IDENTIFIKASI *ITEM FIT* DAN *PERSON FIT* DALAM PENGUKURAN HASIL BELAJAR KIMIA

Rizki Nor Amelia
Universitas Negeri Yogyakarta
Email: rizkinoramelia@gmail.com

ABSTRAK

Tes kimia buatan guru yang berkualitas sangat dibutuhkan mengingat keputusan yang diambil dari hasil tes tersebut berdampak pada siswa. Untuk memenuhi hal tersebut, model Rasch menawarkan statistik uji yang berperan penting dalam konstruksi pengujian yang berkaitan dengan masalah evaluasi dan pemilihan item serta dalam pengambilan keputusan terhadap skor tes yang dihasilkan. Oleh sebab itu, penelitian ini dilakukan dengan tujuan untuk mengidentifikasi *item fit* dan *person fit* dalam pengukuran hasil belajar kimia menggunakan model Rasch. Data yang berupa respon *multiple choice* dari 40 item penyusun instrumen tes kimia buatan guru diambil dengan teknik dokumentasi dan dianalisis menggunakan *Winsteps Rasch Software* versi 3.73. Hasil analisis menyimpulkan bahwa semua item penyusun instrumen tes kimia buatan guru terbukti fit model. Sementara itu, dari 356 siswa SMA di Kota Yogyakarta yang menjadi responden penelitian, 18 siswa diantaranya teridentifikasi sebagai *person misfit* yang sebaiknya diperiksa lebih lanjut untuk mendapatkan bimbingan guru.

Kata kunci: *item fit, person fit, model Rasch*

PENDAHULUAN

Dalam pengukuran hasil belajar kimia di sekolah, guru berperan penting dalam melakukan proses penilaian dan evaluasi kemampuan siswa pada mata pelajaran yang dia ajarkan. Agar hasil pengukuran kimia mencerminkan keadaan yang sesungguhnya, maka proses evaluasi harus dilakukan dengan tepat. Salah satu alat yang dapat digunakan untuk melakukan evaluasi hasil belajar kimia siswa adalah Model Rasch. Melalui model tersebut, guru mendapatkan beberapa manfaat diantaranya membantu dalam membuktikan validitas instrumen yang digunakan dan memberikan hasil pengukuran kemampuan siswa yang lebih akurat (Bond & Fox, 2007; Linacre, 1997; Runnels, 2012). Pada dasarnya, model Rasch sendiri merupakan model pengukuran yang berbasis teori tes modern yang

dibentuk berdasarkan pertimbangan kemampuan responden dalam menjawab kuisioner atau tes terhadap tingkat kesukaran dari butir-butir penyusun instrumen tersebut (Rasch, 1980). Secara matematis, model Rasch memang setara dengan model 1-PL (1-Parameter Logistik) dalam IRT (*Item Response Theory*), namun dikembangkan secara terpisah dan tidak ditentukan sebagai kasus khusus dari model 2-PL (Rasch 1980).

Kehadiran model Rasch sebagai sistem pengukuran yang baru, bertujuan untuk mengatasi keterbatasan pada sistem pengukuran klasik atau *Classical Test Theory* (CTT) (Ashraf & Jaseem, 2020; Meyer & Zhu, 2013; Yilmaz, 2019). Pada pengukuran klasik, baik parameter orang maupun parameter butir yang berupa hasil analisis dari tingkat kesukaran item dan indeks diskriminasi item bersifat *group dependent* (Fan, 1998). Dalam hal tingkat kesukaran item, klasifikasi tingkat kesukaran item akan berubah ketika diberikan pada kelompok sampel yang berbeda (Magno. 2009), sedangkan dalam hal indeks diskriminasi, nilai yang lebih tinggi cenderung diperoleh dari sampel heterogen dan nilai yang lebih rendah diperoleh dari sampel yang homogen (Bichi, 2016). Ketergantungan tersebut berakibat pada CTT yang kurang dapat menggambarkan kemampuan sampel dan membatasi pengembangan tes karena mempersulit analisis (Hambleton & Jones, 1993), serta timbulnya kesulitan teoretis dalam penerapan CTT untuk beberapa situasi pengukuran misalnya saat *equating* maupun *computerized adaptive testing* (Hambleton, Swaminathan, & Rogers, 1991). Dalam teori tes modern, parameter butir tidak berubah meskipun diestimasi dari kelompok sampel yang berbeda (Rezaee, Shafiayan, Jafari, & Zarifsanaiey, 2018). Ini berarti teori tes modern menyediakan skala pengukuran yang seragam (Magno, 2009), sehingga kelompok sampel dapat diuji dengan serangkaian item yang berbeda, sesuai dengan tingkat kemampuan mereka dan skornya dapat dibandingkan secara langsung (Anastasi & Urbina, 2002).

Model Rasch adalah model unidimensional probabilistik yang menyatakan bahwa (1) semakin mudah pertanyaan, maka semakin besar pula kemungkinan siswa akan merespons pertanyaan itu dengan benar, dan (2) semakin besar abilitas yang dimiliki siswa, semakin besar kemungkinan dia akan menjawab pertanyaan

dengan benar dibandingkan dengan siswa yang kurang mampu (Magno, 2009). Oleh sebab itu, dalam model ini hanya dikenal satu parameter butir yang berupa tingkat kesukaran butir, sedangkan parameter indeks diskriminasi diasumsikan sama dengan satu (Maier, 2001). Secara umum, kecocokan model dengan data merupakan perhatian utama saat menerapkan analisis menggunakan pendekatan tes modern (Reise, 1990). Jika data sangat menyimpang dari model Rasch, penyebabnya perlu dipertimbangkan dan orang (*person*) atau item yang tidak sesuai tersebut mungkin saja perlu dihapus (Boone & Noltemeyer, 2017). Oleh sebab itu, khusus model Rasch, terdapat dua jenis kecocokan (fit) yakni *item fit* dan *person fit*, yang menggambarkan validitas pengukuran model Rasch (Wright & Stone, 1999) dan dapat digunakan untuk mendeteksi perbedaan antara data empiris dengan data model Rasch (Bond & Fox, 2015; Zubairi & Kassim, 2006). *Item fit* menerangkan sejauhmana pola sampel respon terhadap suatu item itu konsisten seperti respon orang lain dalam menanggapi item-item yang lain, sedangkan *person fit* menunjukkan sejauhmana pola kinerja sesorang pada tes tersebut itu konsisten melalui item-item yang juga direspon oleh orang lain (Wright & Stone, 1999; Razak, Khairani, & Thien, 2012).

Urgensi penelitian *item fit* dan *person fit* tidak dapat ditawar lagi mengingat keduanya memiliki hubungan yang kuat secara simetri (Reise, 1990), dimana keduanya berperan penting dalam konstruksi pengujian terutama berkaitan dengan masalah evaluasi dan pemilihan item dan dalam pengambilan keputusan terhadap skor tes yang didasarkan dari hasil respon individu (Rost & von Davier, 1994). Oleh sebab itu, melalui *item fit*, kesalahan yang terjadi selama fase kalibrasi pada pengembangan sebuah instrumen dapat terdeteksi dengan jelas. Misalnya, jika terdapat sebuah butir yang memiliki parameter daya beda yang tidak baik, maka statistik *item fit* akan mengidentifikasi masalah ini (Reise, 1990). Sementara melalui *person fit*, akan dapat ditunjukkan ada tidaknya penyimpangan pola respon (mengarah pada skor yang terlalu tinggi atau terlalu rendah) akibat adanya *cheating, carelles responding, lucky guessing, carelles responding*, dan *random responding* (Karabatsos, 2003; Meijer, 1996). Ini berarti bahwa

responden yang termasuk *person fit* hanya mampu menjawab item dengan benar manakala item tersebut memiliki tingkat kesukaran dibawah kemampuan mereka.

Secara historis, penelitian terkait *item fit* dimulai oleh Andersen pada tahun 1973 dan Yen pada tahun 1981 (Reise, 1990); sementara penelitian terkait *person fit* dipelopori oleh Spearman pada tahun 1910, Thurstone pada tahun 1927, maupun Cronbach pada tahun 1946 (Karabatsos, 2003). Pada penelitian ini, *item fit* dan *person fit* diidentifikasi secara khusus menggunakan instrumen tes kimia buatan guru. Identifikasi ini diperlukan guna menentukan dan memastikan item-item penyusun tes kimia sudah layak sehingga skor yang dihasilkan nantinya benar-benar merefleksikan kemampuan kimia siswa secara utuh. Selain itu, dapat diperoleh pula gambaran terkait instrumen tes kimia yang dibuat guru, sehingga guru dapat melakukan evaluasi agar ke depannya dapat membuat instrumen tes kimia yang lebih baik. Hal ini dikarenakan tes kimia buatan guru yang berkualitas sangat dibutuhkan mengingat keputusan yang diambil dari hasil tes tersebut berdampak pada siswa.

METODE

Pendekatan yang digunakan dalam penelitian adalah pendekatan kuantitatif, dimana data penelitian didapatkan melalui teknik dokumentasi. Instrumen yang digunakan adalah instrumen tes kimia buatan guru yang berbentuk *multiple choice* dan dikerjakan oleh 356 siswa SMA di Kota Yogyakarta. Item-item penyusun instrumen tes kimia yang berjumlah 40 butir berfungsi untuk mengukur kemampuan kognitif siswa yang memuat 72,5% pemahaman, 17,5% penerapan, dan 10% penalaran. Adapun ranah materi kimia yang diuji meliputi Kimia Dasar (struktur atom, sistem periodik unsur, ikatan kimia, tata nama senyawa anorganik dan organik, persamaan reaksi sederhana, dan hukum-hukum dasar kimia), Kimia Analisis (larutan (non)-elektrolit, asam-basa, stoikiometri larutan, larutan penyangga, hidrolisis garam, kelarutan dan hasil kali kelarutan), Kimia Fisik (termokimia, laju reaksi, kesetimbangan kimia, ikatan kimia koloid, dan sifat koligatif larutan, reaksi redoks dan elektrokimia), Kimia Organik (senyawa karbon, minyak bumi, dan makromolekul: polimer, karbohidrat

dan protein, serta cara analisis kuantitatifnya, lemak-minyak), dan kimia anorganik (ikatan kimia, unsur kimia yang terdapat di alam termasuk radioaktif, sifatnya, manfaatnya, kereaktifannya, dan produksinya). Data respon yang diperoleh selanjutnya dianalisis menggunakan *Winsteps Rasch software* versi 3.73 (Linacre, 2009) untuk mengidentifikasi ada tidaknya *item misfit* dan *person misfit*, serta mengetahui karakteristik psikometrik berupa Tingkat Kesulitan (*b*) item dan Koefisien Reliabilitasnya.

HASIL DAN PEMBAHASAN

Dalam bagian ini disajikan hasil penelitian yang dipaparkan secara ringkas sebagaimana dalam Tabel 1 dan Tabel 2. Tabel 1 memaparkan statistik outfit MNSQ (*outlier-sensitive or information-weighted fit Mean Square*) yang digunakan untuk mengindentifikasi apakah suatu item termasuk *fit* atau *misfit* terhadap model Rasch beserta karakteristik parameter item berupa tingkat kesukaran item (*b*) bagi masing-masing item penyusun instrumen tes kimia buatan guru. Berdasarkan tabel tersebut telihat bahwa statistik outfit MNSQ berkisar antara 0,76 sampai 1,54; sedangkan tingkat kesukaran item berkisar antara -2,44 logit sampai 2,7 logit.

Tabel 1. Ringkasan Statistik Outfit MNSQ dan Tingkat Kesukaran Item (*b*)

No	MNSQ	b	No	MNSQ	b
1	1,24	0,05	21	0,97	-0,18
2	1,05	-1,36	22	0,90	0,27
3	0,89	-1,72	23	0,76	-2,44
4	1,09	-0,62	24	1,13	-0,51
5	0,97	0,11	25	1,20	0,28
6	0,95	0,14	26	0,88	1,19
7	1,05	0,01	27	1,14	0,48
8	1,06	-1,27	28	1,25	-1,34
9	1,01	-0,36	29	1,54	-1,43
10	0,76	-1,23	30	1,01	-1,75
11	1,08	1,74	31	1,01	-0,54
12	0,98	1,09	32	0,78	1,84
13	0,92	-0,33	33	0,87	1,07
14	0,85	0,42	34	0,93	-0,81
15	0,81	-2,00	35	1,00	0,77
16	1,04	1,13	36	0,99	0,66
17	1,02	0,38	37	1,14	1,22
18	1,43	2,72	38	0,9	0,35
19	0,82	0,45	39	1,08	0,45
20	1,02	0,87	40	1,31	0,21

Tabel 2 memamparkan statistik outfit MNSQ dan abilitas (θ) khusus bagi responden yang termasuk dalam *person misfit*. Berdasarkan tabel tersebut telihat bahwa statistik outfit MNSQ berkisar antara 1,53 sampai maksimum (tak hingga); sedangkan abilitas siswa berkisar antara -1,53 logit sampai 5,47 logit.

Tabel 2. Ringkasan Statistik Outfit MNSQ dan Ability (θ) untuk Person Misfit

Kode Subjek	MNSQ	θ	Kode Subjek	MNSQ	θ
7	1,62	2,11	96	1,88	1,37
16	2,26	2,11	98	1,53	-0,11
26	1,53	0,14	100	max	5,47
34	1,55	1,53	125	2,82	2,99
36	1,69	2,64	136	2,35	-1,72
60	3,54	2,99	145	1,54	0,01
63	3,70	3,46	232	1,65	0,39
70	1,60	3,46	349	2,06	-1,05
77	1,62	-1,53	353	1,55	-0,24

Menurut Linacre (2002), terdapat dua statistik yang dapat digunakan untuk menilai kecocokan data terhadap model Rasch, yakni infit (*inlier-sensitive or information-weighted fit*) dan outfit (*outlier-sensitive or information-weighted fit*), dimana statistik tersebut umumnya dilaporkan dalam bentuk rerata kuadrat (MNSQ) dan z-terstandardisasi (ZSTD). MNSQ adalah rata-rata dari residu kuadrat untuk suatu item, sebaliknya ZSTD (bentuk standar) adalah transformasi dari nilai rata-rata kuadrat dengan koreksi ukuran sampel (Bond & Fox, 2007). Oleh sebab itu, dalam penelitian ini, untuk mengidentifikasi apakah suatu item maupun responden (*person*) terbukti *fit* atau *misfit* terhadap model Rasch, maka output dari Winsteps Rasch software yang berupa statistik *Outfit Mean Square* (MNSQ) perlu diinterpretasikan. Statistik MNSQ dipilih karena statistik tersebut independen dari ukuran sampel.

Linacre (2002) memberikan *rule of thumb* untuk menilai implikasi kecocokan model terhadap pengukuran, yakni MNSQ > 2,0 berarti merusak sistem pengukuran; 1,5 < MNSQ ≤ 2,0 berarti tidak memiliki makna bagi pengukuran; 0,5 ≤ MNSQ ≤ 1,5 berarti bermanfaat bagi pengukuran; dan MNSQ < 0,5 berarti tidak bermanfaat bagi pengukuran meskipun tidak merusak sistem pengukuran. Dalam Tabel 1 disajikan ringkasan tingkat kesukaran dan statistik outfit MNSQ. Berdasarkan kriteria Linacre (2002) terlihat bahwa semua item penyusun instrumen tes kimia buatan guru terbukti fit dengan model Rasch. Ini berarti tidak ada *item misfit* (item yang memiliki nilai statistik fit yang terlalu tinggi maupun terlalu rendah) dalam instrumen yang dianalisis. Dalam suatu

instrumen pengukuran, *item misfit* haruslah diwaspadai karena item-item tersebut merupakan item yang tidak banyak berkontribusi pada keandalan skor tes (Zubairi & Kassim, 2006). Jika suatu item ditemukan *misfit* terhadap model Rasch maka ada indikasi jika konstruksi itemnya memang cacat atau bermasalah (misalnya karena memiliki daya diskriminasi yang buruk) (Zubairi & Kassim, 2006) atau parameter item tersebut memiliki validitas yang dipertanyakan (item justru mengukur kemampuan lain) (Reise, 1990).

Untuk melengkapi informasi tentang karakteristik item penyusun instrumen tes kimia buatan guru, maka diidentifikasi pula parameter item yang berupa tingkat kesukaran. Dalam pendekatan IRT, tingkat kesukaran (parameter *b*) didefinisikan sebagai suatu titik atau lokasi dimana kurva berbentuk S memiliki kemiringan yang paling curam pada suatu skala kemampuan (Al-khadher & Albursan, 2017; Adedoyin & Mokobi, 2013), yang besarnya berkisar antara logit $-\infty$ sampai logit $+\infty$, meskipun umumnya hanya -2 logit sampai 2 logit sehingga tidak terlalu mudah atau terlalu sulit untuk subjek uji yang dituju (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; DeMars, 2010). Oleh sebab itu, pada penelitian ini, item dikatakan memiliki tingkat kesukaran yang rendah (item mudah) apabila $b < 2{,}0$ logit; tingkat kesukaran sedang (item sedang) apabila $-2{,}0 \text{ logit} \leq b \leq 2{,}0$ logit; dan tingkat kesukaran tinggi (item sukar) apabila $b > 2{,}0$ logit. Berdasarkan acuan tersebut, hanya item nomor 18 yang tergolong item sukar (2,72 logit) dan hanya item nomor 23 yang tergolong item mudah (-2,44 logit), sedangkan 38 item sisanya tergolong item sedang ($-2{,}0 \text{ logit} \leq b \leq 1{,}84$ logit). Secara umum, dapat disimpulkan bahwa instrumen tes kimia buatan guru memiliki item-item yang dikatakan baik karena telah memenuhi persyaratan sebagaimana yang ditulis Hambleton & Swaminathan (1985) bahwa item dikatakan "baik" jika memiliki tingkat kesukaran yang baik yaitu $-2 \text{ logit} \leq b \leq +2$ logit.

Dengan menggunakan kriteria yang sama, statistik *person fit* diinterpretasikan. Hasilnya adalah 18 siswa yang termasuk dalam *person misfit*, yakni siswa dengan nomor urut 7, 16, 26, 34, 36, 60, 63, 70, 77, 96, 98, 100, 125, 136, 145, 232, 349, dan 353. Ini berarti, kemampuan (abilitas) 18 siswa yang

diestimasi tersebut memiliki pola respons yang tidak dapat diprediksi oleh model (Smith, 2001). Padahal melalui pola respon, ketepatan respon dari tiap siswa terhadap setiap item soal dapat tergambarkan (Sumintono & Widhiarso, 2015). Salah satu cara mengidentifikasi penyebab *person misfit* adalah melalui matriks Guttman atau *scalograms*. Matriks Guttman mampu memberikan informasi yang berharga karena item soal telah diurutkan dari item termudah (nomor 23) hingga item tersukar (nomor 18). Matriks ini juga bisa menunjukkan unidimensionalitas data (Hambleton & Swaminathan, 1985). Di bawah ini disajikan identifikasi dari lima contoh siswa yang tergolong *person misfit* berdasarkan matriks Guttman:

GUTTMAN SCALOGRAM OF RESPONSES:

```
Person |Item
       |213 2 2 13 32 12    42231113233231123131
       |35039288044149317156025874997650326671 28
       |----------------------------------------
  100 +11111111111111111111111111111111111111  100
   63 +11110111111111111111111111111011111111   63
   70 +11111111111101111111111111111111110111   70
   60 +11111011110111101111111111111111111111   60
   16 +11011111101111111111111101101110111101 11   16
```

Berdasarkan paparan matriks Guttman di atas, dapat disimpulkan bahwa siswa nomor urut 63, 70, 60, dan 16 tergolong *person misfit* dalam model Rasch. Hal ini dikarenakan siswa-siswa tersebut memiliki pola respon yang tidak lazim, yaitu mampu menjawab benar pada butir sukar (butir nomor 18) namun menjawab salah pada butir yang relatif mudah (butir-butir sebelumnya). Jika didasarkan pada definisi Rasch model, yakni siswa dengan kemampuan lebih rendah tidak akan mempunyai peluang menyelesaikan taraf soal yang lebih sukar, maka dapat disimpulkan bahwa jawaban yang diberikan siswa-siswa di atas mungkin saja adalah tebakan yang kebetulan benar (*lucky guessing*) atau bahkan hasil dari

cheating. Hasil identifikasi tersebut selaras dengan Meijer (1996) dan Karabatsos (2003) yang menyebutkan bahwa setidaknya ada lima hal penyebab *person misfit* yakni *cheating* (misalnya menyalin jawaban dari *testee* lain) mengacu pada perilaku yang tidak adil atas jawaban benar pada item soal yang sebenarnya tidak dapat dia jawab dengan benar, *careless responding* terjadi saat *testee* menjawab dengan benar item sukar namun dengan cara yang tidak jelas justru menjawab salah pada item mudah, *lucky guessing* terjadi jika *testee* menebak dengan benar item yang sebenarnya dia tidak tau jawaban mana yang benar, *creative responding* hanya terjadi pada *testee* yang memiliki kemampuan tinggi saat merespon dengan salah pada item yang sebenarnya mudah karena mereka mengartikan item tersebut dengan cara yang unik dan kreatif, terakhir, *random responding* mengacu pada situasi dimana *testee* memilih opsi pilihan ganda secara acak dalam merespon item soal. Sementara itu, siswa dengan nomor urut 100 juga diidentifikasi oleh model Rasch sebagai *person misfit* karena berhasil menjawab 40 item dengan benar (memperoleh ekstrim skor) sehingga statistik fit nya tidak dapat diukur (*over fit*). Menurut Meijer & Sitsma (2001), pengukuran *person fit* tidak hanya mengidentifikasi pola respon yang tidak mungkin, tetapi juga pola respon yang terlalu mungkin. Model Rasch memprediksi ketidakpastian dan terlalu banyak kepastian justru menunjukkan kendala pada respon.

Terakhir, informasi terkait keandalan skor tes (reliabilitas) hasil estimasi menggunakan model Rasch menunjukkan bahwa keandalan termasuk dalam kategori baik. Simpulan tersebut didasarkan pada dua koefisien reliabilitas yang menjadi ciri khusus dari hasil analisis menggunakan model Rasch, yakni *person reliability* dan *item reliability* yang ditafsirkan sama dengan koefisien reliabilitas pada pengukuran klasik (misalnya KR-20 atau *Cronbach Alpha*) (Linacre, 2009), dimana nilai mendekati 1 menunjukkan ukuran yang lebih konsisten secara internal (Boone, Staver, & Yale, 2014). Terdapat dua statistik *person reliability* dan *item reliability* yang dimunculkan oleh hasil analisis, yakni *model person reliability* dan *model item reliability* yang memberikan batas atas keandalan pada *person* dan item, serta *real person reliability* dan *item reliability* yang memberikan batas bawah keandalan pada *person* dan item (Boone, Staver, &

Yale, 2014). Dalam penelitian ini, berturut-turut koefisien *model person reliability* dan *real person reliability* adalah 0,86 dan 0,85; sedangkan koefisien *model item reliability* dan *real item reliability* adalah sama-sama 0,99.

SIMPULAN

Berdasarkan hasil analisis, semua item penyusun instrumen tes kimia buatan guru terbukti fit dengan model Rasch. Ini berarti bahwa semua item dalam instrumen tes bekerja bersama untuk mendefinisikan kemampuan kimia siswa. Hasil estimasi reliabilitas baik ditinjau dari segi *person reliability* maupun *item reliability*, menunjukkan bahwa instrumen memberikan hasil pengukuran yang andal dan analisis tingkat kesukaran item juga mendukung bahwa item-item memiliki karakteristik psikometrik yang baik. Sementara itu, teridentifikasi 18 siswa yang termasuk dalam *person misfit* yang sebaiknya diperiksa lebih lanjut untuk mendapatkan bimbingan guru.

DAFTAR PUSTAKA

Adedoyin, O.O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*. 3(4). 992-1011.

Al-khadher, M.M.A., & Albursan, I.S. (2017). Accuracy of measurement in the classical and the modern test theory: An empirical study on a children intelligence test. *International Journal of Psychological Studies, 9*(1), 71-80.

Anastasi, A. & Urbina, S. (2002). *Psychological testing*. Prentice Hall: New York.

Ashraf, Z.A., & Jaseem, K. (2020). Classical and modern methods in item analysis of test tools. *International Journal of Research and Review, 7*(5), 397-403.

Bichi, A.A. (2016). Classical test theory: An introduction to linear modelling approach to test and item analysis. *International Journal for Social Studies, 2*, 27-33

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.).* Mahwah, NJ: Lawrence Erlbaum Associate.

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

Boone, W.J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners, *Cogent Education*, *4*(1), 1-13. https://dx.doi.org/10.1080/2331186X.2017.1416898

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

DeMars, C. (2010). *Item response theory: understanding statistics measurement*. New York: Oxford University Press, Inc.

Fan, X. (1998). Item response theory and classical test theory:An empirical comparison of their item/person statistics. *Education and Psychological Measurement, 58*, 357-381.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement Issues and Practice*, *12*, 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., & Swaminathan, H. (1985). *Items response theory: Principles and application.* Boston: Kluwer-Nijjhoff Publish.

Hambleton, R.K., Swaminathan, H., & Rogers H.J. (1991). *Fundamental of item response theory*. London: Sage Publication.

Karabatsos, G. (2003). Comparing the abberant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Linacre, J. M. (1997). KR-20/Cronbach alpha or Rasch person reliability: Which tells us the truth? *Rasch Measurement Transactions, 11*, 580-581.

Linacre, J.M. (2002). What do infit and outfit mean-square and standardized mean?. *Rasch Measurement Transaction, 16*, 878.

Linacre, J.M. (2009). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*, 1-11.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307-331. https://dx.doi.org/10.3102/10769986026003307

Meijer, R.R. (1996). Person-fit research: an introduction. *Applied Measurement in Education*, *9*(1). 3-8.

Meijer, R.R., & Sitsma, K. (2001). Person fit statistic: What is their purpose?. *Rasch Measurement Transactions*, *15*(2), 823.

Meyer, J.P., & Zhu, Shi (2013). Fair and equitable measurement of student learning in MOOCs: an introduction to item response theory, scale linking, and score equating. *Journal of Research and Practice in Assessment, 8*, 26-39.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment test*. Chicago, IL: University of Chicago Press.

Razak, N. bin Abd, Khairani, A.Z. bin, & Thien, L.M. (2012). Examining quality of mathemtics test items using rasch model: Preminarily analysis. *Procedia - Social and Behavioral Sciences*, *69*, 2205-2214. https://dx.doi.org/10.1016/j.sbspro.2012.12.187.

Reise, S.P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127-137. https://dx.doi.org/10.1177/014662169001400202.

Rezaee, R., Shafiayan, M., Jafari, P., & Zarifsanaiey, N. (2018). Invariance of item difficulty parameter estimates based on classical test theory and item response theory. *Journal of Advance Pharmacy Education and Research*, *8*, 156-161.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for rasch model. *Applied Psychological Measurement*, *18*(2), 171-182. https://dx.doi.org/10.1177/014662169401800206.

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies, 6*(4), 141-153.

Smith, E.V.Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2(3), 281-311

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi permodelan rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Wright, B., & Stone, M. (1999). *Measurement essentials (2nd ed.)*. Wilmington: Wide Range, Inc.

Yilmaz, H.B. (2019). A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test. *International Electronic Journal of Elementary Education*, *11*(5), 539-545. https://dx.doi.org/10.26822/iejee.2019553350

Zubairi, A.M., & Kassim, N.L.A. (2006). Classical and rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, *2*(1), 1-20.