

ANALISIS ITEM BAGI UJIAN SIRAH TAHUN EMPAT DI SEKOLAH AGAMA NEGERI SABAH MENGGUNAKAN MODEL PENGUKURAN RASCH Item Analysis for Year Four Sirah Test in State Religious Schools of Sabah Using the Rasch Measurement Model

NORAZLY NORDIN

Department of Social Science and Hospitality, Tunku Abdul Rahman University of
Management and Technology, Sabah Branch
norazly@tarc.edu.my

HARRIS SHAH ABD HAMID

Department of Psychology, Faculty of Medicine, Universiti Kuala Lumpur, Royal College of
Medicine, Perak

MUHAMMAD DAOH

Faculty of Education and Social Science, Universiti Selangor

Abstract

This study aims to analyze the quality of the questions in the Sirah test set for Year Four students at a state religious school in the state of Sabah. The test set was developed by the teacher at the school and consists of 40 multiple-choice questions. The test was administered to 22 students and then analyzed using WINSTEP software to obtain reliability values and fit values for each question. The analysis results showed that the item reliability of the test set was acceptable (0.76), and there were four items that did not meet the fit criteria as determined by the Rasch measurement model, namely items 19, 23, 38, and 37. In terms of item difficulty ranking, item 40 was the most difficult, with the highest logit value (2.85), while the easiest item was item 15, with the lowest logit value (-1.98). To further improve the quality of this test set, refinement and revision can be done by the teacher, and the Rasch Measurement Model is seen as an appropriate approach to analyze the psychometric characteristics of a test.

Keywords: *Item Analysis, Rasch Measurement Model, Achievement Test*

Abstrak

Kajian ini bertujuan untuk menganalisis kualiti soalan yang terdapat di dalam set ujian Sirah untuk murid-murid tahun empat di sebuah sekolah agama negeri di negeri Sabah. Set ujian telah dibangunkan oleh guru di sekolah berkenaan yang mengandungi 40 butir soalan aneka jawapan. Set ujian tersebut ditadbir ke atas 22 orang murid dan kemudian dianalisis menggunakan perisian WINSTEP untuk mendapatkan nilai kebolehpercayaan dan nilai kesesuaian bagi setiap butir soalan. Hasil analisis mendapati nilai kebolehpercayaan item bagi set ujian adalah diterima (0.76) dan terdapat tiga item yang tidak menepati nilai kesesuaian (fit) yang ditetapkan model pengukuran Rasch iaitu item 19, 38, dan 37. Bagi susunan aras kesukaran item, item 40 merupakan item yang paling sukar kerana menunjukkan nilai logit yang paling tinggi (2.85) manakala item yang paling mudah adalah item 15 dengan nilai logit yang paling rendah (-1.98). Bagi meningkatkan lagi kualiti set ujian ini, pemurnian dan

pembaikan boleh dilakukan oleh guru dan Model Pengukuran Rasch dilihat sebagai satu pendekatan yang sesuai untuk menganalisis ciri-ciri psikometrik bagi sesuatu ujian.

Kata kunci: *Analisis Item, Model Pengukuran Rasch, Ujian Pencapaian.*

PENGENALAN

Sekolah Agama Negeri atau singkatannya SAN adalah sebuah institusi pendidikan formal yang disediakan oleh Kerajaan Negeri Sabah menerusi Jabatan Hal Ehwal Agama Islam Negeri Sabah (JHEAINS). Bahagian Pendidikan JHEAINS adalah badan yang bertanggungjawab dalam mengurus dan memantau segala aktiviti pendidikan di semua sekolah agama negeri di Sabah. Sehingga kini, terdapat 105 buah sekolah agama negeri di seluruh negeri Sabah yang meliputi enam bahagian iaitu bahagian Tawau, Sandakan, Pedalaman Atas, Pantai Barat Utara, Pedalaman Bawah dan Pantai Barat Selatan. Kesemua sekolah agama negeri di Sabah mempunyai kelas untuk murid tahun satu hingga tahun enam sama seperti sekolah-sekolah formal di bawah Kementerian Pendidikan Malaysia.

Sama seperti institusi pendidikan formal yang lain, SAN juga mempunyai kurikulum dan sistem penilaian formalnya sendiri yang berbentuk peperiksaan akhir atau ujian Sijil Rendah Ugama atau SRU. SRU ditadbir pada akhir pembelajaran iaitu pada murid tahun enam. Set peperiksaan SRU dikawalselia oleh Unit Peperiksaan di bawah Bahagian Pendidikan JHEAINS. Namun begitu, peperiksaan bulanan juga boleh diadakan di peringkat sekolah dan guru mata pelajaran pula diberikan autonomi untuk membangunkan set soalnya sendiri. Sehubungan dengan itu, guru perlu memastikan set soalan yang dibangunkan mematuhi kriteria psikometrik untuk menjamin kualiti set ujian dalam mengukur pengetahuan, kemahiran dan penguasaan murid (Autaeva et al, 2020). Justeru, kajian ini membuat satu penilaian psikometrik terhadap satu ujian bagi mata pelajaran Sirah yang termasuk dalam kurikulum SRU.

TINJAUAN LITERATUR

Menurut Mohamad Najib Abdul Ghafar (2011), ujian merupakan satu alat atau prosedur yang sistematik dalam mengukur sesuatu sampel tingkah laku. Ujian juga digunakan sebagai satu alat untuk menentukan prestasi murid dalam sesuatu bidang. Ujian memainkan peranan yang penting dalam proses pendidikan terutamanya dalam penentuan pencapaian murid, penyediaan maklum balas tentang proses pendidikan, penambahbaikan kandungan kurikulum serta penting dalam merangka intervensi dalam membantu melonjakkan lagi potensi murid (Yang et al., 2021; Bilyakovska, 2022; Imankasymova, 2022; Berman, Feuer, & Pellegrino, 2019). Oleh itu, ujian menjadi perkara yang penting dalam institusi pendidikan bermula dari peringkat rendah hinggalah ke universiti.

Sehubungan dengan itu, ujian perlulah dibangunkan dengan penuh teliti dan bersistematik. Ujian yang baik mempunyai beberapa kriteria seperti mana yang telah dinyatakan oleh beberapa pengkaji lalu. Kriteria-kriteria ujian yang baik adalah seperti yang berikut:

- i. Mempunyai nilai kebolehpercayaan yang baik. Kebolehpercayaan merujuk kepada tahap kestabilan dan konsistensi sesuatu alat pengukuran atau penilaian dalam memberikan hasil yang sama apabila digunakan dalam keadaan yang sama. Dalam konteks pendidikan dan psikologi, kebolehpercayaan adalah penting untuk memastikan bahawa hasil ujian, kajian, atau penilaian adalah boleh dipercayai dan konsisten dari satu masa ke masa yang lain (Fathin & Retnawati, 2020).

- ii. Mempunyai nilai kesahan yang baik. Kesahan merujuk kepada sejauh mana suatu alat pengukuran, ujian, atau penilaian mengukur apa yang sepatutnya diukur. Dalam erti kata lain, ia menilai keupayaan alat tersebut untuk memberikan hasil yang tepat dan relevan berkaitan dengan tujuan pengukurannya (Fathin & Retnawati, 2020).
- iii. Sesuai dengan Objektif Pembelajaran. Kandungan ujian harus selaras dengan objektif pembelajaran dan kurikulum bagi sesuatu mata pelajaran (Bilyakovska, 2022).
- iv. Mempunyai aras kesukaran yang pelbagai. Set ujian perlu mempunyai kepelbagaian dalam aras kesukaran soalan untuk tujuan ketepatan dalam mengukur kebolehan sebenar murid (Ikhsanudin et al., 2023).
- v. Ujian juga mempunyai satu jawapan yang standard untuk menjamin ketelusan dan keadilan (Bilyakovska, 2022).

Dalam membina ujian yang baik terdapat beberapa langkah yang dicadangkan oleh pengkaji-pengkaji yang lepas seperti menetapkan butiran soalan mengikut kandungan sukatan pelajaran di dalam Jadual Spesifikasi Ujian, menentukan format dan bilangan soalan, menjalankan kajian rintis untuk menilai kualiti ujian seperti mengenal pasti nilai kesahan, kebolehpercayaan dan analisis butir-butir ujian (Hughes, 2002; Foertsch, 2014; Poehner & Lantolf, 2005).

Analisis Butir dalam Ujian Pencapaian

Analisis butir ujian merujuk kepada proses menilai dan meneliti item-item atau soalan-soalan dalam ujian untuk memastikan kualiti dan keberkesannya. Tujuannya adalah untuk menilai sejauh mana setiap item dalam ujian memenuhi tujuan penilaian dan memberi maklumat berguna tentang prestasi murid. Analisis butir atau juga dikenali analisis item ialah proses yang membantu guru menilai kualiti item ujian dari segi kesukaran, diskriminasi, kesahan dan kebolehpercayaan.

Karim, Sudiro dan Sakinah (2021) menyatakan bahawa kebanyakan guru tidak menjalankan analisis item dalam set ujian yang dibangunkan. Perkara ini disebabkan guru tidak memahami proses dan kepentingan menjalankan analisis tersebut. Oleh itu, setelah membina item, set soalan akan terus ditadbir kepada murid dan tidak menjalani analisis yang penting ini (Taufiq et al., 2019). Tambahan pula, kekurangan latihan terhadap pembinaan ujian dan analisis item menjadi salah factor penyumbang kepada permasalahan ini. Arhin, Annan-Brew, dan Owusuaah (2023) mengatakan bahawa guru yang berpengalaman dan mendapat latihan cenderung untuk menghasilkan item yang lebih baik berbanding dengan guru yang kurang berpengalaman.

Kebanyakan kajian menggunakan pendekatan Teori Ujian Klasik dalam menjalankan analisis item (Ashraf & Jaseem, 2020; Saat, 2020; Arhin et al., 2023). Analisis item yang berpandukan teori ini akan menjalankan pengiraan terhadap aras kesukaran dan indeks diskriminasi item. Mohamad Najib Abdul Ghafar (2011) menyatakan bahawa aras kesukaran adalah susunan bagi soalan susah ataupun senang manakala indeks diskriminasi bermaksud kemampuan soalan dalam membezakan kebolehan antara murid. Secara mudahnya, analisis item dalam satu set ujian dijalankan untuk menimbangtara kualiti item dengan berpandukan hasil analisis terhadap aras kesukaran item, indeks diskriminasi, kesahan dan kebolehpercayaan (Nuroini et al., 2021; Ashraf & Jaseem, 2020; Xie & Cobb, 2020).

Model Pengukuran Rasch dalam Analisis Item

Model pengukuran Rasch ialah model statistik kebarangkalian yang digunakan dalam psikometrik untuk menganalisis data daripada penilaian untuk mengukur pembolehubah yang terpendam seperti kebolehan, sikap, dan ciri personaliti. Model ini mendapat perhatian kerana

keupayaannya untuk mengubah data ordinal kepada pengukuran aras selang (*interval level measurements*) untuk memastikan bahawa skala pengukuran adalah tidak berubah merentas kumpulan dan konteks yang berbeza (Briggs, 2019). Model ini juga digunakan dalam pelbagai bidang seperti psikometrik, pendidikan, sains kesihatan, pentaksiran bahasa, rehabilitasi, sains sosial untuk membangunkan instrumen pengukuran yang sah dan dipercayai (Aryadoust et al., 2020; Xie, 2023; Engelhard, & Wang, 2019; Tesio et al., 2023).

Dalam konteks pengujian, Model Pengukuran Rasch juga dilihat sebagai sebuah model yang berkesan dalam menentukan nilai kebolehpercayaan dan kesahan dalam setiap item. Beberapa kajian telah dijalankan untuk menentukan aras kebolehpercayaan dan kesahan item bagi ujian matapelajaran di sekolah seperti matematik dan Bahasa inggeris seperti kajian yang dijalankan oleh Atikah et al. (2022), Mokshein et al. (2019), Asriadi dan Hadi (2021), dan Rizbudiani et al. (2021). Hasil daripada kajian-kajian tersebut mendapati bahawa model pengukuran Rasch berkemampuan untuk mengenalpasti item-item yang tidak sesuai bagi satu-satu instrumen. Selain itu, model ini juga dapat mengkategorikan item berdasarkan aras kesukaran item dan seterusnya menjadikan sesuatu ujian menjadi lebih seimbang dan tepat dalam mengukur kepelbagaian kemampuan murid. Selain itu, pengkategorian ini juga membantu menyemak kualiti ujian samaada mengandungi soalan yang mudah ataupun sukar (Rizbudiani et al., 2021).

Sumintono dan Widhiarso (2015) menyatakan beberapa kriteria yang perlu diperhatikan dalam menggunakan model pengukuran Rasch dalam menjalankan analisis item. Kriteria-kriteria tersebut dijelaskan di dalam Jadual 1 di bawah:

Jadual 1: Kriteria Analisis Item Menggunakan Model Pengukuran Rasch

Analisis	Kriteria Model Rasch
Aras kesukaran Item	<i>Item Measure</i> (Nilai Logit yang tinggi menunjukkan item adalah sukar)
Aras Kesesuaian Item	Digunakan untuk menentukan item sesuai atau fit dan berfungsi dengan baik di dalam ujian. Item dianggap fit apabila Nilai <i>Outfit Mean Square</i> (MNSQ) di dalam julat 0.5 hingga 1.5; Nilai <i>Outfit Z-Standard</i> (ZSTD) di dalam julat -2.0 hingga 2.0; dan Nilai <i>Point Measure Correlation</i> (PTMEA): 0.4 hingga 0.85. (Boone et al, 2014)
Kebolehpercayaan ujian	Nilai Kebolehpercayaan Individu (<i>person</i>) dan Item >0.67 dan Indeks pemisah ≥ 2

Matlamat utama kajian ini ialah menilai ciri-ciri psikometrik ujian mata pelajaran Sirah bagi pelajar tahun empat di SAN di bawah kelolaan JHEAINS. Penilaian ini merangkumi kualiti ujian secara keseluruhan dan kualiti item yang digunakan.

METODOLOGI

Kajian ini merupakan kajian kualitatif yang menggunakan reka bentuk tinjauan. Instrumen merupakan ujian bulanan bagi matapelajaran Sirah bagi murid tahun empat di sebuah Sekolah Agama Negeri (SAN) di Negeri Sabah.

Responden Kajian

Seramai 22 orang murid tahun empat di sebuah Sekolah Agama Negeri (SAN) di Negeri Sabah telah menduduki ujian ini. Kaedah persampelan menyeluruh digunakan bagi mengumpul data daripada sekolah berkenaan. Setelah menduduki ujian ini, kertas jawapan murid diambil kembali untuk disemak dan direkodkan jawapan murid di dalam perisian *Microsoft Excel*. Jumlah responden yang diperolehi ini sudah memadai untuk mencapai aras keyakinan 95% dalam model Rasch bagi item dikotomus (Linacre, 1994).

Instrumen Kajian

Instrumen kajian adalah berbentuk ujian yang mengandungi 40 butir soalan aneka jawapan. Setiap soalan dibekalkan dengan tiga pilihan jawapan yang mana hanya ada satu jawapan yang betul dan dua lagi pilihan disifatkan sebagai pengganggu. Xie et al. (2022) menyatakan bahawa pengganggu adalah penting untuk menilai pengetahuan pelajar dengan tepat dengan menyediakan pilihan yang salah tetapi munasabah yang boleh mengelirukan pelajar yang belum memahami bahan pembelajaran sepenuhnya.

Soalan-soalan dibangunkan oleh guru di sekolah berkenaan dengan berpanduan Jadual Spesifikasi Ujian yang telah dibangunkan terlebih dahulu yang meliputi tajuk-tajuk di dalam sukatan pelajaran Sirah untuk tahun empat di Sekolah Agama Negeri, Negeri Sabah. Taburan soalan mengikut topik yang terkandung di dalam sukatan pelajaran adalah seperti Jadual 2 yang berikut:

Jadual 2: Jumlah Butir Soalan Mengikut Tajuk

Tajuk	Bilangan soalan
Kewafatan Rasulullah SAW	9
Khulafa Ar Rasyidin	9
Saidina Abu Bakar As-Siddiq	9
Saidina Umar bin Al-Khattab	10
Soalan KBAT (Kemahiran Berfikir Aras Tinggi)	3
Jumlah Soalan	40 butir soalan

Analisis Data

Data yang direkodkan dalam perisian *Microsoft Excel* kemudian dianalisis menggunakan perisian WINSTEP untuk mendapatkan nilai kebolehpercayaan individu dan item, indeks pemisah, nilai kesesuaian item dan juga nilai logit bagi setiap item untuk tujuan menentukan aras kesukaran item bagi set ujian ini. Model pengukuran Rasch bagi data dikotomus digunakan dalam analisis yang dilakukan.

DAPATAN KAJIAN DAN PERBINCANGAN

Hasil kajian mendapati nilai kebolehpercayaan individu bagi set ujian ini adalah baik dan boleh diterima (0.84) dan indeks pemisah pula juga menunjukkan nilai yang baik (2.29). Manakala bagi kebolehpercayaan item pula pada tahap diterima (0.76) manakala indeks pemisah item menunjukkan nilai yang lemah (1.77).

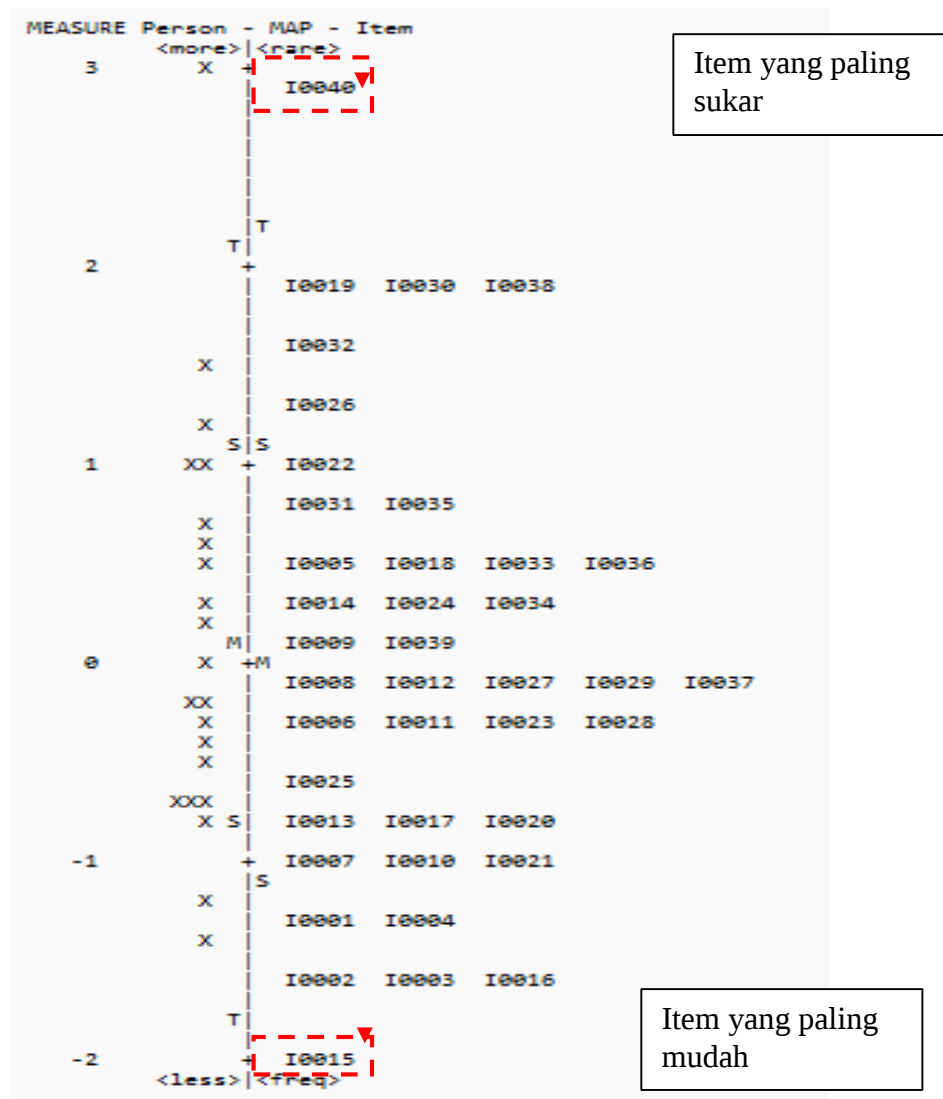
Bagi analisis kesesuaian item pula, terdapat empat item yang tidak menepati nilai kesesuaian (*fit*) yang ditetapkan model pengukuran Rasch iaitu item nombor 19 (MNSQ: 1.59; ZSTD: 1.10

dan PTMEA: .26), Item nombor 37 (MNSQ: 2.49; ZSTD: 4.03 dan PTMEA: -.34), dan Item Nombor 38 (MNSQ: 2.88; ZSTD: 2.47 dan PTMEA: -.23). Jadual 3 di bawah menerangkan dapatan bagi analisis kesesuaian item bagi setiap item di dalam ujian ini.

Jadual 3: Analisis Kesesuaian Item

Item	Nilai MNSQ	Nilai ZSTD	Nilai PTMEA	Interpretasi
1	.81	-.20	.34	Sesuai dan dikekalkan
2	.71	-.27	.36	Sesuai dan dikekalkan
3	.85	-.02	.30	Sesuai dan dikekalkan
4	1.05	.28	.28	Sesuai dan dikekalkan
5	.91	-.27	.46	Sesuai dan dikekalkan
6	.91	-.22	.42	Sesuai dan dikekalkan
7	.83	-.25	.36	Sesuai dan dikekalkan
8	.71	-1.09	.61	Sesuai dan dikekalkan
9	1.16	.70	.26	Sesuai dan dikekalkan
10	.90	-.06	.35	Sesuai dan dikekalkan
11	.63	-1.32	.66	Sesuai dan dikekalkan
12	.69	-1.21	.62	Sesuai dan dikekalkan
13	1.24	.70	.13	Sesuai dan dikekalkan
14	.87	-.47	.53	Sesuai dan dikekalkan
15	.84	.06	.26	Sesuai dan dikekalkan
16	.97	.17	.23	Sesuai dan dikekalkan
17	.79	-.43	.46	Sesuai dan dikekalkan
18	1.32	1.22	.08	Sesuai dan dikekalkan
19	1.59	1.10	.26	Tidak sesuai. Perlu dimurnikan
20	.66	-.86	.57	Sesuai dan dikekalkan
21	.82	-.25	.34	Sesuai dan dikekalkan
22	.87	.46	.46	Sesuai dan dikekalkan
23	.96	-.17	.89	Sesuai dan dikekalkan
24	.87	-.47	.53	Sesuai dan dikekalkan
25	.96	.00	.37	Sesuai dan dikekalkan
26	.59	-1.17	.69	Sesuai dan dikekalkan
27	1.06	.30	.30	Sesuai dan dikekalkan
28	.82	-.52	.48	Sesuai dan dikekalkan
29	1.06	.32	.28	Sesuai dan dikekalkan
30	.95	.09	.44	Sesuai dan dikekalkan
31	.96	-.05	.41	Sesuai dan dikekalkan
32	1.06	.27	.32	Sesuai dan dikekalkan
33	1.01	.13	.37	Sesuai dan dikekalkan
34	.88	-.41	.47	Sesuai dan dikekalkan
35	.79	-.69	.53	Sesuai dan dikekalkan
36	1.07	.34	.38	Sesuai dan dikekalkan
37	2.49	4.03	-.34	Tidak sesuai. Perlu dimurnikan
38	2.88	2.47	-.23	Tidak sesuai. Perlu dimurnikan
39	.86	-.50	.48	Sesuai dan dikekalkan
40	1.32	.62	.09	Sesuai dan dikekalkan

Bagi aras kesukaran soalan pula, susunan aras kesukaran dapat diperhatikan di dalam Peta Wright yang dihasilkan menerusi perisian WINSTEP. Item yang terletak pada kedudukan yang paling tinggi merupakan item yang sukar kerana mempunyai nilai logit yang tinggi dan begitu juga sebaliknya. Di dalam set ujian ini, item yang paling sukar adalah Item 40 (2.85) dan paling mudah pula adalah item 15 (-1.98). Taburan aras kesukaran item bagi ujian ini ditunjukkan pada Rajah 1 berikut:



Rajah 1: Taburan Aras Kesukaran Item

Susunan aras kesukaran item bagi ujian ini bermula dari soalan yang paling sukar kepada yang paling mudah adalah seperti yang ditunjukkan di dalam Jadual 3 yang berikut:

Jadual 3: Susunan Aras Kesukaran Item (Sukar-Mudah)

Item	Nilai logit (Aras Kesukaran)
40	2.85
19	1.91
30	1.91
38	1.91

32	1.57
26	1.28
22	1.02
31	.77
35	.77
5	.54
18	.54
33	.54
36	.54
14	.32
24	.32
34	.32
9	.10
39	.10
8	-.11
12	-.11
27	-.11
29	-.11
37	-.11
6	-.33
11	-.33
23	-.33
28	-.33
25	-.55
13	-.79
17	-.79
20	-.79
7	-1.03
10	-1.03
21	-1.03
1	-1.31
4	-1.31
2	-1.61
3	-1.61
16	-1.61
15	-1.98

Mean: 0. 00

S.D: 1.09

PERBINCANGAN DAN KAJIAN LANJUTAN

Ujian Sirah yang dibangunkan ini mempunyai nilai kebolehpercayaan yang baik tetapi nilai indeks pemisah bagi item pada tahap yang rendah. Indeks pemisah item yang ditetapkan oleh Model Pengukuran Rasch adalah >2.0 (Boone et al., 2014). Indeks pemisah item ialah metrik penting dalam model pengukuran Rasch, mencerminkan keupayaan ujian untuk membezakan antara tahap kesukaran item yang berbeza. Terdapat beberapa cadangan daripada kajian-kajian lepas tentang strategi untuk meningkatkan indeks pemisahan yang boleh diberikan perhatian

seperti membuat penambahbaikan pada item yang dilihat tidak sesuai di dalam ujian tersebut, menggunakan saiz yang sampel yang lebih besar, memvariasikan aras kesukaran item, dan memastikan ujian tersebut bersifat unidimensi (Tunç, 2023; Truong et al., 2021; Aryadoust & Sayama, 2020).

Bagi item yang dikenalpasti sebagai tidak sesuai atau tidak fit dalam set ujian ini iaitu item nombor 19, 37, dan 38 pula, item-item ini perlulah dimurnikan kerana boleh mengganggu pada kualiti set ujian ini. Item yang fit akan mampu meningkatkan kebolehpercayaan dan kesahan bagi satu-satu instrumen (Smith, 1991). Sehubungan dengan itu, item di dalam ujian perlulah dibangunkan dengan baik dan seterusnya dianalisis untuk memastikan item lebih berkualiti. Item yang berkualiti akan dapat menghasilkan keputusan pengukuran yang lebih tepat (Salam, Yousuf, & Bakar, 2020). Kajian ini mempunyai limitasinya dan hanya berfokuskan kepada analisis item. Terdapat beberapa analisis lagi boleh dilanjutkan untuk mengenalpasti kualiti set ujian Sirah secara menyeluruh. Dicadangkan kajian lanjutan seperti analisis dimensionaliti untuk melihat pergerakan semua item pada satu dimensi yang sama. Selain itu, kajian mengenalpasti kebolehfungsian pengganggu juga boleh dilaksanakan pada masa akan datang.

KESIMPULAN

Secara kesimpulannya, ujian Sirah ini mempunyai nilai kebolehpercayaan yang baik dan boleh digunakan untuk mengukur pengetahuan murid tahun empat dalam mata pelajaran Sirah. Namun demikian, terdapat empat item yang perlu dimurnikan iaitu item nombor 19, 23, 37, dan 38. Manakala item-item yang lain boleh disimpan di dalam Bank Soalan Sirah sekiranya perlu kerana mempunyai ciri-ciri psikometrik yang baik. Kajian ini membuktikan juga Model Pengukuran Rasch merupakan satu model yang komprehensif dalam menjalankan analisis item bagi ujian pencapaian.

Rujukan

- A. Z. Fathin & H. Retnawati (2020). How good are the characteristics of the mathematics national standardized school examination items in vocational school? *Journal of Physics: Conference Series*, 1581. <https://doi.org/10.1088/1742-6596/1581/1/012029>
- Abdus Salam, Rabeya Yousuf & Sheikh Muhammad Abu Bakar (2020). Multiple choice questions in medical education: how to construct high quality questions. *International Journal of Human and Health Sciences (IJHHS)*, 4(2), 79-88. <https://doi.org/10.31344/ijhhs.v4i2.180>
- Adilla Desy Rizbudiani, Amat Jaedun, Abdul Rahim & Arief Nurrahman (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, 12(2). <https://doi.org/10.24042/ajpm.v12i2.9939>
- AM Muh Asriadi & Samsul Hadi. (2021). Implementation of item response theory at final exam test in physics learning: rasch model study. *Proceedings of the 6th International Seminar on Science Education (ISSE 2020)*. Atlantis Press <https://doi.org/10.2991/ASSEHR.K.210326.048>
- Arhin, D., Annan-Brew, R., & Owusuaah, R. (2023). Exploring the quality of multiple-choice question type of test items in information and communication technology using item analysis. *E-Journal of Humanities, Arts and Social Sciences*. 4(1), 50-58 <https://doi.org/10.38159/ehass.2023414>

- Aryadoust, V., Ng, L., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1), 6-40. <https://doi.org/10.1177/0265532220927487>
- Ashraf, Z. A., & Jaseem, K. (2020). Classical and modern methods in item analysis of test tools. *International Journal of Research*, 7(5), 397-403. https://www.ijrrjournal.com/IJRR_Vol.7_Issue.5_May2020/IJRR0058.pdf
- Askhabul Kirom & Mochamad Hasyim (2021). Analisis butir soal sebagai standarisasi mutu sekolah dasar pada mata pelajaran PAI dengan menggunakan pendekatan Rasch Model di SD Ma'arif NU Kecamatan Pandaan Pasuruan. *Jurnal Al-Murabbi*, 6(2), 92-98.
- Atikah, Sudiyatno, Abdul Rahim, & Marlina, (2022). Assessing the item of final assessment mathematics test of junior high school using Rasch model. *Jurnal Elemen*. 8(1), 117–130. <https://doi.org/10.29408/jel.v8i1.4482>
- Autaeva, A., Satova, A., Butabayeva, L., Bulabayeva, S., & Sardarova, Z. (2020). Testing method for evaluating academic achievements of students with OOP in inclusive education. *Bulletin of the Abai Kazakh National Pedagogical University. Series "Pedagogical Sciences"*, 3(64), 70-80. <https://doi.org/10.51889/2020-3.1728-7847.13>
- Bambang Sumintono & Wahyu Widhiarso. (2015). *Aplikasi Permodelan Rasch Pada Assessment Pendidikan*. Penerbit Trim Komunika.
- Berman, A. I., Feuer, M. J., & Pellegrino, J. W. (2019). What use is educational assessment? *The ANNALS of the American Academy of Political and Social Science*, 683(1), 20-8. <https://doi.org/10.1177/0002716219843871>
- Bilyakovska, O. O. (2022). Test as an effective means of assessing the quality of students' knowledge. *Academic Notes Series Pedagogical Science*. (204), 16-20. <https://doi.org/10.36550/2415-7988-2022-1-204-16-20>
- Boone, W.J., Staver, J.R., dan Yale, M.S. (2014). *Rasch Analysis in the Human Sciences*. Springer.
- Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch Model. *Measurement*. 146, 961-971 <https://doi.org/10.1016/J.MEASUREMENT.2019.07.035>
- Engelhard, G. & Wang, J. (2020). Developing a concept map for Rasch measurement theory. In: Wiberg, M., Molenaar, D., Gonzalez, J., Bockenholt, U., Kim, JS. (Eds.) *Quantitative Psychology. IMPS 2019. Springer Proceedings in Mathematics & Statistics* (19-29). Cham. https://doi.org/10.1007/978-3-030-43469-4_2
- Estrada, S. (2020). Evaluating the impact of multidimensionality on type i and type ii error rates using the Q-index item fit statistic for the Rasch model. *Journal of Applied Measurement*, 21(4), 496-514. <https://doi.org/10.31219/osf.io/kh7vq>
- Foertsch, M. (2014). Five steps to effective test development. *Educação médica* 13(4), 79-83. <https://doi.org/10.12957/RHUPE.2014.13961>
- Hughes, A. (2002). *Testing for Language Teachers: Stages of test development*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732980.008>
- Ikhsanudin, Novaliah, Hidayatullah, & Memi Almizi (2023). A practical using of the quest program to analyze the characteristics of the test items in educational measurement. *JISAE: Journal of Indonesian Student Assessment and Evaluation*. 9(1), 37-43. <https://doi.org/10.21009/jisae.v9i1.31163>
- Imankasymova, T.K. (2022). Methods of organization of computer testing in general educational institutions. *The Heralds of KSUCTA*, 75(1)128-134. <https://doi.org/10.35803/1694-5298.2022.1.128-134>
- Linacre, J. M (1994). *Sample size and item calibration stability*. https://www.researchgate.net/publication/235361463_Sample_Size_and_ItemCalibration_Stability

- M. Aries Taufiq, Rahmi Eka Putri, Agustina Agustina, Irwan Irwan, M. Zaim, Jasmienti Jasmienti & Syahrul Ramadhan. (2019). Item analysis and teachers' factors in designing a test. *Proceedings of the Proceedings of The 1st EAI Bukittinggi International Conference on Education, BICED 2019*, 17-18 October, 2019, Bukittinggi, West Sumatera, Indonesia. <https://doi.org/10.4108/eai.17-10-2019.2289772>
- Mohamad Najib Abdul Ghafar. (2011). *Pembinaan dan Analisis Ujian Bilik Darjah*. Edisi kedua. Penerbit UTM.
- Nazatin Nuroini, Sulisty Saputro & Suryadi Budi Utomo (2021). Item quality analysis of chemistry final semester test in the academic year of 2017/2018, 2018/2019 and 2019/2020. *JKPK (Jurnal Kimia dan Pendidikan Kimia)*, 6(3), <https://doi.org/10.20961/jkpk.v6i3.54999>
- Nor Aisyah Saat (2020). Summative test items analysis using classical test theory (CTT)/ analisis item kertas peperiksaan sumatif menggunakan teori ujian klasik (TUK). *Sains Humanika*, 12(2-2), <https://doi.org/10.11113/sh.v12n2-2.1788>
- Poehner, M. E. & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, 9(3), 233-265. <https://doi.org/10.1191/1362168805lr166oa>
- Sayit Abdul Karim, Suryo Sudiro & Syarifah Sakinah. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*. 6(2), 256-269. <https://doi.org/10.30659/e.6.2.256-269>
- Siti Eshah Mokshein, Haliza Ishak & Hishamuddin Ahmad. (2019). The use of Rasch measurement model in english testing. *Jurnal Cakrawala Pendidikan*, 3(1), <https://doi.org/10.21831/cp.v38i1.22750>
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51(3), 541-565. <https://doi.org/10.1177/0013164491513003>
- Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation*, 46(3), 591-603. <https://doi.org/10.1080/09638288.2023.2169771>
- Truong, Q., Choo, C., Numbers, K., Merkin, A., Sachdev, P., Feigin, V., Brodaty, H., Kochan, N., & Medvedev, O. (2024). Enhancing precision of the 16-item Informant questionnaire on cognitive decline in the elderly (IQCODE-16) using Rasch methodology. *International Psychogeriatrics*, 36(3), 1-11. <https://doi.org/10.1017/S1041610221002568>
- Tunç, E. B. (2023). A Review of Measurement Tools developed and adapted based on the Rasch model. *İnsan ve Sosyal Bilimler Dergisi*, 6(2), 249-275. <https://doi.org/10.53048/johass.1369336>
- Xie, D., & Cobb, C. (2020). Item Analysis. *The Wiley Encyclopedia of Personality and Individual Differences*. <https://doi.org/10.1002/9781118970843.ch97>
- Xie, J., Peng, N., Cai, Y., Wang, T., & Huang, Q. (2022). Diverse distractor generation for constructing high-quality multiple-choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 280-291. <https://doi.org/10.1109/taslp.2021.3138706>
- Xie, X. (2023). Book review of Rasch models for solving measurement problems: invariant measurement in the social sciences by Engelhard Jr. and Wang. *International Journal of Education and Literacy Studies*, 11(3), <https://doi.org/10.7575/aiac.ijels.v.11n.3p.306>

Yang, C., Luo, L., Vadillo, M., Yu, R., & Shanks, D. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological bulletin*, 147(4), 399–435. <https://doi.org/10.1037/bul0000309>