2006). Examinations and tests have become one of the prevalent assessment methods in schools. Consistent with students' level of maturity, the measurement of the cognitive level at the Form Six level demands students reflect creatively, critically, analytically, and in higher-order (MPM, 2012). Following the Malaysian Certificate of Education, economics subjects are also offered for Form Six students. Therefore, the institution and advancement of high-quality economic achievement test paper measuring instruments should be emphasised because it is instrumental in defining the level of mastery of student learning. Hopefully, economics students will have the capacity to consider creatively and critically, practice problem-solving initiatives, foster self-confidence, maintain resilience, operate positive thinking and technological skills in the face of the dynamic dynamics of 21st-century calls (Gordanier et al., 2019; Lopes et al., 2015).

Annually, teachers in schools produce achievement test measuring tools to gauge the level of student achievement. The skill of building and directing student achievement measures is one of the components of assessment competencies that must be mastered by every teacher (American Federation of Teachers National Council on Measurement in Education National Education Association, 2009). Nevertheless, Rosmawati (2008) revealed that teachers often use questions that do not follow the procedure to students. Even though a decade has passed, researchers Arumugham (2020); Sumaryanta (2018); Yan et al (2021) still discovered that there were weaknesses of teachers in implementing student assessment, and teachers still need guidance. Thus, the results obtained were less valid, and interpretations of the level of student ability and follow-up actions became less accurate (Mclellan, 2007). Test results did not provide information on what has been mastered and what has not been mastered by a student to help teachers better teaching and learning in the classroom. The test instrument development process did not follow the proper standards. This situation authenticated that there were still gaps that need improvement and given attention by researchers regarding existing teacher practices. Therefore, this article aimed to discuss developing and evaluating achievement test instruments focused on economics subjects (Semester 1).

In schools, the most conventional method used is through examinations or tests to measure the extent of students' mastery of a learning topic (DeLuca & Volante, 2016). Measurement of student achievement is generally implemented formatively or summatively depending on the purpose of the assessment (Shiel, 2017). Broadfoot and Black (2004) considered that high-quality assessment recognises the extent to which students' mastery of a learning topic, while teachers implement reflections on current teaching methods to promote teaching approaches in the future. In addition, the findings provided information to parents regarding children's performance in school. Thus, developing a good instrument can provide valid and reliable information (Wu et al., 2015). The use of achievement tests as a measure of students has become commonplace by teachers in schools. Achievement tests are measurement tools that are objective (Koretz, 2002). These methods involve a clear learning syllabus, objectives and learning objectives, offering easy and quality feedback and involving a variety of learning approaches (Jimaa, 2011). Ebel and Frisbie (1991) maintained that a good test holds easy items for low-ability students and difficult items for high-ability students. The questions of a test should include low and high-level questions (Cecilio-Fernandes et al., 2018). Referring to Bloom's taxonomy, the low-level domain consists of knowledge, understanding and application, while the high-level domain consists of analysis,

synthesis and evaluation (Bloom et al., 1971; Jimaa, 2011). It allows each category of students to be tested fairly. Thus, developing a practical measurement test must satisfy the standard evaluation criteria through specific and systematic procedures.

Once the achievement test is developed, the implementation of testing on the constructed questions is also imperative to identify the capability of the measurement tool. It confirms the extent to which the quality of the questions and tests will be administered. The Rasch measurement model can prove the validity of an instrument and the item quality of an instrument (Azarilah et al., 2013; Boone, 2016). Compared to classical test theory, respondent characteristics and item characteristics are inseparable (Abu Bakar & Bhasah, 2008; Bambang, 2017). It signifies that the ability of the respondent is only obtained based on the test score. Meanwhile, the Rasch measurement model can measure student differences according to ability level, determine the difficulty level of test items, construct interpretation determination, construct item unidimensionality and test determination. Thus, this method is one of the more comprehensive alternatives in instrument testing in education. Preliminary education research such as Huei et al (2020); Siti Mistima (2015); Osadebe (2018); Owi et al (2020); Nordin et al (2012) also administered this approach to test the reliability and validity of test instruments and questionnaires.

Research Objective

The first objective of this study was to develop an economic achievement test instrument. The objective of the second study was to identify the validity value of economic achievement test instruments. At the same time, the objective of the third study was to identify the reliability value of economic achievement test instruments.

Research Methodology

The implementation of this pilot study was not intended to make generalisations but instead focused on the clarity of the questions, items, formats, and measurement scales used before the actual study was conducted (Moore et al., 2011). Hence, the selection of the sample was by purposeful sampling. The suggested sample size of Browne's (1995) pilot study is 30 people, while Kieser and Wassmer (1996) mentioned that 30 to 40 people is sufficient. The study sample consisted of 40 Form Six economics students (Semester 1). This study used the Rasch measurement approach through Winsteps 3.72.3 software to construct validity values and instrument reliability. The performance test instrument construction implementation went through several phases: instrument development, expert validity, qualitative analysis, and instrument reliability analysis. Figure 1 shows the instrument testing flow chart used in the study.

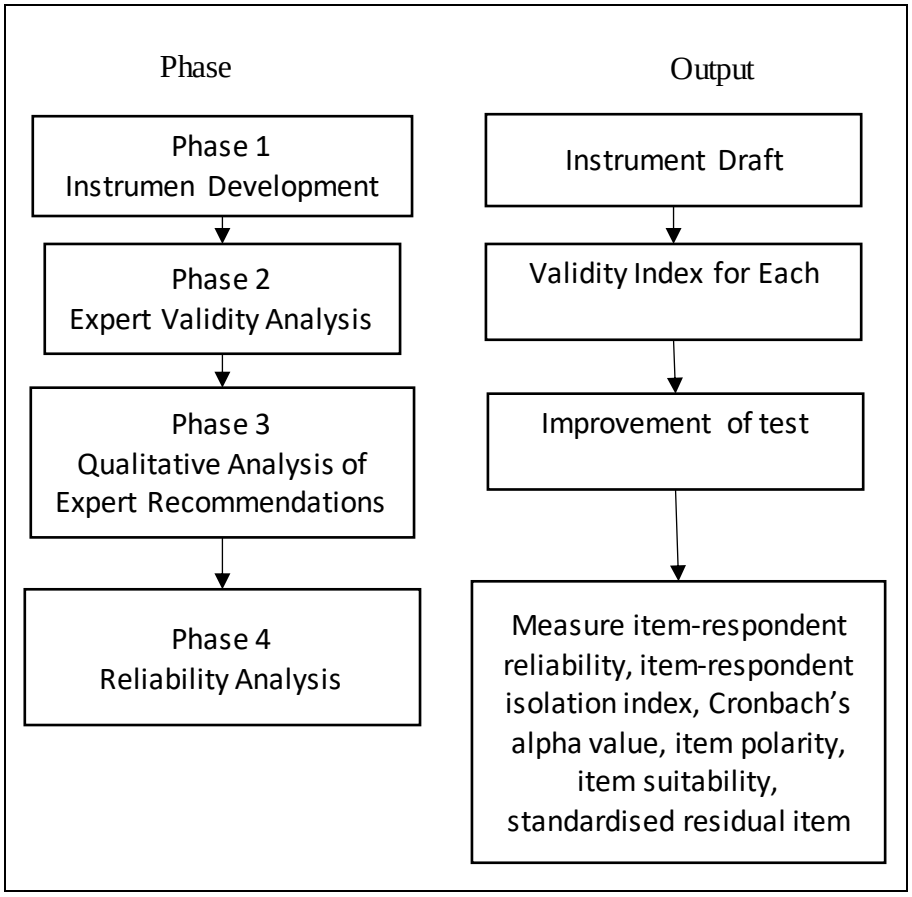

Figure 1: *Instrument testing flow chart*
Source: Modifications of Zaharah & Nurulwahida (2021)

Finding

The researcher presents instrument development, content validity analysis, instrument improvement and reliability analysis in this sub-topic.

Instrument Development

According to Adom et al (2020), developing a test specification table supports improving the content validity of the developed achievement test instruments. The economic achievement test contained 30 objective questions with multiple choices A, B, C, and D and considered six Bloom's cognitive domain levels. The question structure consisted of 40% low-level questions (two knowledge level questions, three comprehension level questions, and five application-level questions) and 60% high-level questions (14 analysis level questions, five synthesis level questions and one assessment level question) according to student level. The time to answer the given question was 60 minutes. Table 1 displays the distribution of questions by taxonomic level.

Table 1
Economic Achievement Test Specification Table

Topic	Question Level						Number of questions
	Low			High			
	Knowledge	Comprehension	Application	Analysis	Synthesis	Assessment	
Introduction	2	2	2				6
Goods Market and Prices		1	2	5	1		9
Production Theory and Production Cost			1	3	1		4
Market Structure, Pricing and Output				4	3	1	8
Factor Market and Distribution				2	1		3
Number of Questions	2	3	5	14	5	1	30
Percentage	40			60			

Expert Validity Analysis

Next, the researchers appointed five experts as instrument validators. This measure can improve the content validity and interface of the instrument (Osadebe, 2015). Adapted from Lynn (1986) and Polit et al. (2007) the selection of experts based on the field and expertise of the experts was made based on the recommendations it includes three or more people. The instrument validation experts consisted of two economics education lecturers, a SISC+ technical and professional officer and two competent economics teachers. The selection of expert samples was made by purposive sampling. Experts were asked to indicate agreement levels 1 (highly irrelevant) to 4 (highly relevant). To obtain the value of the item validity index (I-CVI), the researchers determined the average value of the scale points by summing the scores given by each expert and dividing that value by the number of experts. An acceptable I-CVI value is 0.78 and above, while a value of 0.90 indicates an excellent validity value (Polit et al., 2007; Stewart & Haswell, 2013). The I-CVI value of the economic achievement test in this study was 0.97 that verified high content validity. Table 2 reports the I-CVI economic performance test obtained.

Table 2
Item Validity Index (I-CVI) Economic Achievement Test

Question	Expert A	Expert B	Expert C	Expert D	Expert E	I-CVI
1	1.00	1.00	1.00	0.00	1.00	0.80
2	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	0.00	1.00	1.00	0.80
9	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	0.00	1.00	0.80
11	1.00	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	0.00	1.00	0.80
15	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00	1.00
26	1.00	1.00	1.00	1.00	1.00	1.00
27	1.00	1.00	1.00	1.00	1.00	1.00
28	1.00	1.00	1.00	1.00	1.00	1.00
29	1.00	1.00	1.00	1.00	1.00	1.00
30	1.00	1.00	1.00	1.00	1.00	1.00
AVERAGE	1.00	1.00	0.97	0.90	1.00	0.97

Indication:
1 = expert agreement on constructed items; 0 = disagree on constructed items

Qualitative Analysis of Expert Recommendations

Face validity remains a requirement as a filter for instrument items, even if it has a slightly different purpose than content validity. The feedback received in facial validity will help provide information regarding difficulty level, respondent comprehension, item ambiguity, language levels and grammar (Mousazadeh et al., 2017; Torabizadeh et al., 2016). Therefore, the researchers improvised the instrument by considering the experts' views while reviewing the instrument's items.

Reliability Analysis

The instrument reliability analysis based on Rasch model measurements in this study included item-respondent reliability analysis, item-respondent separation index, Cronbach's alpha. In comparison, item validity involved item polarity, item fit, standardised residual item correlation and item-ability difficulty level distribution of respondents.

Item-Respondent Reliability Index, Item-Respondent Separation Index and Cronbach's Alpha Value

Respondent reliability signifies the probability of repetition of respondent results when given the same instrument (Bambang & Wahyu, 2015). At the same time, the item reliability value symbolises the item's adequacy to measure something to be measured (Azrilah et al., 2013). Based on McMillan and Schumacker (1984), alpha values of 0.70 to 0.90. proves that instrument reliability is good. In this study, the reliability values of the respondents and items were 0.80 and 0.83 were sufficient to signify the reliability of the respondents and the test question items were accepted for use in the actual study. The respondent separation index estimates the separation or difference of a group of individuals according to the level of ability in the measured variables (Bambang & Wahyu, 2014). Whereas the item separation index indicates separation for item difficulty level, it refers to the number of item difficulty strata (Jones & Fox, 1998). Separation values exceeding the value of 2 are good (Fisher, 2007; Linacre, 2007). In this study, the separation index of respondents was 2.02. It validated that the instrument used could distinguish two student ability levels (low-ability and high-ability students). In comparison, the item separation index was 2.21, which confirmed that the item separation consisted of difficult and easy items. Overall, the test questions developed met the reliability standard because Cronbach's alpha value was 0.80. Table 3 summarises the item-respondent reliability index, the item-respondent separation index and Cronbach's alpha values.

Table 3
Summary of Item-Respondent Reliability Index, Item-Respondent Separation Index and Cronbach's Alpha Value

	Reliability	Separation	Alpha Cronbach
Respondent	0.80	2.02	
Item	0.83	2.21	
Overall			0.80

Item Polarity

Item polarity analysis showed that the response correlation on the item or respondent conflicted with the construct, i.e. the item did not work correctly with another item to measure the construct when the value was negative or zero on the Point Measure Correlation (PTMEA CORR.) value (Bond & Fox, 2015). In Table 4, all items show positive value. It proved that each item for the construct could measure the construct to be measured.

Table 4
The Polarity of Economic Achievement Test Instrument Items

Outfit		PT-Measure		Exact	Match	Question
MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1.55	1.72	.07	.41	62.5	71.8	s17
1.36	1.14	.17	.41	70.0	73.9	s05
1.26	.78	.19	.41	65.0	77.6	s03
1.20	.61	.20	.40	75.0	79.4	s04
1.58	2.18	.23	.42	55.0	66.2	s20
1.18	.80	.30	.42	57.5	66.4	s16
1.04	.26	.34	.41	55.0	66.8	s23
1.05	.29	.36	.42	57.5	67.0	s07
.92	-.12	.37	.41	67.5	75.8	s09
.96	-.09	.38	.41	57.5	66.2	s11
.93	-.23	.39	.41	57.5	66.2	s21
.91	-.15	.42	.41	77.5	75.8	s18
. 70	-.12	.44	.33	95.0	90.3	s01
.78	-.51	.48	.41	75.0	77.6	s08
.82	-.68	.49	.42	75.0	67.7	s14
.79	-.88	.51	.42	70.0	66.6	s15
.76	-.79	.51	.41	72.5	71.8	s24
.79	-.74	.51	.41	80.0	70.3	s12
.94	.06	.51	.38	92.5	84.6	s02
.69	-.64	.51	.40	85.0	81.2	s19
.78	-.97	.52	.42	75.0	66.2	s06
.75	-1.07	.54	.42	80.0	66.6	s13
.71	-1.23	.58	.42	85.0	67.7	s25
.65	-1.39	.61	.41	80.0	70.3	s22
.65	-1.62	.62	.42	85.0	66.2	s10

Item Fit

Item fit analysis intends to ascertain the fit of items that measure a construct or latent variable. If there is an item value outside that range, it should be modified or discarded. Wright and Linacre (1994) declared that the accepted index or range for dichotomous data (multiple choice) is 0.7 to 1.3. In addition to the mean squared (MNSQ) infit and outfit values, the z-standardized (ZSTD) infit and outfit values need to be in the range of -2 to +2 (Bond & Fox, 2015). Nevertheless, if the infit and outfit values of the MNSQ are good, then the ZSTD index can be neglected (Linacre, 2012). Based on Table 5, six questions need to be modified, specifically questions 5, 10, 17, 19, 20 and 22, because the values are out of range.

Table 5
MNSQ Infit and Outfit Table

Infit		Outfit		Exact	Match	Question
MNSQ	ZSTD	MNSQ	ZSTD	OBS%	EXP%	
1.18	1.56	1.58	2.18	55.0	66.2	S20
1.41	2.49	1.55	1.72	62.5	71.8	S17
1.31	1.73	1.36	1.14	70.0	73.9	S05
1.30	1.45	1.26	.78	65.0	77.6	S03
1.28	1.24	1.20	.61	75.0	79.4	S04
1.14	1.27	1.18	.80	57.5	66.4	S16
1.10	.87	1.04	.26	55.0	66.8	S23
1.09	.53	.92	-.12	67.5	75.8	S09
1.08	.70	1.05	.29	57.5	67.0	S07
1.07	.68	.96	-.09	57.5	66.2	S11
1.05	.51	.93	-.23	57.5	66.2	S21
1.00	.08	.91	-.15	77.5	75.8	S18
.74	-.89	.94	.06	92.5	84.6	S02
.93	-.29	.78	-.51	75.0	77.6	S08
.92	-.65	.82	-.68	75.0	67.7	S14
.90	-.87	.79	-.88	70.0	66.6	S15
.89	-.78	.79	-.74	80.0	70.3	S12
.89	-.70	.76	-.79	72.5	71.8	S24
.88	-1.06	.78	-.97	75.0	66.2	S06
.87	-.47	.69	-.64	85.0	81.2	S19
.86	-1.32	.75	-1.07	80.0	66.6	S13
.82	-.34	.70	-.12	95.0	90.3	S01
.80	-1.76	.71	-1.23	85.0	67.7	S25
.75	-2.49	.65	-1.62	85.0	66.2	S10
.75	-1.86	.65	-1.39	80.0	70.3	S22

Standardised Residual Item Correlation

Standardised residual item correlation analysis aims to recognise whether the item depends or not between items with other items (Ellyza & Kamisah, 2018). If the correlation value between the items produces a value above 0.7, indicating that such items are interdependent and not singular, then one of the items should be dropped (Linacre, 2012). Based on the findings in Table 6, all items have less than 0.7. These data confirmed that all items could measure the construct to be measured, and no items were overlapping.

Table 6
Correlation of Economic Achievement Test Items

Correlation	*Entry*		*Entry*	
	Number	*Question*	*Number*	*Question*
.49	4	s04	20	s20
.43	3	s03	25	s25
.41	7	s07	15	s15
.41	6	s06	10	s10
-.58	12	s12	17	s17
-.56	4	s04	5	s5
-.54	7	s07	13	s13
-.51	11	s11	24	s24
-.45	3	s03	18	s18
-.44	10	s10	23	s23
-.44	3	s03	13	s13
-.44	16	s16	19	s19
-.43	1	s01	18	s18
-.43	5	s05	10	s10
-.42	6	s06	24	s24
-.42	8	s08	18	s18
-.41	5	s05	8	s8
-.41	12	s12	19	s19
-.40	7	s07	22	s22
-.39	7	s07	14	s14

Distribution of Item Difficulty Levels and Respondents' Abilities

The item difficulty map and the respondent's abilities confirm whether the tests performed are appropriate to the respondent's abilities. Figure 2 shows that the distribution of test questions is almost balanced and can test low-ability and high-ability students. It is crucial to measure each level of student mastery fairly, that is, on students of high and low cognitive ability.

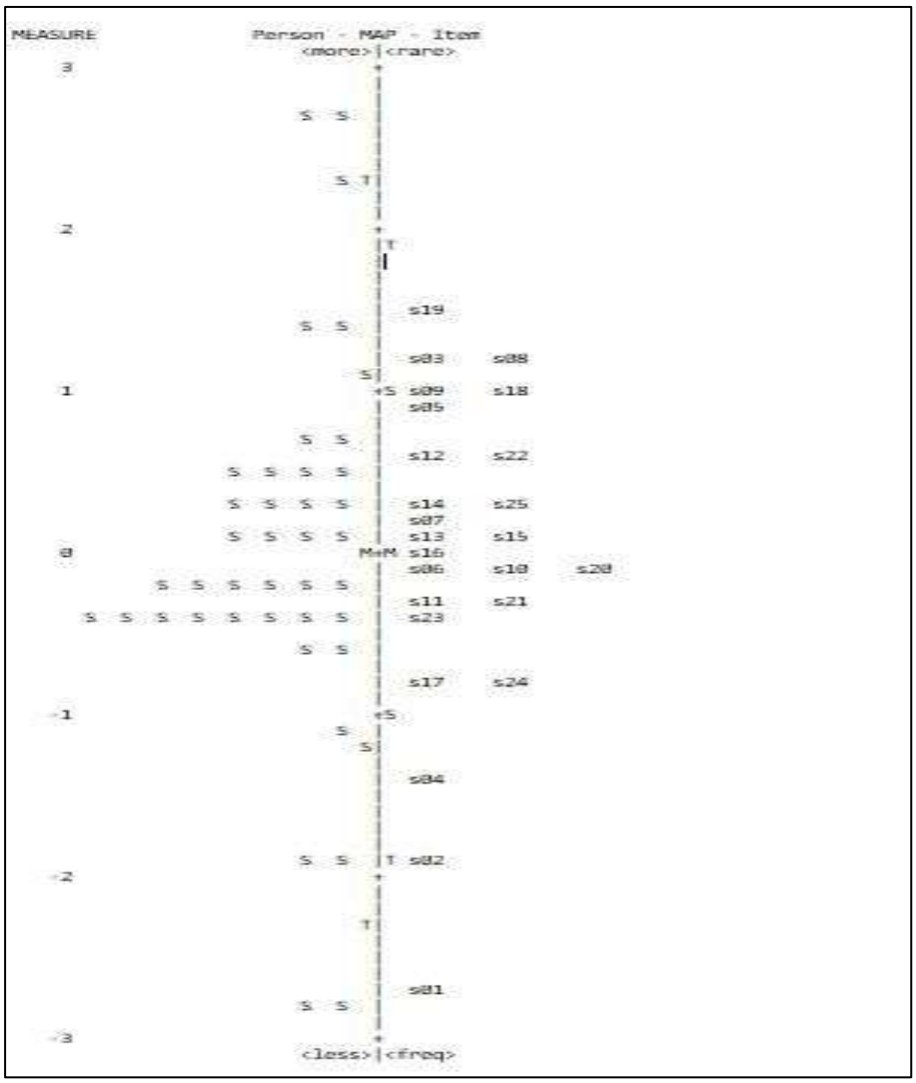

Figure 2: *Wright Map*

Discussion

Developing achievement tests is a systematic process so that the quality of the tests produced is excellent. Empirical analysis proved that there were six questions out of 30 objective questions that needed to be modified. Based on the findings, the Rasch measurement model successfully tested the validity and reliability of the economic performance test instrument. The Rasch approach is a more comprehensive alternative approach than the classical testing approach. It is because this measurement could measure the questions and the respondents' ability at the same time. The study's conclusions established that the developed achievement test questions achieved a good level of validity and reliability. This achievement test instrument was valid and reliable for real students to test the level of mastery of economic learning in Form Six.

However, this study was limited to the construction of 30 objective multiple-choice Economics questions on a small number of respondents of 40 students. The sample was selected through purposive sampling. Based on the findings, the researchers suggested that Economics teachers build a bank of formative and summative economic questions for students' use in the classroom. The production of test questions that meet the standard is important. This is to ensure the quality of the test set administered to the students.

Transparency and fairness in assessment help teachers identify topics of weakness and the individual students who need additional guidance. This allows teachers to perform continuous assessment and is not just at the end of a topic. Indirectly, students can be familiar with the skills of answering actual exam-level questions. This step should also be implemented when developing the tests on other subjects so that the results obtained measure student mastery of the learning topics. In addition, the construction procedure of the test instrument can also be extended to the matriculation diploma and pre-university levels that offer economics courses. Current achievement results help students prepare for their studies in higher education later.

This study has implications for researchers, in general, and economics teachers, in particular, to develop quality test instruments in the future. Furthermore, this approach should be practiced by the teachers who make examination questions at the school, district, and state levels to improve the existing methods. Indirectly, it also enhances teacher assessment competencies in constructing and administering student achievement tests (Okolie et al., 2020). This approach directly contributes to the increasing diversity of assessment alternatives in the development and evaluation of achievement test instruments. Transparency in achievement measurement helps teachers and school administrators plan appropriate student development programs.

Conclusion

A valid and reliable measuring tool can afford accurate information on students' level of mastery of learning topics. Nonetheless, this research was limited to constructing 30 multiple-choice objective economic questions on a few respondents. Sample selection was made by simple sampling. Therefore, the researchers suggested that economics teachers build a bank collection of economics questions for students' use in the classroom. This instrument is not only limited to use among Form Six students. It could also be extended to the Matriculation and Diploma levels that offer economics courses at the pre-university level. Hence, this research could guide teachers and researchers to compose high-quality achievement tests.

References

Abu Bakar, N., & Bhasah, A. B. (2008). *Penaksiran dalam pendidikan & sains sosial*. Penerbit Universiti Pendidikan Sultan Idris.

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement and evaluation: Understanding and use of the concepts in education*. International Journal of Evaluation and Research in Education,* 9(1), 109-119. https://doi.org/10.11591/ijere.v9i1.20457

Amua-Sekyi, E. T. (2016). Assessment, student learning and classroom practice: A review. *Journal of Education and Practice,* 7 (21).

Arumugham, K. S. (2020). Curriculum, teaching and assessment in the perspective of classroom assessment. *Asian People Journal*, 3(1), 152-161.

Azarilah, A. A., Saidfudin, M., & Azami, Z. (2013). *Asas model pengukuran rasch: Pembentukan skala & struktur pengukuran.* Penerbit Universiti Kebangsaan Malaysia.

Bambang, S. (2017). Rasch Model Measurement as Tools in Assessment for Learning. *International Conference on Educational Innovation (ICEI 2017), Wyndham Hotel, Surabaya, Indonesia*. https://doi.org/10.2991/icei-17.2018.11

Bambang, S., & Wahyu, W. (2014). Aplikasi model rasch untuk penelitian ilmu-ilmu sosial. Trim Komunikata Publishing House.

Bambang, S., & Wahyu, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim Komunikata Publishing House.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds) (1971). *Handbook on the formative and summative evaluation of student learning.* McGraw-Hill.

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences (3rd ed.).* Lawrence Erlbaum Associates.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education,* 15(4). https://doi.org/10.1187/cbe.16-04-0148

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice*, 11(1), 7-26, 10.1080/0969594042000208976

Browne, R.H. (1995). On the use of a pilot study for sample size determination. Statistics in Medicine, 14, 1933-1940.

Cecilio-Fernandes, D., Cohen-Schotanus, J., & Tio, R. A. (2018). Assessment programs to enhance learning. Physical Therapy Reviews, 23(1), 17-20. https://doi.org/10.1080/10833196.2017.1341143

DeLuca, C., & Volante, L. (2016). Assessment for learning in teacher education programs: Navigating the juxtaposition of theory and praxis. *Journal of the International Society for Teacher Education,* 20 (1), 19-31.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement (5th edition),* Prentice-Hall, Englewood Cliffs.

Ellyza, K., & Kamisah, O. (2018). Kesahan dan kebolehpercayaan ujian kemahiran proses sains untuk murid sekolah rendah berdasarkan model pengukuran rasch. *Jurnal Pendidikan Malaysia,*1-9. http://dx.doi.org/10.17576/JPEN-2018-43.03-01

Fisher Jr., W.P. (2007). *Rating scale instrument quality criteria*. Rasch Measurement Transaction, 21, 1095. http://www.rasch.org/rmt/rmt211a.htm

Gordanier, J., Hauk, W., & Sankaran, C. (2019). Early intervention in college classes and improved student outcomes. *Economics of Education Review*, 72, 23–29. https://doi.org/10.1016/j.econedurev.2019.05.003

Huei, O. K., Rus, R. C., & Kamis, A. (2020). Knowledge of design and technology subject: A rasch measurement model approaches for pilot study. *International Journal of Academic Research Business and Social Sciences,* 10(3), 599–613.

Jimaa, S. (2011). The impact of assessment on students learning. *Procedia - Social and Behavioral Sciences*, 28, 718–721. https://doi.org/10.1016/j.sbspro.2011.11.133

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*, 8, 941-949.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources,* 37(4), 752. https://doi.org/10.2307/3069616

Linacre, J. M. (2007). *A user's guide to WINSTEPS Rasch-model computer programs*. MESA Press.

Linacre, J.M. (2012). *User's guide and program manual to WINSTEPS: Rasch model computer programs.* MESA Press.

Lopes, J. C., Graça, J. C., & Correia, R. G. (2015). Effects of economic education on social and political values, beliefs and attitudes: Results from a survey in Portugal. *Procedia Economics and Finance*, 30, 468–475. https://doi.org/10.1016/S2212-5671(15)01314-3

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, 35, 378-382.

Majlis Peperiksaan Malaysia (MPM). (2012). *Huraian sukatan pelajaran ekonomi.* Majlis Peperiksaan Malaysia.

Mclellan, E. (2007). What is a competent "competence standard"? *Quality Assurance in Education,* 15(4), 437–448. https://doi.org/10.1108/09684880710829992

McMillan, J. H., & Schumacher, S. (1984). *Research In Education*. Little, Brown & Company Limited.

Moore, C. G., Carter, R. E., Nietert, P. J., & Stewart, P. W. (2011). Recommendations for planning pilot studies in clinical and translational research. *Clinical and Translational Science,* 4(5), 332–337. https://doi.org/10.1111/j.1752-8062.2011.00347.x

Mousazadeh, S., Rakhshan, M., & Mohammadi, F. (2017). Investigation of content and face validity and reliability of sociocultural attitude towards appearance questionnaire-3 (SATAQ-3) among female adolescents. *Iranian Journal of Psychiatry*, 12(1), 15–20.

Nordin, A. R., Zamri, A. K., & Lei, M. T. (2012). Examining quality of mathematics test item using rasch model: Preminarily analysis. *Procedia-Social and Behavioral Sciences*, 69, 2205-2214.

Okolie, U. C., Igwe, P. A., Nwajiuba, C. A., Mlanga, S., Binuomote, M. O., Nwosu, H. E., & Ogbaekirigwe, C. O. (2020). Does PhD qualification improve pedagogical competence? A study on teaching and training in higher education. *Journal of Applied Research in Higher Education,* 12(5), 1233–1250. https://doi.org/10.1108/JARHE-02-2019-0049

Osadebe, P. U. (2015). Construction of valid and reliable test for assessment of students. *Journal of Education and Practice*, 6(1).

Osadebe, P. U. (2018). Assessment of test items with rasch measurement model. *Journal of Applied Measurement*, 19(1), 106–112.

Owi, K. H., Ridzwan, C. H., & Arasinah, K. (2020). Knowledge of design and technology subject: A rasch measurement model approaches for pilot study. *International Journal of Academic Research Business and Social Sciences,* 10(3), 599–613. http://dx.doi.org/10.6007/IJARBSS/v10-i3/7075

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467. https://doi.org/10.1002/nur.20199

Rosmawati, M. (2008). *Pengesanan dan penggunaan ujian matematik tahun empat sekolah rendah: Analisis rasch* [Unpublished doctoral dissertation]. University of Science Malaysia.

Shiel, T. (2017). *Chapter 2 building the base: begin with the end in mind. In Designing and Using Performance Tasks: Enhancing Student Learning and Assessment*, 25-40. Corwin. https://www-doi-org.ezplib.ukm.my/10.4135/9781506343402.n3

Siti Mistima, M. (2015). Psychometric evaluation on mathematics beliefs instrument using rasch model. *Creative Education*, 6, 1797-1801.

Stewart, J., & Haswell, K. (2013). Assessing readiness to work in primary health care: The content validity of a self-check tool for physiotherapists and other health professionals. *Journal of Primary Health Care*, 5(1), 70–73.

Sumaryanta, Mardapi, D., Sugiman, & Herawan, T. (2018). Assessing teacher competence and its follow-up to support professional development sustainability. *Journal of Teacher Education for Sustainability,* 20 (1), 106-123.

Torabizadeh, C., Yousefinya, A., Zand, F., Rakhshan, M., & Fararooei, M. (2016). A nurses' alarm fatigue questionnaire: development and psychometric properties. *Journal of Clinical Monitoring and Computing,* 31(6), 1305–1312. https://doi.org/10.1007/s10877-016-9958-x

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Wu, X. V., Enskär, K., Lee, C. C. S., & Wang, W. (2015). A systematic review of clinical assessment for undergraduate nursing students. *Nurse Education Today*, 35(2), 347–359. https://doi.org/10.1016/j.nedt.2014.11.016

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, 28(3), 228-260. https://doi.org/10.1080/0969594X.2021.1884042

Zaharah, C. I., & Nurulwahida, A. (2021). Analisis statistik kesahan dan kebolehpercayaan ujian pencapaian reka bentuk elektrik. *Malaysian Journal of Social Sciences and Humanities*, 6(8), 196-206.