Jurnal Pendidikan Malaysia 47(1)(2022): 1-14
DOI: http://dx.doi.org/10.17576/JPEN-2022-47.01-01

Analisis Kesahan Kandungan Instrumen Kompetensi Guru untuk Melaksanakan Pentaksiran Bilik Darjah Menggunakan Model Rasch Pelbagai Faset

(Content Validity Analysis of an Instrument to Measure Teacher's Competency for Classroom Assessment Using Many Facet Rasch Model)

ROSYAFINAZ MOHAMAT*, BAMBANG SUMINTONO, & HARRIS SHAH ABD HAMID

ABSTRAK

Kesahan kandungan instrument adalah penting bagi memastikan item-item yang dibina berupaya mengukur perkara yang sepatutnya diukur dan membincangkan sejauh mana item mewakili kandungan yang dimaksudkan. Kajian ini bertujuan menguji kesahan kandungan Instrumen Kompetensi Guru untuk Melaksanakan Pentaksiran Bilik Darjah (IkomGuruPBD) dengan menggunakan analisis Model Rasch Pelbagai Faset (MRPF). IkomGuruPBD terdiri daripada 56 item yang dibina berdasarkan 3 konstruk utama iaitu Pengetahuan PBD, Kemahiran PBD dan Sikap terhadap PBD. Reka bentuk kajian adalah kaedah tinjauan kuantitatif dengan pendekatan penilai berganda (*multirater*) menggunakan soal selidik IkomGuruPBD yang telah diedarkan kepada 12 orang panel pakar. Setiap pakar perlu menentukan tahap persetujuan mereka tentang kesesuaian setiap item dengan menggunakan skala Likert 3 mata terhadap tiga aspek penilaian bagi setiap item. Dapatan menunjukkan antara kelebihan MRPF adalah boleh menentukan tahap ketegasan penilai dan konsistensi penilai, juga mendapati dapatan respon luar jangkaan. Walaupun setiap penilai diberi instrumen, aspek penilaian dan kategori skala yang sama, namun MRPF boleh membandingkan tahap ketegasan secara individu bagi setiap penilai. MRPF juga boleh mengesan kualiti item untuk membantu pengkaji mengenal pasti item yang baik dan item yang lemah untuk ditambah baik. MRPF mempunyai kelebihan untuk memberikan maklumat yang lengkap mengenai ciri psikometrik bagi IkomGuruPBD yang diuji.

Key Words: Kesahan Kandungan; Model Rasch Pelbagai Faset; Kompetensi; Pentaksiran Bilik Darjah; Ketegasan penilai

ABSTRACT

Instrument's content validity is important to ensure that the items constructed are able to measure what they should measure and discuss the extent to which the items represent the intended content. This study examined the content validity of an instrument to measure teacher's competency for classroom assessment (IkomGuruPBD) by using the analysis of Many-Facet Rasch Model (MFRM). IkomGuruPBD consists of 56 items built based on 3 main constructs: knowledge in PBD, skills in PBD and attitude towards PBD. This research is a quantitative method with a multirater approach using IkomGuruPBD questionnaire distributed to 12 expert panels. Each expert determined their level of agreement about the suitability of each item by using a 3-point Likert scale on the three aspects of evaluation. Results show that MFRM can determine the severity and consistency level of the raters, also report findings of unexpected bias response. Although all raters were given the same instrument, the same aspects of evaluation and scale category, MFRM has the ability to compare the severity level for each rater individually. Furthermore, MFRM can also detect the quality of the item and help the researcher to identify good items and weak items to be improved. MFRM has the advantage of providing complete information on the psychometric characteristics of the IkomGuruPBD being tested.

Key Words: Content Validity; Many Facet Rasch Model; Competency; Classroom Assessment; Rater severity

PENGENALAN

Pentaksiran Bilik Darjah (PBD) melibatkan proses pengumpulan maklumat oleh guru untuk menentukan tindakan susulan yang perlu diambil agar dapat meningkatkan perkembangan pembelajaran pelajar (Bahagian Pembangunan Kurikulum 2019). Guru merupakan agen pelaksana yang bertanggungjawab memastikan keberkesanan pentaksiran yang dilaksanakan. Namun begitu, masih wujud kekeliruan dan kurang kesediaan dalam kalangan guru untuk melaksanakan PBD (Sh. Siti Hauzimah 2019). Oleh itu, adalah penting untuk guru memiliki kompetensi agar PBD dapat dilaksanakan dengan berkesan.

Kompetensi guru adalah amat penting untuk ditentukan kerana ianya melibatkan kesan jangka panjang, melibatkan persepsi dan membuat pertimbangan dalam pelaksanaan tugasan (Foschi 2000). Guru akan menjadi lebih efektif dalam melaksanakan tugasan sekiranya memiliki kompetensi yang merangkumi aspek pengetahuan, kemahiran dan sikap (Muhd Khaizer et al. 2019). Oleh itu, Instrumen Kompetensi Guru untuk melaksanakan Pentaksiran Bilik Darjah (IkomGuruPBD) adalah perlu dibina dan telah dibina untuk mengukur tahap kompetensi guru untuk melaksanakan PBD.

Berikutan itu, pembinaan instrumen yang berkualiti perlu memiliki ciri psikometrik yang baik agar instrumen tersebut benar-benar mengukur perkara yang sepatutnya diukur. Kebolehpercayaan sesuatu instrumen adalah ciri psikometrik yang sering dilaporkan, tetapi aspek utama yang perlu diutamakan dalam pembinaan instrumen ialah kesahan muka dan kesahan kandungan (Connell et al. 2018). Kesahan merupakan salah satu syarat penting yang perlu dipenuhi oleh pengkaji dalam pembangunan instrumen kajian (Mohamad et al. 2015). Walaupun instrumen yang dibina mempunyai kebolehpercayaan yang tinggi, namun pengukuran yang dibuat tidak akan tepat sekiranya kandungannya tidak mempunyai kesahan yang memuaskan (Mohammed Afandi et al. 2020; Oluwatayo 2012).

Kesahan kandungan boleh menentukan kesesuaian item terhadap konstruk yang diukur (Mohammad Rahim et al. 2017). Pengujian kesahan kandungan yang melibatkan beberapa orang pakar adalah lebih objektif dan boleh mengelakkan ketidakseimbangan atau *bias*. Bukti empirikal mengenai laporan kesahan kandungan oleh panel pakar boleh meningkatkan keyakinan terhadap kualiti pembinaan instrumen (Mohammed Afandi et al. 2020). Kaedah kuantitatif dalam pengujian kesahan kandungan digunakan untuk memudahkan data diubah dalam bentuk angka/nombor dan diuji secara statistik dan boleh digunakan untuk membuat ramalan dengan disokong oleh bukti secara empirikal (Pozzo et al. 2019).

Tujuan artikel ini adalah untuk menunjukkan analisis penilai berganda (*multirater*) bagi kesahan kandungan IkomGuruPBD dengan lebih tepat dan terperinci menggunakan Model Rasch Pelbagai Faset (MRPF) kerana mempunyai beberapa kelebihan untuk mengatasi limitasi yang terdapat dalam kaedah Teori Ujian Klasik (TUK).

PENGUJIAN ITEM OLEH PAKAR

Kesahan kandungan penting bagi memastikan instrumen yang dibina berupaya mengukur perkara yang sepatutnya diukur dan membincangkan sejauh mana item mewakili kandungan yang dimaksudkan untuk menambah baik item tersebut (American Educational Research Association et al. 2014; Baghaei & Amrahi 2011; Beglar 2010; Edmundson & Koch 1993; Fahmina et al. 2019). Bukti kesahan kandungan perlu dinilai secara berhati-hati serta tertumpu kepada penentuan sama ada instrumen yang dibina adalah cukup untuk digunakan dalam pengukuran (Oluwatayo 2012). Bukti empirikal untuk kesahan kandungan boleh menjelaskan kecukupan kandungan dalam instrumen yang dibina untuk mewakili konstruk yang diukur (American Educational Research Association et al. 2014).

Penilaian pakar sering digunakan untuk para pengkaji mendapatkan bukti penentuan kesahan kandungan yang dibuat. Kesahan pakar merupakan aspek penting dalam pembinaan sesuatu instrumen (Zhu et al. 1998). Penilaian pakar terhadap kesahan kandungan bagi setiap item memainkan peranan yang kritikal untuk membuktikan kesesuaian kandungan dan format yang digunakan untuk menunjukkan seberapa baik kandungan instrumen yang dibina dapat menggambarkan situasi dan perkara yang diukur (Messick 1987). Penilaian pakar terus memainkan peranan dalam amalan pengukuran di pelbagai bidang untuk memastikan kandungan item berupaya mengukur konstruk, jumlah item yang mencukupi, dan kesesuaian skala pengukuran yang digunakan (Finch & French 2019).

Dalam penilaian yang melibatkan beberapa orang panel pakar, setiap pakar menilai setiap item dalam instrumen dan seterusnya membantu pengkaji untuk melaporkan pengubahsuaian yang dibuat berdasarkan penilaian oleh panel pakar (Oluwatayo 2012). Kebolehpercayaan dalaman penilai merupakan salah satu pendekatan untuk mengesan penilai yang *bias* serta merupakan prosedur pengukuran yang baik untuk mengukur konsistensi dan persetujuan antara penilai (Saldaña 2013).

Instrumen yang memiliki kualiti yang baik dapat berfungsi dengan berkesan dan mempunyai keyakinan skor yang lebih tinggi untuk mempengaruhi sebarang

keputusan berdasarkan dapatan yang diperolehi (Asdhiani et al. 2020). Terdapat beberapa cadangan bagi meningkatkan komunikasi mengenai kesahan kandungan oleh pakar untuk pengkaji menentukan persetujuan kesesuaian item, iaitu dengan membezakan antara kesahan kandungan pada peringkat item dan peringkat skala serta menentukan kefahaman pakar terhadap penilaian yang dibuat dan mengelakkan berat sebelah (*bias*) terhadap item yang dinilai (Polit & Beck 2006).

KAEDAH STATISTIK YANG BIASA DIGUNAKAN

Terdapat pelbagai kaedah yang telah digunakan secara meluas untuk menentukan persetujuan antara panel pakar berdasarkan pendekatan Teori Ujian Klasik (TUK). Kaedah Cohen Kappa mula diperkenalkan pada tahun 1960 untuk mengukur persetujuan antara dua penilai yang menggunakan skala kategori nominal (Cohen 1960). Kaedah ini mengukur konsistensi antara dua penilai dengan mengecualikan persetujuan antara kedua-dua penilai (Hsu & Field 2003). Bagi mengatasi kekurangan kaedah Cohen Kappa, kaedah Fleiss Kappa telah diperkenalkan pada tahun 1973 untuk menganalisis persetujuan antara lebih daripada dua penilai (Fleiss & Cohen 1973). Selain itu, kaedah Fleiss Kappa menyediakan tafsiran perbandingan statistik yang lebih mudah difahami berbanding kaedah Cohen Kappa yang mempunyai kesesuaian nilai yang sukar ditafsirkan untuk menentukan persetujuan antara penilai (Allen 2017).

Content Validity Index (CVI) pula merupakan kaedah yang boleh digunakan untuk menentukan kesahan kandungan keseluruhan instrumen yang dikira berdasarkan purata *Content Validity Ratio* (CVR) (Lindell & Brandt 1999). CVI memberi maklumat secara langsung mengenai persetujuan pakar dengan menukarkan data skala ordinal menjadi dua kategori; contohnya, relevan atau tidak relevan (Polit & Beck 2006). Terdapat pengkaji yang cenderung untuk menggunakan kaedah CVR kerana berpendapat kaedah ini adalah lebih praktikal, telus, terarah dan mesra pengguna kerana mudah untuk digunakan (Lawshe 1975; Lindell & Brandt 1999; Mohammed Afandi et al. 2020).

Teori Generalizabiliti (Teori G) telah dibangunkan oleh Cronbach untuk mengukur kebolehpercayaan antara penilai dan mempunyai kelebihan untuk mengasingkan dan menganggarkan pelbagai sumber (Brennan 2010; Webb et al. 2018). Teori G yang merupakan teori statistik lanjutan daripada TUK yang membolehkan pengiraan kebolehpercayaan lebih tepat berkaitan dengan pengukuran tingkah laku serta boleh menganggarkan pelbagai punca ralat untuk mengira kebolehpercayaan dengan lebih tepat (Nor Mashitah 2017).

Dalam kaedah Fuzzy Delphi pula, pakar menentukan skala yang sesuai bagi setiap item berdasarkan indikator bagi setiap konstruk. Seterusnya, kriteria persetujuan yang boleh diterima adalah berdasarkan nilai threshold tertentu, (d) ≤ 0.2, peratus persetujuan pakar bagi setiap item ≥ 70 dan nilai *defuzzication* melebihi 0.5 (Chu & Hwang 2008; Rodgers et al. 2016). Dapatan tersebut boleh dikukuhkan lagi dengan menentukan indeks kesahan kandungan Aiken's V dalam julat 0 hingga 1, iaitu kesahan kandungan adalah baik sekiranya nilai indeks Aiken's yang diperoleh adalah tinggi (Aiken 1985).

Kesemua kaedah analisis yang dinyatakan di atas adalah berdasarkan TUK yang bergantung kepada analisis pada skor yang diberikan penilai. Seterusnya adalah Model Rasch yang merupakan model pengukuran dalam Teori Respon Item (TRI) yang berasaskan konsep kebarangkalian (*probabilistic*) yang digunakan dengan menghasilkan anggaran parameter sebagai bebas antara satu sama lain (Bond & Fox 2015; Linacre 1994). Penilai dianggap sebagai pakar bebas yang menerapkan pemahaman mereka untuk menilai prestasi seseorang dan tidak hanya bergantung kepada rubrik pemarkahan yang ditetapkan.

KELEMAHAN KAEDAH SEDIA ADA

Terdapat pelbagai kelemahan dalam kaedah analisis *multirater* yang menggunakan pendekatan TUK. Kaedah Cohen Kappa boleh digunakan sekiranya jumlah panel pakar adalah 2 orang manakala Fleiss Kappa pula boleh digunakan untuk jumlah panel pakar yang melebihi 2 orang untuk kategori data nominal sahaja (Cohen 1960; Fleiss & Cohen 1973). Indeks persetujuan adalah berada dalam julat -1 hingga +1, iaitu indeks persetujuan positif menunjukkan persetujuan yang baik manakala indeks persetujuan negatif menunjukkan persetujuan yang lemah (Fleiss & Cohen 1973). Namun begitu, kaedah Fleiss Kappa adalah dipersoalkan kerana bergantung kepada anggapan homogeniti dan tidak boleh digunakan dalam skala ordinal dan interval (Allen 2017; Bartok & Burzler 2020; Warrens 2010). Kaedah Fleiss Kappa juga tidak mampu mengesan sekiranya terdapat kemungkinan tekaan (*guessing*) yang dilakukan oleh penilai semasa memberi skor dan tidak boleh mengesan tahap ketegasan penilai (Allen 2017).

Seterusnya, kaedah CVI juga mempunyai beberapa limitasi seperti melibatkan skala ordinal dua kategori sahaja iaitu relevan atau tidak relevan, indeks persetujuan pakar berkemungkinan akan menurun apabila jumlah pakar bertambah, menggunakan pendekatan nilai purata untuk menentukan persetujuan pakar, serta hanya fokus kepada kesesuaian item tetapi tidak melibatkan analisis skala yang digunakan untuk

memastikan pengukuran dibuat konstruk dengan tepat (Polit & Beck 2006).

Kaedah CVR hanya terhad kepada penilaian untuk data berbentuk dikotomus (Lindell & Brandt 1999). Pengukuran konsistensi dalaman berdasarkan TUK mempunyai limitasi kerana tidak berupaya mengesan perbezaan antara penilai secara sistematik contohnya apabila tahap ketegasan penilai adalah konsisten terhadap semua item (Newton 2009). Walaupun Teori G mempunyai beberapa kelebihan berbanding kaedah TUK yang biasa digunakan, kaedah Teori G adalah adalah agak kompleks dan rumit yang menyebabkan pembaca sukar untuk menerima dan memahami tafsirannya (Brennan 2010; Webb et al. 2018). Selain itu, Teori G dan Fuzzy Delphi juga mempunyai beberapa limitasi seperti tidak boleh menentukan tahap ketegasan penilai, ini menyebabkan kesan ralat ketegasan penilai tidak dapat diambil kira dalam penjelasan pengujian skala (Zhu et al. 1998).

Suatu kajian skala besar melibatkan kaedah penilai berganda telah dijalankan oleh Scullen, Mount dan Goff (2000) yang menggunakan kaedah TUK, ternyata memperoleh maklumat sebanyak 62% dapatan hanya mengenai penilai (*rater*) sahaja, manakala hanya 21% dapatan yang diperoleh mengenai individu yang dinilai (*ratee*). Kajian tersebut juga menunjukkan kaedah TUK yang digunakan tidak boleh memberi maklumat yang terperinci mengenai item, susunan tahap kebolehan *ratee* dan konsistensi penilai.

ASAS MODEL RASCH

Beberapa kaedah yang telah dijelaskan oleh pengkaji adalah berdasarkan pendekatan TUK yang hanya membuat analisis secara keseluruhan instrumen, berbanding kaedah Teori Respon Item (TRI) yang mempunyai kelebihan untuk menganalisis secara terperinci sehingga peringkat item dan individu (Bond & Fox 2015; Linacre 2018; Siti Rahayah 2008). Model Rasch adalah berdasarkan pendekatan TRI yang membolehkan pengkaji menyemak semula instrumen (menambah atau membuang item), mengesan *bias* yang mungkin wujud dalam pengukuran, serta boleh memudahkan pengkaji untuk berkomunikasi mengenai dapatan kajian seperti menggunakan Peta Wright yang boleh membuat perbandingan mengenai susunan kebolehan individu dan kesukaran item dengan jelas (Boone 2020).

Prinsip asas model Rasch adalah menukar data mentah atau data ordinal kepada data interval berdasarkan kebarangkalian dan menerapkan kaedah logaritma, yang menjadikannya sebagai satu model log-linear kerana menghasilkan pengukuran menggunakan unit logit (Linacre 2006). Dengan menggunakan model Rasch, kesukaran item dan kebolehan individu diletakkan pada satu lajur skala linear yang sama kerana diandaikan semua item memiliki indeks diskriminasi yang berbeza dan setiap item dihitung hanya parameter kesukaran sahaja (Bond & Fox 2015). Analisis model Rasch boleh menangani beberapa kekurangan TUK kerana ia boleh mengesan data hilang, membuat pengukuran kesahan dan kebolehpercayaan terhadap individu dan penentukuran item, pengukuran individu dan item pada metrik yang sama serta tidak bergantung kepada sampel (Bond & Fox 2015; Bradley et al. 2015).

Analisis data penilai berganda dalam TUK hanya dilakukan secara umum sahaja berbanding MRPF yang memberikan maklumat yang lebih tepat, dapatan yang bernilai, mengenal pasti ketegasan penilai serta kualiti instrumen yang digunakan secara serentak (Bond & Fox 2015; Boone et al. 2014; Engelhard & Wind 2018; Linacre 2021; Maryati et al. 2019). Tahap ketegasan penilai merujuk kepada kecenderungan penilai secara konsisten memberi skor yang lebih tinggi atau yang lebih rendah (Engelhard 1994). Perbezaan tahap ketegasan antara penilai berlaku apabila penilai tidak mempunyai tafsiran skala penarafan yang sama atau mempunyai standard atau jangkaan yang berbeza oleh penilai yang berbeza (Noor Lide 2011).

PENGGUNAAN MODEL RASCH PELBAGAI FASET (MRPF) DALAM KAJIAN

MRPF merupakan lanjutan model pengukuran Rasch dan melibatkan lebih daripada dua aspek yang berinteraksi untuk menghasilkan pemerhatian (Linacre 1994). MRPF berupaya menggabungkan lebih variabel atau faset untuk menentukan perkaitan antara faset tersebut, contohnya analisis yang melibatkan tiga faset iaitu item, penilai (*rater*) dan guru (*ratee*) (Eckes 2015). Dalam perbandingan terhadap penilaian penilai, dapatan MRPF berupaya menerangkan dengan jelas tahap ketegasan penilai dalam penilaian item, konsistensi penilai, membetulkan skor penilai berdasarkan model ideal, analisis skala penarafan dan mengesan interaksi *bias* (Bond & Fox 2015; Eckes 2015; Engelhard & Wind 2018).

MRPF digunakan dalam kajian ini untuk mendapatkan dapatan penilaian yang adil dan tepat berdasarkan penilaian penilai. MRPF mempunyai kelebihan untuk memodelkan penilai berdasarkan takrifan skala yang tersendiri, tanpa perlu selari dengan penilaian oleh penilai yang lain (Bond & Fox 2015; Engelhard & Wind 2018). Oleh itu, setiap penilaian dianggap menyumbang kepada dapatan kebarangkalian terhadap empat komponen yang berinteraksi antara satu sama lain iaitu kebolehan individu yang dinilai, tahap ketegasan penilai, kesukaran item dan skala penilaian (Linacre 1994).

Menurut Linacre (1994), MRPF memiliki kelebihan, seperti: i) boleh memastikan kejayaan sesuatu analisis dengan menjelaskan statistik kesepadanan untuk menyingkirkan responden yang *misfit* yang menyebabkan kualiti dapatan analisis yang dijalankan terjejas; ii) boleh memodelkan tingkah laku penilai iaitu dengan menentukan susunan tahap ketegasan antara penilai; dan iii) mengenal pasti sekiranya terdapat penilai yang *bias* kerana penilai yang berkualiti boleh mengekalkan tahap ketegasan yang sama apabila menilai individu yang berbeza.

Antara beberapa kelebihan lain MRPF ialah model ini: i) boleh mengesan masalah berat sebelah dalam pemberian skor dengan membandingkan skor yang menghampiri nilai purata dengan skor yang ekstrem; ii) pengukuran setiap faset adalah tidak bergantung kepada faset yang lain, contohnya pengukuran tahap kebolehan individu adalah tidak dipengaruhi oleh tahap ketegasan penilai; dan iii) boleh mengenal pasti situasi luar jangkaan yang wujud bagi setiap faset dan boleh dinyatakan satu persatu melalui analisis statistik kesepadanan (Linacre 1994; Sahin et al. 2016). Kajian oleh Cai (2015) menunjukkan penilaian yang berat sebelah (*bias*) boleh menjejaskan proses penilaian dalam ujian yang djalankan.

Analisis MRPF telah mendapat perhatian ramai pengkaji dan telah digunakan secara meluas dalam bidang pengujian, pendidikan dan pengukuran psikologi (Barkaoui 2013; Linacre 1994). MRPF juga banyak digunakan dalam bidang lain seperti kajian dalam bidang nutrisi oleh Sunjaya et al. (2020), kajian untuk menentukan kualiti penilaian penilai dalam *The Canadian English Language Benchmark Assessment for Nurses* (CELBAN) oleh Wan et al. (2021), dan kajian untuk menganalisis kesahan kandungan bagi *Computerized Testlet Instrument to Measure Chemical Literacy Capabilities* oleh Fahmina et al. (2019).

MRPF juga mempunyai kelebihan berbanding TUK kerana MRPF boleh mengenal pasti respon yang tidak tepat oleh pakar, corak penilaian yang tidak sesuai, dan boleh mengesan data hilang (*missing data*) (Fahmina et al. 2019; Goodwin & Leech 2003). Antara kelebihan lain MRPF adalah boleh memberi maklumat yang lebih terperinci mengenai individu yang dinilai (*ratee*), penilai (*rater*) dan kriteria, prosedur analisis adalah lebih mudah dan pantas, boleh mengesan data hilang (*missing data*) serta mengambil kira perbezaan antara ketegasan penilai dengan kesukaran kriteria yang diukur (Eckes 2015).

Ini jelas menunjukkan MRPF merupakan model alternatif yang sesuai untuk digunakan untuk mengatasi limitasi dalam model statistik TUK. MRPF menyumbang kepada pemahaman terhadap analisis konsistensi penilaian pakar serta sokongan secara kuantitatif untuk menjelaskan item yang dikekalkan, disingkirkan atau diubah suai (Nor Mashitah et al. 2015; Zhu et al. 1998). MRPF juga mempunyai kelebihan berbanding TUK kerana MRPF boleh mengenal pasti respon yang tidak tepat oleh pakar, corak penilaian yang tidak sesuai, dan boleh mengesan data hilang (*missing data*) (Fahmina et al. 2019; Goodwin & Leech 2003).

METODOLOGI

DRAF INSTRUMEN UNTUK PENILAIAN PANEL PAKAR

Instrumen Kompetensi Guru untuk melaksanakan PBD (IkomGuruPBD) yang digunakan dalam kajian ini adalah merupakan instrumen yang digunakan untuk pengujian kesahan kandungan oleh panel pakar. IkomGuruPBD terdiri daripada 56 item untuk mengukur 3 konstruk utama iaitu pengetahuan PBD, kemahiran PBD dan sikap terhadap PBD. Jumlah item bagi setiap konstruk pula adalah konstruk pengetahuan PBD adalah sebanyak 22 item, kemahiran PBD sebanyak 24 item dan sikap dalam PBD sebanyak 10 item.

Penentuan subkonstruk dalam IKomGuruPBD adalah berdasarkan analisis terhadap 8 model kompetensi dan 13 instrumen kompetensi sedia ada dan disesuaikan dengan Panduan Pelaksanaan Pentaksiran Bilik Darjah (Edisi Kedua). Lapan model kompetensi tersebut adalah *Standards for Teacher Competence in Educational Assessment of Students* (STCEAS) (American Federation of Teachers et al. 1990), *The Attitude Toward Educational Measurement Inventory* (ATEMI) (Bryant & Barnes 1997), Standard Guru Malaysia (SGM) (Bahagian Pendidikan Guru 2009), Model Literasi Guru (Rohaya Talib & Mohd Najib 2008), Model Kompetensi dalam Pentaksiran (Stiggins 1999), Model Literasi dan Amalan Pentaksiran Guru (Suah et al. 2009), *Educational Assessment Knowledge and Skills for Teachers* (Brookhart 2011), dan Model Konseptual Pendidikan Guru (Institut Pendidikan Guru Malaysia 2011).

Senarai instrumen sedia ada yang dirujuk pula adalah *Teacher Assessment Literacy Questionnaire* (Plake, Impara & Fager 1993), *Attitude Toward Educational Measurement Inventory* (Bryant & Barnes 1997), *Assessment Practices Inventory* (Zhang & Burry-Stock 2003), *The Assessment Literacy Inventory* (ALI) (Mertler & Campbell 2005), Ujian Literasi Pentaksiran (Rohaya Talib & Mohd Najib 2008), *Teacher Report of Student Involvement* (Randel et al. 2011), Inventori Amalan Pentaksiran Guru (Suah et al. 2010), *Teacher Report of Student Involvement, Classroom Assessment Practices Questionnaire* (CAP-Q) (Gonzales & Fuggan 2012), *Assessment Practices Survey* (Lyon et al. 2018), Instrumen Kompetensi Guru Pelatih dalam PBD (Zahari 2018), Instrumen Mengukur Pengetahuan, Kemahiran, Sikap dan Masalah Guru dalam Melaksanakan PBD (Sh. Siti

Hauzimah 2019), *Self-Perceived Assessment Competence & Practices Questionnaire* (Al-Bhalani, 2019), dan Instrumen Mengukur Tahap Pelaksanaan PBD Guru (Tan & Husaina 2020).

KRITERIA PENILAIAN INSTRUMEN

Instrumen kajian telah disemak panel pakar untuk menyemak item bagi memastikan kesesuaian item dengan konstruk, penggunaan ayat, ejaan dan arahan yang jelas bagi instrumen yang dibina. Setiap pakar perlu menentukan tahap persetujuan mereka tentang kesesuaian setiap item dengan menggunakan skala Likert 3 mata terhadap tiga aspek penilaian bagi setiap item.

Aspek penilaian yang pertama adalah kesesuaian dengan kandungan konstruk, iaitu panel pakar perlu memberi respon mengikut skala 1: Kurang Tepat, 2: Tepat, 3: Sangat Tepat. Aspek penilaian yang kedua adalah elemen *bias* ataupun melibatkan isu sensitif, iaitu panel pakar perlu memberi respon mengikut skala 1: Ada, 2: Samar-samar, 3: Tiada. Aspek penilaian yang ketiga adalah kejelasan ayat yang digunakan, iaitu panel pakar perlu memberi respon mengikut skala 1: Tidak Jelas, 2: Jelas, 3: Sangat Jelas.

Pakar juga bebas untuk menyatakan pendapat mereka mengenai kandungan atau istilah yang digunakan dengan menulis ulasan atau cadangan pada ruang yang disediakan di setiap item. Pakar juga boleh memberi komen atau cadangan secara keseluruhan bagi setiap konstruk pada ruangan yang disediakan untuk membantu pengkaji membuat penambahbaikan.

PROFIL PENILAI

Pemilihan panel pakar ini adalah berdasarkan kriteria berpengetahuan luas dalam PBD dan/ atau sedang bertugas atau terlibat secara langsung dalam bidang PBD serta pakar psikometrik. Draf instrumen telah diedarkan kepada 14 orang panel pakar yang terpilih, namun hanya 12 orang pakar sahaja (Rujuk Jadual 1) yang bersetuju untuk menjadi panel pakar semakan kesahan pakar IkomGuruPBD. Kesemua pakar telah memberikan persetujuan untuk terlibat dalam kajian ini. Kebenaran pelaksanaan kajian ini telah diperoleh daripada institusi-institusi di mana panel-panel pakar tersebut bertugas.

JADUAL 1. Maklumat Demografi Panel Pakar (N = 12)

Demografi		Frekuensi	Peratus (%)
Jantina	Lelaki	8	66.67
	Perempuan	4	33.33
Kepakaran	PBD	8	66.67
	Psikometrik	4	33.33
Kelulusan	Sarjana Muda	2	16.67
	Sarjana	2	16.67
	Kedoktoran	8	66.67

MODEL PENGUKURAN

MRPF merupakan model alternatif yang sesuai untuk digunakan untuk mengatasi limitasi dalam model statistik TUK. Data yang dikumpul akan dianalisis dengan menggunakan MRPF bagi menentukan tahap ketegasan penilai (*severity*) dan konsistensi dalaman penilai dalam memberikan skor bagi setiap item. Statistik kesepadanan penilaian penilai boleh ditentukan dengan melihat kepada nilai infit MNSQ dan nilai outfit MNSQ. Statistik kesepadanan penting untuk membantu pengkaji mengetahui sejauh mana ketepatan atau kebolehramalan data berpadanan dengan model Rasch (Siti Rahayah 2008). Nilai statistik kesepadanan (*infit* dan *outfit*) MNSQ menunjukkan konsistensi penilai dalam membuat penilaian. Nilai MNSQ = 1 menunjukkan data adalah ideal mengikut spesifikasi model Rasch. Julat statistik kesepadanan 0.5 hingga 1.5 adalah diterima (Bond & Fox 2015).

Nilai kebolehpercayaan adalah diterima sekiranya lebih besar daripada 0.65 (Bond & Fox 2015). Analisis indeks pengasingan dijalankan untuk mendapatkan andaian/ anggaran pengasingan atau perbezaan responden berdasarkan tahap kebolehan pada pemboleh ubah yang diukur (Wright & Masters 1982). Jika indeks pengasingan yang diperoleh adalah melebihi 2, ia menunjukkan nilai yang baik dan diterima (Linacre 2006). Analisis Rasch memerlukan sekurang-kurangnya pencapaian minimum 40% *raw variance explained by measures* sebagai tanda ukuran instrumen *unidimensionality* yang sangat baik sebanyak 60% (Bond & Fox 2015).

DAPATAN DAN PERBINCANGAN

Hasil analisis menunjukkan jumlah respon yang terlibat dalam analisis panel pakar adalah sebanyak 2016 (12 penilai × 56 item × 3 aspek) yang menunjukkan tiada data hilang (*missing data*). Data yang dikumpul direkodkan ke dalam Microsoft Excel dan kemudian dianalisis dengan menggunakan perisian FACETS versi 3.71.3 yang melibatkan tiga faset iaitu penilai (pakar), kriteria dan item-item IkomGuruPBD.

KEBOLEHPERCAYAAN DAN KESAHAN

Bagi menentukan kebolehpercayaan penilaian panel pakar, pengkaji melihat nilai kebolehpercayaan penilai daripada dapatan analisis MRPF (Jadual 2).

Nilai min bagi faset penilai adalah -3.03, menunjukkan kecenderungan penilai untuk mudah memberi skor (*lenient*) tetapi mempunyai taburan yang luas (SD=1.78) yang menunjukkan para penilai mempunyai tahap ketegasan yang berbeza. Nilai min bagi faset kriteria dan item pula adalah baik iaitu 0.0.

Ketiga-tiga faset menunjukkan nilai ralat standard (*standard error*) yang baik kerana tidak melebihi 0.5 logit. Oleh kerana analisis ini dijalankan untuk penilaian kesahan kandungan oleh panel pakar, maka pengkaji hanya tertumpu kepada dapatan mengenai faset penilai sahaja.

Nilai kebolehpercayaan penilai adalah tinggi iaitu 0.98, indeks pengasingan penilai adalah baik kerana melebihi 3 iaitu 7.26. Nilai signifikan penilai p=0.00 menunjukkan terdapat perbezaan signifikan terhadap ketegasan penilai, iaitu terdapat konsistensi dalaman yang tinggi dalam penilaian oleh penilai. Ini menunjukkan panel pakar mempunyai ketegasan yang berbeza apabila membuat penilaian. Peratus sebenar bagi persetujuan penilai ialah 57.6% manakala peratus jangkaan bagi persetujuan penilai 58.6%. Kedua-dua nilai peratus yang hampir sama menunjukkan penilaian yang dibuat oleh panel pakar adalah tidak homogen dan baik kerana menepati jangkaan oleh Model Rasch.

Seterusnya, ciri unidimensionaliti instrumen dikenal pasti untuk memastikan keupayaan instrumen untuk mengukur dalam satu arah sahaja. Dalam analisis Rasch, peratus *Variance explained by Rasch measures* perlu mencapai sekurang-kurangnya minimum 40% untuk menunjukkan unidimensionaliti yang baik (Engelhard & Wind 2018). Dapatan analisis menunjukkan instrumen yang dibina mempunyai ciri unidimensionalti yang baik kerana peratus Variance *explained by Rasch measures* adalah 56.04%.

Dapatan menunjukkan MRPF berupaya menganalisis kebolehpercayaan dan kesahan instrumen yang dibina secara menyeluruh dan terperinci. Sebagai contoh, kajian oleh Wan et al. (2021) merupakan kajian pertama dalam konteks Kanada yang menggunakan MRPF untuk menentukan kualiti penilaian penilai terhadap 8 kriteria berdasarkan konsistensi penilai, tahap ketegasan penilai, penilaian *bias* dan skala penilaian dalam *The Canadian English Language Benchmark Assessment for Nurses* (CELBAN). Kajian tersebut memberi bukti positif mengenai kebolehpercayaan penilai dan kesahan terhadap CELBAN serta memberi sumber informatif untuk latihan, penilaian penilai dan penyelarasan rubrik.

STATISTIK KESEPADANAN

Analisis Rasch menghasilkan pengukuran untuk mengetahui kualiti alat pengukuran dan memberikan maklumat yang sistematik terhadap analisis statistik kesepadanan (Bradley et al. 2015). Statistik kesepadanan membantu pengkaji untuk mengesan percanggahan antara model Rasch dengan data kajian berdasarkan data *misfit* yang dikenal pasti (Bond & Fox 2015). Bagi memastikan data adalah serasi (*fit*) dengan model pengukuran Rasch, pengkaji meneliti nilai *outfit* MnSq bagi setiap penilai. MnSq (*Mean Square*) merupakan statistik kesesuaian min kuasa dua yang menentukan saiz kerawakan (*randomness*) terhadap sesuatu sistem pengukuran. Penilai yang *misfit* ialah penilai yang memberi skor yang tidak dapat diramal merentasi kriteria (Eckes 2011) atau *noisy rating* (Engelhard 1994).

Nilai MnSq yang sempurna adalah 1.00 logits yang menunjukkan nilai yang dijangka, nilai kurang daripada 0.5 bermaksud bahawa data yang diperoleh adalah mudah diramal (*data overfit model*) manakala nilai lebih daripada 1.5 bermaksud bahawa data yang diperoleh adalah sukar diramal (*data underfit model*) (Azrilah et al. 2013; Bond & Fox 2015). Menurut Bond dan Fox (2015), nilai MnSq adalah diterima sekiranya berada dalam julat 0.5 hingga 1.5.

Dapatan analisis statistik penilaian 12 orang panel pakar adalah seperti yang dinyatakan dalam Jadual 3. Pengukuran logit (*logit measure*) pakar menganggarkan tahap ketegasan penilai, dimana nilai pengukuran logit yang lebih besar menunjukkan penilai yang mempunyai tahap ketegasan yang tinggi manakala nilai pengukuran logit yang lebih kecil menunjukkan penilai yang mempunyai tahap ketegasan yang rendah. Seperti yang ditunjukkan dalam Jadual 3, Penilai 2 merupakan penilai yang mempunyai tahap ketegasan yang paling tinggi dengan nilai logit -0.57 (SE= 0.16) manakala penilai 12 merupakan penilai yang mempunyai tahap ketegasan yang paling rendah dengan nilai logit -6.09 (SE = 0.46).

Dari aspek statistik kesepadanan, 8 daripada 12 orang pakar menunjukkan kesepadanan yang bagus (konsisten) dengan atribut psikometrik yang sesuai dengan julat yang diterima. Namun begitu, terdapat 4 orang penilai yang *nilai* outfit MNSQ yang tidak menepati julat iaitu penilai 6 (*outfit* MNSQ = 0.25), penilai 9 (*outfit* MNSQ = 0.13), penilai 4 (*outfit* MNSQ = 4.18) dan penilai 12 (*outfit* MNSQ = 7.73). Disamping itu juga, 3 daripada 4 orang penilai tersebut mempunyai nilai index *outfit* ZStd yang berada diluar julat yang diterima (Jadual 3). Keempat penilai ini memberikan respon yang berbeza dengan model ideal Rasch yang menunjukkan cara penilaian yang kurang konsisten.

Dapatan ini adalah selaras dengan kajian dalam bidang nutrisi oleh Sunjaya et al. (2020) menggunakan MRPF yang melibatkan penilaian 30 wanita hamil terhadap 3 jenis biskut berdasarkan 4 kriteria penilaian mendapati tahap persetujuan penilai adalah berada pada tahap baik. Kajian tersebut juga menunjukkan MRPF memberikan maklumat terperinci mengenai kriteria biskut yang paling sukar dicapai, jenis biskut yang menjadi pilihan penilai serta susunan tahap ketegasan penilai.

JADUAL 2. Dapatan Analisis MRPF

	Penilai	Kriteria	Item
N	12	3	56
Min logit	-3.03	0.00	0.00
Sisihan Piawai (SD)	1.78	2.29	0.48
Ralat Standard (SE)	0.22	0.14	0.40
Indek Pengasingan	7.26	13.42	0.67
Strata	10.02	18.23	1.23
Kebolehpercayaan	0.98	0.99	0.31
Signifikan (p)	0.00	0.00	0.01
Peratus persetujuan penilai sebenar (%)		57.6	
Peratus persetujuan penilai jangkaan (%)		58.6	
Variance explained by Rasch measures (%)		56.04	

JADUAL 3. Dapatan Analisis Statistik Kesepadanan

Penilai	Model		Infit		Outfit		*Correlation*
	Measure	S. *Error*	MnSq	ZStd	MnSq	ZStd	PtMea
1	-2.32	0.16	0.86	-1.20	0.58	-1.00	0.63
2	-0.57	0.16	0.83	-1.50	0.91	-0.40	0.80
3	-5.59	0.37	1.15	0.50	0.49	0.30	0.21
4	-2.17	0.16	1.24	1.90	*4.18*	*4.90*	0.50
5	-4.17	0.22	1.96	4.90	1.21	0.50	0.27
6	-1.36	0.15	0.34	-7.60	*0.25*	*-4.20*	0.88
7	-5.35	0.33	0.89	-0.20	0.43	0.20	0.24
8	-1.20	0.15	0.78	-1.90	0.68	-1.40	0.78
9	-1.63	0.16	0.17	-9.00	*0.13*	*-4.90*	0.94
10	-3.26	0.18	2.07	7.10	1.29	0.60	0.37
11	-2.68	0.16	1.61	4.60	1.07	0.30	0.48
12	-6.09	0.46	1.40	0.90	*7.73*	1.90	0.09

JADUAL 4. Ringkasan Dapatan Analisis Respon Luar Jangkaan

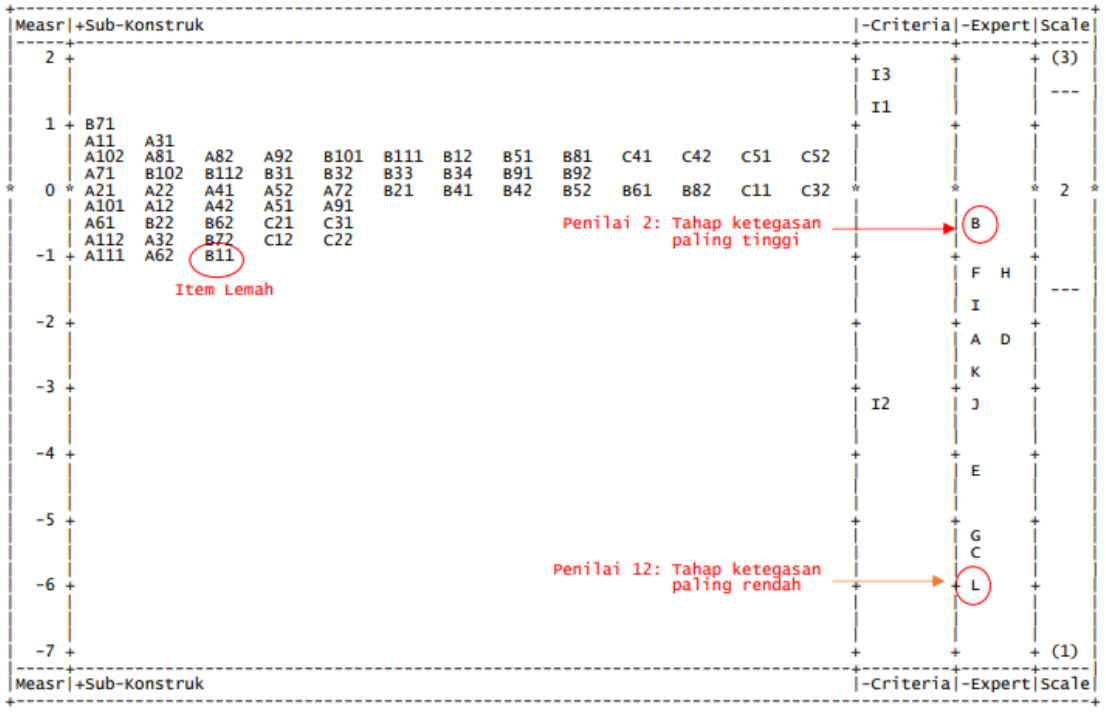
Penilai		Item		Kriteria	
Penilai	Kekerapan	Item	Kekerapan	Kriteria	Kekerapan
2, 3, 4, 7	≥ 5	A12, A32, A111	≥ 3	I1	18
5, 11	≥ 10	A112, B11, B82,		I2	12
10	≥ 20	C22, C31		I3	51

RESPON LUAR JANGKAAN

Antara kelebihan menggunakan MRPF adalah boleh memaparkan respon luar jangkaan yang memberi maklumat mengenai fungsi elemen yang terlibat sebagai contoh sekiranya wujud masalah untuk penilai memahami dan menggunakan kriteria yang diberikan (Eckes 2015; Kudiya et al. 2018). Dapatan mengenai respon luar jangkaan menunjukkan MRPF boleh mengesan konsistensi bagi setiap penilai terhadap item tertentu (Rujuk Lampiran). Terdapat 77 respon yang menunjukkan penilai memberi skor yang lebih rendah berbanding skor jangkaan (*under-value*) dan 4 respon yang menunjukkan penilai memberi skor yang lebih tinggi berbanding skor jangkaan (*over-value*). Jumlah respon luar jangkaan yang dikesan adalah sangat kecil iaitu hanya 4.02% (81 daripada 2016 respon) menunjukkan kesemua pakar telah membuat penilaian secara teliti dan berhati-hati.

Jadual 4 menunjukkan sebahagian respon luar jangkaan yang mempunyai kekerapan yang tinggi bagi faset penilai, item dan kriteria, yang boleh memberi maklumat mengenai konsistensi penilai serta kualiti item yang diuji. Penilai 10 merupakan penilai yang kurang konsisten kerana mempunyai kekerapan respon luar jangkaan yang paling tinggi. Item A12, A32, A111, A112, B11, B82, C22 dan C31 adalah antara item yang menimbulkan kekeliruan kepada penilai semasa membuat penilaian kerana mempunyai kekerapan respon luar jangkaan yang tinggi berbanding item yang lain. Bagi faset kriteria pula, I3 (kejelasan ayat) merupakan kriteria yang paling sukar untuk ditetapkan oleh penilai kerana mempunyai kekerapan yang paling tinggi dalam respon luar jangkaan.

Oleh itu, terbukti MRPF berupaya menghasilkan dapatan yang terperinci seperti mengesan respon luar jangkaan. Kajian oleh Sunjaya et al. (2020) juga telah menunjukkan kelebihan MRPF kerana berupaya mengesan 15 respon luar jangkaan yang boleh menjelaskan konsistensi penilaian yang dibuat oleh penilai.



RAJAH 1. Taburan Kualiti Item

PETA WRIGHT

Nilai logit menunjukkan penilaian yang telah diberikan oleh panel pakar untuk menentukan kesahan kandungan bagi semua item dalam instrumen yang dibina.

Rajah 1 menunjukkan item yang berada di bahagian atas adalah item yang dikatakan baik manakala item yang berada dibahagian bawah adalah dikatakan kurang baik berdasarkan penilaian penilai. Bagi faset penilai pula, Rajah 1 menunjukkan penilai B (penilai 2) sebagai penilai yang mempunyai tahap ketegasan yang paling tinggi manakala penilai L (penilai 12) mempunyai tahap ketegasan yang paling rendah.

Untuk mengkelaskan kualiti item berdasarkan penilaian oleh penilai, pengukuran nilai logit digunakan sebagai dasarnya. Pengkaji menggunakan maklumat yang diperoleh dalam Jadual 1 iaitu nilai min logit = 0.00 dan nilai sisihan piawai = 0.48. Item-item IkomGuruPBD yang mempunyai skor logit 0.96 dikelaskan sebagai item yang sangat baik manakala item yang berada dalam julat skor logit < 0.96 hingga 0.00 dikelaskan sebagai item yang baik. Item yang berada dalam julat skor < 0.0 hingga -0.96 dikelaskan sebagai item yang boleh diterima manakala item yang berada dalam julat skor < -0.96 dikelaskan sebagai item lemah yang perlu diubahsuai atau disingkirkan. Hasil analisis dalam Jadual 5 menunjukkan sebanyak 28 item adalah baik kerana berada dalam julat skor logit < 0.96 hingga 0.00 manakala sebanyak 27 item adalah diterima kerana berada dalam julat skor logit < 0.00 hingga -0.96.

Namun begitu, terdapat 1 item yang lemah dan perlu diubahsuai kerana berada dalam julat skor logit < -0.96 iaitu item B11. Pengkaji telah membuat keputusan untuk tidak membuang item lemah tetapi membuat penambahbaikan seperti yang dicadangkan oleh beberapa panel pakar agar item yang dibina lebih tepat dan jelas. Pernyataan asal bagi item B11 ialah ‘Menjelaskan kaedah pentaksiran yang dipilih’ telah diubah suai menjadi pernyataan baru iaitu ‘Menjajarkan standard pembelajaran dengan standard prestasi untuk merancang PBD’.

Secara keseluruhannya, MRPF boleh digunakan secara meluas untuk meningkatkan kualiti ciri psikometrik bagi instrumen yang dibina. Kelebihan MRPF juga telah terbukti dalam kajian oleh Fahmina et al. (2019) yang menggunakan MRPF untuk menganalisis kesahan kandungan bagi *Computerized Testlet Instrument to Measure Chemical Literacy Capabilities* yang dinilai oleh 9 penilai terhadap 21 item berdasarkan 5 aspek penilaian. Dapatan kajian tersebut menunjukkan MRPF berkebolehan memberi maklumat terperinci mengenai kebolehpercayaan dan indeks pengasingan item, peratus persetujuan penilai, aspek yang paling sukar atau mudah dicapai oleh item, susunan tahap ketegasan penilai, respon luar jangkaan serta susunan kualiti dan tahap kesukaran item.

JADUAL 5. Laporan Pengukuran Item

Kualiti Item	Logit	Item	N
Baik	< 0.96 hingga 0.00	B71 A11 A31 A102 B12 B51 B81 B101 C42 A81 A82 A92 B111 C41 C51 C52 A71 B31 B32 B33 B34 B91 B92 B102 B112 A72 B61 C11	28
Diterima	< 0.00 hingga -0.96	A21 A22 A41 A52 B21 B41 B42 B52 B82 C32 A12 A42 A51 A91 A101 B22 B62 C31 A61 C21 A32 C22 A112 B72 C12 A62 A111	27
Lemah	< -0.96	B11	1

KESIMPULAN

Kajian ini telah menguji kesahan kandungan Instrumen Kompetensi Guru untuk Melaksanakan Pentaksiran Bilik Darjah (IkomGuruPBD) dengan menggunakan analisis Model Rasch Pelbagai Faset (MRPF). Dapatan analisis statistik kesepadanan menunjukkan penilai 6, penilai 9, penilai 4 dan penilai 12 adalah penilai yang kurang baik dan tidak sepadan dengan model pengukuran Rasch. Oleh itu, empat penilai tersebut telah dicadangkan untuk disingkirkan untuk memperoleh dapatan yang lebih berkualiti. Penyingkiran penilai yang *misfit* adalah menyumbang ke arah kawalan kualiti untuk penyaringan dan pengesahan data yang diperolehi. Ini menunjukkan secara keseluruhannya, terdapat lapan orang panel pakar yang membuat penilaian dengan baik. Dapatan analisis penilaian panel pakar terhadap kesahan kandungan instrumen yang dibina menunjukkan hasil yang menarik dan maklumat lengkap mengenai ciri psikometrik. Implikasinya, MRPF boleh menghindari penilaian yang *bias*, berupaya mengenal pasti panel pakar yang tidak berkualiti dan mengesan respon luar jangkaan. Selain itu juga, MRPF juga telah mengesan terdapat 81 respon luar jangkaan dan 1 item yang lemah berdasarkan penilaian 12 orang panel pakar. Peta Wright juga menunjukkan ketiga-tiga faset mempunyai taburan yang baik. Ini menunjukkan MRPF boleh menyediakan pengukuran tepat dan menghasilkan maklumat yang diperlukan secara terperinci. Tidak seperti pendekatan TUK yang memberi penekanan kepada statistik berpusatkan kumpulan, MRPF menghasilkan maklumat yang lebih terperinci mengenai corak kecenderungan penilai, tahap ketegasan penilai dan seterusnya meningkatkan proses kesahan tersebut (Zuliana et al. 2021). Rumusannya, dapatan kajian menunjukkan MRPF sebagai rangka kerja psikometrik yang sesuai berbanding kaedah TUK untuk mengambil kira kesan penilai (*rater effects*) kerana MRPF lebih umum dan menyediakan analisis yang terperinci terhadap penilaian penilai (Eckes 2019).

RUJUKAN

Aiken, L. R. 1985. Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*, 131–141. Retrieved from file:///D:/SKRIPSI/E-SKRIPSI/ejurnal/uji coba produk/validitas/33.pdf

Al-Bhalani, S. M. 2019. *Assessment literacy : A study of EFL teachers ' assessment knowledge, perspectives , and classroom behaviors* (The University of Arizona). Retrieved from https://repository.arizona.edu/handle/10150/633240

Allen, M. 2017. *The SAGE encyclopedia of communication research methods* (Volume 1). Retrieved from United States of America

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. 2014. *Standards for educational and pychological testing*. Washington, D.C.: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. 1990. *Standards for teacher competence in educational assessment of students*. Washington, D.C.

Asdhiani, Y., Saptono, A., & Komarudin. 2020. Profesional supervision model: Development of clinical supervision instruments for teachers of islamic education through a multi-faceted Rasch model. *International Conference on Humanities, Education, and Social Sciences*, 323–333. https://doi.org/10.18502/kss.v4i14.7890

Azrilah Abdul Aziz, Mohd Saidfudin Masodi, & Azami Zaharim. 2013. *Asas model pengukuran Rasch: Pembentukan skala dan struktur pengukuran*. Bangi: Universiti Kebangsaan Malaysia.

Baghaei, P., & Amrahi, N. 2011. Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research*, *2*(5), 1052–1060. https://doi.org/10.4304/jltr.2.5.1052-1060

Bahagian Pembangunan Kurikulum. 2019. *Panduan pelaksanaan pentaksiran bilik darjah edisi Ke-2*. Putrajaya: Kementerian Pendidikan Malaysia.

Bahagian Pendidikan Guru. 2009. Standard guru Malaysia. In *Sektor Pembangunan Profesionalisme Keguruan*.

Barkaoui, K. 2013. Multifaceted Rasch analysis for test evaluation. *The Companion to Language Assessment*, 1–46. https://doi.org/10.1002/9781118411360.wbcla070

Bartok, L., & Burzler, M. A. 2020. How to assess rater

rankings? A theoretical and a simulation approach using the sum of the Pairwise Absolute Row Differences (PARDs). *Journal of Statistical Theory and Practice*, *14*(37). https://doi.org/10.1007/s42519-020-00103-w

Beglar, D. 2010. A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bond, T. G., & Fox, C. M. 2015. *Applying the Rasch Model. In Applying the Rasch model: Fundamental Measurement in the Human Sciences*. https://doi.org/10.1017/CBO9781107415324.004

Boone, W. J. 2020. Rasch basics for the novice. In *Rasch measurement: Applications in quantitative educational research* (pp. 9–30). Singapore: Springer Nature Singapore Pte Ltd.

Boone, W. J., Yale, M. S., & Staver, J. R. 2014. Rasch analysis in the human sciences. In *Rasch Analysis in the Human Sciences*. https://doi.org/10.1007/978-94-007-6857-4

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. M. 2015. Rating scales in survey research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(1), 1–12. https://doi.org/10.29115/sp-2015-0001

Brennan, R. L. 2010. Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brookhart, S. M. 2011. Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Bryant, N. C., & Barnes, L. L. B. 1997. Development and validation of the attitude toward educational measurement inventory. *Educational and Psychological Measurement*, *57*(5), 870–875.

Cai, H. 2015. Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*(3), 262–282. https://doi.org/10.1080/15434303.2015.1053134

Chu, H. C., & Hwang, G. J. 2008). A Delphi-based approach to developing expert systems with the cooperation of multiple experts. *Expert Systems with Applications*, *34*(4), 2826–2840. https://doi.org/10.1016/j.eswa.2007.05.034

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A. D., Ricketts, T., … Brazier, J. 2018. The importance of content and face validity in instrument development: Lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research*, *27*, 1893–1902. https://doi.org/10.1007/s11136-018-1847-y

Eckes, T. 2011. *Introduction to many-facet Rasch model measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang Edition.

Eckes, T. 2015. *Introduction to Many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessment*. Frankfurt: Peter Lang Edition.

Eckes, T. 2019. Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques* (pp. 153–176). https://doi.org/10.4324/9781315187815-2

Edmundson, E. W., & Koch, W. R. 1993. A facet analysis approach to content and construct validity. *Educational and Psychological Measurement*, *53*.

Engelhard, G., & Wind, S. 1994. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., & Wind, S. 2018. *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. https://doi.org/10.4324/9781315766829

Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. 2019. Content validity uses Rasch model on computerized testlet instrument to measure chemical literacy capabilities. *AIP Conference Proceedings*, *2194*(020023). https://doi.org/10.1063/1.5139755

Finch, W. H., & French, B. F. 2019. *Educational and psychological measurement*. https://doi.org/10.4324/9781315650951

Fleiss, J. L., & Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. https://doi.org/10.1177/001316447303300309

Foschi, M. 2000. Double standards for competence: theory and research. *Annual Review of Sociology*, *26*(1), 21–42. https://doi.org/10.1146/annurev.soc.26.1.21

Gilbert, G. E., & Prion, S. 2016. Making sense of methods and measurement: Lawshe's content validity index. *Clinical Simulation in Nursing*, *12*(12), 530–531. https://doi.org/10.1016/j.ecns.2016.08.002

Gonzales, R. D., & Fuggan, C. G. 2012. Exploring the Conceptual and Psychometric Properties of Classroom Assessment. *The International Journal of Educational and Psychological Assessment*, *9*(2), 45–60. Retrieved from http://tijepa.books.officelive.com/Documents/A4_V9.2_TIJEPA.pdf

Goodwin, L. D., & Leech, N. L. 2003. The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, *36*(3), 181–191. https://doi.org/10.1080/07481756.2003.11909741

Hsu, L. M., & Field, R. 2003. Interrater agreement measures: Comments on Kappa n , Cohen's Kappa, Scott's π, and Aickin's α. *Understanding Statistics*, *2*(3), 205–219. https://doi.org/10.1207/s15328031us0203_03

Institut Pendidikan Guru Malaysia, K. P. M. 2011. *Panduan pelaksanaan program bina insan guru PPG*.

Kudiya, K., Sumintono, B., Sabana, S., & Sachari, A. 2018. Batik artisans' judgement of batik wax quality and its criteria: An application of the many-facets Rasch model. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings* (pp. 27–38). https://doi.org/10.1007/978-981-10-8138-5

Lawshe, C. H. 1975. a Quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. 1994. *Many-facet Rasch Measurement*.

Chicago: MESA PRESS.

Linacre, J. M. 2006. *A user's guide to Winsteps/ Ministep Rasch-model computer programs*. Chicago: www.winsteps.com.

Linacre, J. M. 2018. *User's guide to winsteps Rasch-Model computer program*. Chicago: MESA Press.

Linacre, J. M. 2021. A user's guide to Facets Rasch-model computer programs. In *European University Institute*. Retrieved from https://winsteps.com/a/Facets-Manual.pdf

Lindell, M. K., & Brandt, C. J. 1999. Assessing interrater agreement on the job relevance of a test: A comparison of the cvi, t, rwg(j), and r*wg(j) indexes. *Journal of Applied Psychology*, *84*(4), 640–647. https://doi.org/10.1037/0021-9010.84.4.640

Lyon, C. J., Wylie, E. C., Brockway, D., & Mavronikolas, E. 2018. Formative assessment and the role of teachers' content area. *School Science and Mathematics*, *1*(12). https://doi.org/10.1111/ssm.12277

Maryati, Prasetyo, Z. K., Wilujeng, I., & Sumintono, B. 2019. Measuring teachers' pedagogical content knowledge using many-facet Rasch model. *Cakrawala Pendidikan*, *38*(3). https://doi.org/10.21831/cp.v38i3.26598

Mertler, C. A., & Campbell, C. 2005. Measuring teachers' knowledge & application of classroom assessment concepts: Development of the "Assessment Literacy Inventory". In *American Educational Reasearch Association*. Quebec, Canada.

Messick, S. 1987. *Validity*. New Jersey: Educational Testing Service.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. 2015. Measuring the validity and reliability of research instruments. *Procedia - Social and Behavioral Sciences*, *204*, 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

Mohammad Rahim Kamaluddin, Rohany Nasir, Wan Shahrazad Wan Sulaiman, Rozainee Khairudin, & Zainah Ahmad Zamani. 2017. Validity and psychometric properties of malay translated religious orientation scale-revised among Malaysian adult samples. *Akademika*, *87*(2), 133–144. https://doi.org/10.17576/akad-2017-8702-10

Mohammed Afandi Zainal, Mohd Effendi Ewan Mohd Matore, Wan Norshuhadah W Musa, & Noor Hashimah Hashim. 2020. Kesahan kandungan instrumen pengukuran tingkah laku inovatif guru menggunakan kaedah nisbah kesahan kandungan (CVR). *Akademika*, *90*(Isu Khas 3), 43–54.

Muhd Khaizer Omar, Farah Nadia Zahar, & Abdullah Mat Rashid. 2020. Knowledge, skills, and attitudes as predictors in determining teachers' competency in Malaysian TVET institutions. *Universal Journal of Educational Research*, *8*(3C), 95–104. https://doi.org/10.13189/ujer.2020.081612

Newton, P. E. 2009. The reliability of results from national curriculum testing in England. *Educational Research*, *51*(2), 181–212. https://doi.org/10.1080/00131880902891404

Noor Lide Abu Kassim. 2011. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, *11*(3), 179–197.

Nor Mashitah, Mariani, Jain Chee, Mohamad Ilmee, Hafiza, & Rosmah. 2015. Penggunaan model pengukuran Rasch many-facet (MFRM) dalam penilaian perkembanagn kanak-kanak berasaskan prestasi. *Jurnal Pendidikan Awal Kanak-Kanak*, *4*, 1–21.

Nor Mashitah Mohd Radzi. 2017. *Pembinaan dan pengesahan instrumen pentaksiran prestasi standard awal pembelajaran dan perkembangan awal kanak-kanak*. Universiti Malaya.

Oluwatayo, J. A. 2012. Validity and reliability issues in educational research. *Journal of Educational and Social Research*, *2*(May), 391–400. https://doi.org/10.5901/jesr.2012.v2n2.391

Plake, B. S., Impara, J. C., & Fager, J. J. 1993. Assessment Competencies of Teachers: A National Survey. *Educational Measurement: Issues and Practice*, *12*(4), 10–12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Polit, D. F., & Beck, C. T. 2006. The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Reseacrh in Nyrsing & Health*, *29*, 489–497. https://doi.org/10.1038/s41590-018-0072-8

Pozzo, M. I., Borgobello, A., & Pierella, M. P. 2019. Using questionnaires in research on universities: analysis of experiences from a situated perspective. *Revista d'Innovació i Recerca En Educació*, *12*(2), 1–16. https://doi.org/10.1344/reire2019.12.227010

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. 2011. Classroom assessment for student learning: Impact on elementary school mathematics in the central region. Final report. In *National Center for Education Evaluation and Regional Assistance*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED517969&site=ehost-live

Rodgers, M., Thomas, S., Harden, M., Parker, G., Street, A., & Eastwood, A. 2016. Developing a methodological framework for organisational case studies: a rapid review and consensus development process. *Journal of Epidemiology and Community Health*, *4*(1), 1–142. https://doi.org/10.3310/hsdr04010

Rohaya Talib, & Mohd Najib Abd Ghafar. 2008. Pembinaan dan pengesahan instrumen bagi mengukur tahap literasi pentaksiran guru sekolah menengah di Malaysia. *Seminar Penyelidikan Pendidikan Pasca Ijazah 2008, 25-27 November 2008, Universiti Teknologi Malaysia*. https://doi.org/10.4103/nah.NAH_106_16

Sahin, M. G., Teker, G. T., & Güler, N. 2016. An analysis of peer assessment through many facet Rasch model. *Journal of Education and Practice*, *7*(32), 172–181.

Saldaña, J. 2013. *The coding manual for qualitative researchers*. Retrieved from www.sagepublications.com

Scullen, S. E., Mount, M. K., & Goff, M. 2000. Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Sh. Siti Hauzimah Wan Omar. 2019. *Pengetahuan, kemahiran, sikap dan masalah guru dalam melaksanakan pentaksiran bilik darjah bahasa melayu di sekolah rendah*. *9*(1), 56–67.

Siti Rahayah Ariffin. 2008. *Inovasi dalam pengukuran dan penilaian*. Bangi: Fakulti Pendidikan, Universiti Kebangsaan Malaysia.

Stiggins, R. J. 1999. Evaluating classroom assessment training i n Teacher Education. *Educational*

Measurement: Issues and Practice, *18*.

Suah See Ling, Ong Saw Lan, & Shuki Osman. 2009. Pentaksiran pembelajaran pelajar: amalan guru-guru di Malaysia. *Majlis Dekan Pendidikan Malaysia*, *5(6)*, 1–22. https://doi.org/10.1017/CBO9781107415324.004

Suah See Ling, Ong Saw Lan, & Shuki Osman. 2010. Pentaksiran pembelajaran pelajar: Amalan guru-guru di Malaysia. *Malaysian Education Deans' Council Journal*, *5*, 68–83.

Sunjaya, D. K., Herawati, D., Puteri, D. P., & Sumintono, B. 2020. Development and sensory test of eel cookies for pregnant women with chronic energy deficiency using many facet Rasch model: a preliminary study. *Progress in Nutrition*, *22*(3), 1–11. https://doi.org/10.23751/pn.v22i3.10040

Tan Jia Yuh, & Husaina Banu Kenayathulla. 2020. Pentaksiran bilik darjah dan prestasi murid sekolah jenis kebangsaan cina di Hulu Langat, Selangor. *Jurnal Kepimpinan Pendidikan*, *7*(3), 53–64.

Wang, P., Coetzee, K., Strachan, A., Monteiro, S., & Cheng, L. 2021. Examining rater performance on the CELBAN speaking : A many-facets Rasch measurement analysis. *Canadian Journal of Applied Linguistics*, *23*(2), 73–95.

Warrens, M. J. 2010. A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification*, *27*(3), 322–332. https://doi.org/10.1007/s00357-010-9060-x

Webb, N. M., Shavelson, R. J., & Steedle, J. T. 2018. Generalizability theory in assessment contexts. In *Handbook on measurement, assessment, and evaluation in higher education* (pp. 284–305). https://doi.org/10.4324/9780203142189

Wright, B. D., & Masters, G. N. 1982. *Rating scale analysis: Rasch measurement*. https://doi.org/10.2307/2288670

Zahari Suppian. 2018. *Pembinaan instrumen kompetensi guru pelatih dalam pentaksiran bilik darjah* (Universiti Kebangsaan Malaysia). https://doi.org/10.1017/CBO9781107415324.004

Zhang, Z., & Burry-Stock, J. A. 2003. Classroom Assessment Practices and Teachers' Self-Perceived Assessment Skills. *Applied Measurement in Education*, *16*(4), 323–342. https://doi.org/10.1207/S15324818AME1604_4

Zhu, W., Ennis, C. D., & Chen, A. 1998. Many-faceted Rasch modeling expert judgment in test development. *Measurement in Physical Education and Exercise Science*, *2*(1), 21–39.

Zuliana Mohd Zabidi, Sumintono, B., & Zuraidah Abdullah. 2021. Enhancing analytic rigor in qualitative analysis : Developing and testing code scheme using many facet Rasch model. *Quality & Quantity*, *55*(2). https://doi.org/10.1007/s11135-021-01152-4

LAMPIRAN

Hasil Analisis Respon Luar Jangkaan

Cat	Score	Exp.	Resd	StRes	Nu	E	Nu	Sub-	N	Cr
1	1	3.0	-2.0	-9.0	4	D	3	A21	2	I2
1	1	3.0	-2.0	-9.0	4	D	17	A91	2	I2
2	2	3.0	-1.0	-9.0	12	L	21	A111	2	I2
2	2	3.0	-1.0	-7.0	4	D	8	A42	2	I2
2	2	3.0	-1.0	-7.0	4	D	19	A101	2	I2
2	2	3.0	-1.0	-6.0	4	D	11	A61	2	I2
2	2	3.0	-1.0	-5.8	12	L	39	B81	3	I3
1	1	2.9	-1.9	-5.7	12	L	21	A111	3	I3
1	1	2.9	-1.9	-5.6	5	E	37	B71	3	I3
2	2	3.0	-1.0	-4.9	4	D	12	A62	2	I2
1	1	2.8	-1.8	-4.8	3	C	48	C12	3	I3
2	2	3.0	-1.0	-4.6	8	H	25	B21	2	I2
2	2	3.0	-1.0	-4.5	3	C	39	B81	3	I3
1	1	2.8	-1.8	-4.4	5	E	45	B111	3	I3
2	2	2.9	-.9	-4.3	2	B	45	B111	2	I2
1	1	2.8	-1.8	-4.1	5	E	30	B34	3	I3
1	1	2.8	-1.8	-4.1	5	E	41	B91	3	I3
1	1	2.8	-1.8	-4.1	5	E	42	B92	3	I3
1	1	2.8	-1.8	-4.1	5	E	46	B112	3	I3
2	2	2.9	-.9	-4.0	2	B	46	B112	2	I2
2	2	2.9	-.9	-3.9	12	L	2	A12	3	I3
1	1	2.7	-1.7	-3.6	5	E	7	A41	3	I3
1	1	2.7	-1.7	-3.5	10	J	1	A11	3	I3
2	2	2.9	-.9	-3.4	2	B	52	C32	2	I2
2	2	2.9	-.9	-3.4	7	G	13	A71	3	I3
2	2	2.9	-.9	-3.2	3	C	4	A22	3	I3
2	2	2.9	-.9	-3.1	7	G	14	A72	3	I3
2	2	2.9	-.9	-3.1	7	G	35	B61	3	I3
1	1	2.6	-1.6	-3.1	10	J	15	A81	3	I3
1	1	2.6	-1.6	-3.1	10	J	16	A82	3	I3
1	1	2.6	-1.6	-3.1	10	J	18	A92	3	I3
1	1	2.6	-1.6	-3.1	10	J	40	B82	1	I1
3	3	1.5	1.5	2.9	2	B	22	A112	1	I1
2	2	2.9	-.9	-2.9	2	B	51	C31	2	I2
1	1	2.6	-1.6	-2.9	5	E	50	C22	3	I3
1	1	2.6	-1.6	-2.9	10	J	44	B102	3	I3
1	1	2.5	-1.5	-2.8	10	J	26	B22	1	I1
1	1	2.5	-1.5	-2.8	11	K	27	B31	1	I1
1	1	2.5	-1.5	-2.8	11	K	28	B32	1	I1
1	1	2.5	-1.5	-2.8	11	K	29	B33	1	I1
1	1	2.5	-1.5	-2.7	5	E	38	B72	3	I3
1	1	2.5	-1.5	-2.7	5	E	48	C12	3	I3
2	2	2.9	-.9	-2.7	7	G	6	A32	1	I1
1	1	2.5	-1.5	-2.7	10	J	47	C11	3	I3
2	2	2.9	-.9	-2.6	3	C	49	C21	3	I3
1	1	2.5	-1.5	-2.6	5	E	12	A62	3	I3
1	1	2.5	-1.5	-2.6	10	J	4	A22	3	I3
1	1	2.5	-1.5	-2.6	10	J	10	A52	3	I3
1	1	2.5	-1.5	-2.6	10	J	25	B21	3	I3
1	1	2.5	-1.5	-2.6	10	J	32	B42	3	I3
1	1	2.5	-1.5	-2.6	10	J	40	B82	3	I3
1	1	2.4	-1.4	-2.5	5	E	23	B11	3	I3
2	2	2.9	-.9	-2.5	7	G	22	A112	1	I1
1	1	2.4	-1.4	-2.5	11	K	40	B82	1	I1
2	2	2.9	-.9	-2.4	3	C	50	C22	3	I3
1	1	2.4	-1.4	-2.4	10	J	2	A12	3	I3
1	1	2.4	-1.4	-2.4	10	J	22	A112	1	I1
1	1	2.4	-1.4	-2.4	11	K	9	A51	1	I1
1	1	2.4	-1.4	-2.4	11	K	27	B31	3	I3
1	1	2.4	-1.4	-2.4	11	K	28	B32	3	I3
1	1	2.4	-1.4	-2.4	11	K	29	B33	3	I3
1	1	2.4	-1.4	-2.3	10	J	26	B22	3	I3
1	1	2.4	-1.4	-2.3	10	J	36	B62	3	I3
1	1	2.4	-1.4	-2.3	10	J	51	C31	3	I3
1	1	2.3	-1.3	-2.3	11	K	36	B62	1	I1
1	1	2.3	-1.3	-2.3	11	K	51	C31	1	I1
1	1	2.3	-1.3	-2.2	1	A	15	A81	3	I3
2	2	2.8	-.8	-2.2	7	G	23	B11	1	I1
3	3	1.7	1.3	2.2	8	H	50	C22	1	I1
1	1	2.3	-1.3	-2.2	10	J	11	A61	3	I3
1	1	2.3	-1.3	-2.2	10	J	49	C21	3	I3
3	3	1.8	1.2	2.1	1	A	23	B11	3	I3
2	2	2.8	-.8	-2.1	7	G	6	A32	3	I3
3	3	1.7	1.3	2.1	8	H	40	B82	3	I3
1	1	2.3	-1.3	-2.1	10	J	6	A32	3	I3

```
|  1     1     2.3  -1.3 -2.1 | 11 K 40 B82  3 I3  |
|  2     2     2.8   -.8 -2.0 |  3 C 21 A111 3 I3  |
|  1     1     2.2  -1.2 -2.0 |  4 D  2 A12  1 I1  |
|  1     1     2.2  -1.2 -2.0 |  4 D 17 A91  1 I1  |
|  1     1     2.2  -1.2 -2.0 |  4 D 19 A101 1 I1  |
|  1     1     2.2  -1.2 -2.0 | 10 J 22 A112 3 I3  |
|----------------------------+--------------------|
| Cat  Score  Exp.  Resd StRes| Nu E Nu Sub- N Cr  |
+-----------------------------------------------+
```

Rosyafinaz Binti Mohamat
Fakulti Pendidikan
Universiti Malaya
Emel: rosyafinazmohamat@gmail.com

Bambang Sumintono
Fakulti Pendidikan
Universitas Islam Internasional Indonesia
Emel: bambang.sumintono@uiii.ac.id

Harris Shah Abd Hamid
Fakulti Pendidikan
Universiti Malaya
Emel: harris75@um.edu.my

*Pengarang untuk surat-menyurat, emel:
rosyafinazmohamat@gmail.com