Direktorat Jenderal Pendidikan Islam
Kementerian Agama Republik Indonesia

MADRASAH REFORM 2020 2024
Realizing Education's Promise
Madrasah Education Quality Reform
(IBRD 8992-ID)

THE WORLD BANK

KEMENTERIAN AGAMA REPUBLIK INDONESIA
DIREKTORAT JENDERAL PENDIDIKAN ISLAM

Sertifikat

diberikan Kepada

suwahono

Sebagai PEMAKALAH

SEMINAR NASIONAL

PENGELOLAAN MADRASAH SWASTA YANG BERMUTU

Jakarta, 9-10 November 2021
A.n Direktur Jenderal Pendidikan Islam
Direktur Guru dan Tenaga Kependidikan Madrasah

KEMENTERIAN AGAMA REPUBLIK INDONESIA
DIREKTORAT JENDERAL PENDIDIKAN ISLAM

Dr. Muhammad Zain, M.Ag.

NIP. 197202062000031001

DIREKTORAT KURIKULUM, SARANA DAN PRASARANA,
KELEMBAGAAN, DAN KESISWAAN MADRASAH
KEMENTERIAN AGAMA REPUBLIK INDONESIA

Prosiding Seminar Nasional

PENGELOLAAN MADRASAH SWASTA YANG BERMUTU

Penjaminan Mutu Madrasah Swasta

Double Tree By Hilton Cikini, Jakarta
9-10 November 2021

Pemodelan *Rasch:* Pengukuran Rater yang Objektif dan Adil dalam Uji Keterampilan Dasar Kimia

Dr. Suwahono, M.Pd
UIN Walisongo Semarang, kota semarang
Email: suwahono@walisongo.ac.id

Abstrak

Tujuan penelitian untuk untuk melihat keobjektifan dan keadilan rater dalam asesmen uji keterampilan dasar kimia. Penelitian ini merupakan penelitian kuantitatif deskriptif. Pemodelan Rasch dalam asesmen digunakan dasar keobjektifan dan dalam mengolah data. Responden dalam penelitian ini adalah 2 kelompok yaitu Rater dalam dan rater luar. Berdasarkan Fakta di lapangan diketahui bahwa tidak semua rater menilai semua peserta uji secara objektif dan adil. Ada perbedaan karakteristik antara rater luar dan dalam. Hasil analisis data berdasarkan pemodelan Rash menunjukkan bahwa rater dalam memiliki nilai ketajaman lebih tinggi dibandikan rater luar,dan Rater luar lebih pelit dibandingkan rater dalam. Pada rater dalam terjadi bias kemurahan. Kesemua rater tidak terjadi efek halo baik rater dalam atau rater luar.

Kata kunci: objektif, adil, rater, bias penilai

I. Pendahuluan

Salah satu upaya yang dilakukan meningkatkan hasil pembelajaran adalah dengan penilaian yang objektif dan adil. Pengujian dan Penilaian uji keterampilan dasar kimia harus melibatkan dua pihak yang berbeda. Pihak yang dilibatkan adalah pihak dari dalam pengelola tempat uji kompetensi dan pihak luar dari *stakeholders* sebagai rater eksternal. Perpaduan kedua kelompok rater diharapkan menghasilkan suatu hasil evaluasi yang bermutu (Dunbar, Koretz, & Hoover, 1999, Linn, Baker, & Dunbar, 1991, Mehrens, 1992). Persoalan yang muncul kemudian adalah adanya perbedaan pandangan antara rater internal dan rater eksternal (Holroyd, 2000, Garraway, 2006). Rater eksternal cenderung fokus pada hasil pembelajaran, analisis tugas, dan pembelajaran tuntas. Berbeda dari pandangan rater eksternal, rater internal fokus pada aktivitas kelas, holistik, dan perkembangan siswa sebagai calon pekerja.

Rater eksternal lebih pasti memberi nilai sebab Rater eksternal memakai acuan analitik dan tuntas, sedangkan rater internal mengambang karena rater internal memakai acuan holistik. Di samping hal tersebut terdapat efek kesalahan dari rater yang berkaitan dengan efek ketajaman, ke-

murahan, halo efek, sentral tendensi dan resensi (Congdon, & MeQueen, 2000, Wilson, & Case, 2000, Weigle, 1998, Lunz, & Stahl, 1993). Hal ini perlu diuji karakteristik rater dalam pengujian uji keterampilan dasar kimia. Pengujian dilakukan untuk memberikan kepastian penilaian terhadap nilai yang telah diberikan oleh rater.

Kepastian Penilaian oleh rater dapat dilihat dari karakteristiknya. Salah satu cara mengungkap karakteristik rater adalah menggunakan Pemodelan Rasch berbasis multi facet (Engelhard, 1994, 1992, Lunz, & Stahl, 1993, Weigle, 1998, Lynch, & McNamara, 1998, Bonk, W. J., & Ockey, G. J. 2003). Hasil penelitian pendahulu adalah akurasi adanya efek-efek kesalahan dari rater yang dapat diungkap. akurasi dapat diperoleh melalui hasil pengukuran penilai

Pengukuran kompetensi hukum dasar kimia terdiri atas beberapa uji, salah satunya adalah keterampilan dasar kimia. Tes keterampilan dasar kimia menggunakan tes performansi berupa uji petik kerja. Uji petik kerja kimia dikembangkan dengan lembar observasi kerja kimia dan lembar perintah kerja sebagai alat ukur. Penyusunan lembar observasi kerja kimia telah dilakukan dengan model analitik yang disusun dan ditetapkan melalui *focused group discussion* antara pakar, rater internal, rater eksternal dan *stakeholders*. Penyekoran uji petik kerja keterampilan dasar kimia tidak terlepas dari pembobotan. Pembobotan suatu butir kerja ditetapkan berdasarkan tingkat kesulitan, tingkat kepentingan dan lama kerja (Cacciatore, & Sevian, 2009, Bernholt, & Parchmann, 2011, Potgieter, Davidowitz, & Venter, 2008). Materi tugas otentik keterampilan dasar kimia, yaitu: (1) aspek pengamatan langsung keterampilan dasar kimia dan (2) aspek pengamatan tidak langsung tugas. Aspek-aspek keterampilan dasar ini menampilkan kegiatan yang berkaitan dengan kegiatan kimia di laboratorium menggunakan panca indra dan alat bantu percobaan.

Titik tolak instrumen berbentuk tes keterampilan dasar menggunakan tugas otentik dengan melibatkan rater pernah dikembangkan oleh McCurry dan Bryce (1997; 2000). Mc Curry dan Bryce membentuk satu panel rater dan memberikan training kepada rater terlebih dahulu dengan harapan terdapat penyamaan persepsi hingga para rater dapat memberikan penilaian konsisten terhadap kemampuan dasar peserta uji. Kelemahan dari penelitian mc Curry dan Bryce adalah tidak adanya uji legalitas atau fokus kajian terhadap karakteristik rater. Hal yang sama berkaitan dengan pengembangan instrumen keterampilan dasar kimia yang melibatkan rater juga dilakukan oleh Sussiana (2012). Sussiana mengukur keterampilan dasar kimia berbasis tes otentik berupa uji perfomans menggunakan rater diri sendiri dan teman sejawat. Penelitian Susana tidak melibatkan rater yang lebih berkompeten dan dilakukan oleh para siswa sendiri yang cenderung kurang tajam, murah dalam menilai adanya efek bias dan halo efek terhadap penilaian.

Penelitian terbaru tentang efek penilaian dipublikasikan oleh Wiyanto dan Supartono (2016) yang mengukur aspek keterampilan dasar kimia untuk kemampuan dasar kimia bidang pangan dengan menggunakan 2 rater. Rater yang digunakan merupakan rater internal (dosen dan asisten laboratorium), efek

rater dalam penelitian Wiyanto dan Supartono tidak diperhitungkan sangat dalam, lebih dimunculkan atau ditonjolkan kemampuan hasil pengukuran rater. Penelitian Susiana, Wiyanto dan Supartono memiliki kesamaan dalam mengukur keterampilan dasar kimia yaitu penggunaaan rater tetapi fokus kepada karakteristik rater tidak diteliti dengan baik. Penelitian ini bertujuan mengungkap karaktersitik rater berdasarkan model Rasch. Selain itu, manfaat praktis hasil penelitian ini adalah dapat digunakan sebagai dasar mengambil keputusan tentang kualitas keterandalan suatu butir tes performansi keterampilan dasar kimia. Secara teoretis penelitian ini memberikan gambaran atau profil keterampilan dasar kimia calon pekerja dan peranan rater internal atau pun eksternal dalam proses uji kompetensi tersebut.

II. METODE

Penelitian menggunakan metode kuantitatif deskriptif. Uji karakteristik rater diukur dengan menggunakan pemodelan Rasch dengan alat bantu aplikasi minifacet. Responden atau Peserta uji adalah siswa-siswa yang sedang diuji keterampilan dasar kimianya. Ujian dilakukan di tempat uji kompetensi keterampilan dasar kimia dengan tempat uji di SMK Texmaco Semarang. Jumlah total peserta uji sebesar 43. Distribusi peserta meliputi jurusan: Teknik Pembuatan Serat Buatan (TPSB) dan Garmen. Rater dalam penilitian ini terdiri atas rater internal dan rater eksternal. Secara keseluruhan rater ber jumlah 31 orang. Rater internal meliputi 12 guru kimia dan 12 guru kompetensi. Rater eksternal terdiri atas 7 orang praktisi berasal dari bagian kimia industri tekstil. Syarat dasar sebagai rater yang dipenuhi yaitu 24 guru telah mengajar lebih dari 5 tahun dengan bidang ilmu yang linear. 7 orang profesional kimia tekstil yang memiliki pengalaman bekerja antara 12 sampai dengan 15 tahun.

Tabel 1. Jumlah peserta dan penilai

<table>
<tr><td colspan="3">Rater uji KKGK</td><td colspan="4">Calon Pekerja</td></tr>
<tr><td colspan="2">Internal</td><td>Eksternal</td><td colspan="4">SMK Texmaco Semarang</td></tr>
<tr><td>Guru Kimia</td><td>Guru Kompetensi</td><td>Praktisi</td><td colspan="2">TPSB</td><td colspan="2">Garmen</td></tr>
<tr><td rowspan="10">12</td><td rowspan="10">12</td><td rowspan="10">7</td><td colspan="4">Berdasarkan Asal Kelas/ Rombel</td></tr>
<tr><td>X</td><td>6</td><td>X</td><td>5</td></tr>
<tr><td>XI</td><td>6</td><td>XI</td><td>5</td></tr>
<tr><td>XII</td><td>11</td><td>XII</td><td>10</td></tr>
<tr><td colspan="4">Berdasarkan Gender</td></tr>
<tr><td>LK</td><td>15</td><td>LK</td><td>18</td></tr>
<tr><td>PR</td><td>8</td><td>PR</td><td>2</td></tr>
<tr><td colspan="2">23</td><td colspan="2">20</td></tr>
</table>

Instrumen yang digunakan dalam penelitian adalah instrumen keterampilan dasar kimia yang sedang dikembangkan oleh Suwahono, Budiyono & Prodjosantosa (2015). Instrumen berupa lembar perintah tugas atau kerja untuk peserta uji, lembar observasi dan lembar bantuan observasi untuk rater. Sebanyak 31 rater telah berpartisipasi aktif dalam pengukuran. Rater mengobservasi tugas yang dikerjakan oleh peserta uji. Setiap sesi ujian terdiri dari 4 rater dan 10 peserta uji. Empat rater dalam sesi ujian terdiri atas 2 rater internal dan 2 Rater eksternal. untuk rater internal yang berasal dari sekolah calon pekerja diberlakukan secara bergilir artinya guru kelas X menilai calon pekerja

kelas XI sedangkan kelas XI menilai kelas XII guru kelas XII menilai calon pekerja kelas X. Ada 2 aspek keterampilan dasar kimia yang digunakan yaitu aspek pengamatan langsung dan pengamatan tidak langsung. Dari 2 aspek tersebut ada 10 tugas yang dikerjakan atau ditampilkan oleh peserta uji yaitu: (1) mengukur menggunakan thermometer, (2) menimbang menggunakan triple beam balance, (3) memanaskan larutan, (4) menyaring suspensi (larutan kasar), (5) membuka dan menutup botol zat padat, (6) menuang larutan dalam wadah, (7) meneteskan lalutan, (8) mencampur larutan, (9) mencium bau hasil reaksi dan (10) mengamati endapan hasil reaksi. Rater memberi skor pada peserta uji dengan cara mencentang apa yang dilakukan oleh peserta uji sesuai dengan kenyataan yang dikerjakan. Pengamatan dilakukan secara terstruktur manakala ada urutan kerja yang tidak sesuai standar maka langkah selanjutnya tidak dinilai. Hasil skor kemudian dikalikan bobot kerja dan dikonsulkan dengan klasifikasi kompeten penilaian.

III. HASIL DAN PEMBAHASAN

1. Hasil Identifikasi Unidimensionalitas Instrumen yang digunakan

Analisisis dimensionalitas butir-butir tugas uji petik kerja keterampilan dasar kimia menggunakan pemodelan Rasch. Pengujian dimensionalitas pemodelan Rasch menggunakan Analisis Komponen Utama (*Principal Component Analysis*, PCA) dari residual. Hasil identifikasi unit dimensional instrumen untuk uji petik kerja di kelompokan berdasarkan cara dan model penskoran yang digunakan. Hasil uji dimensional untuk unidimensi keluaran software Minifacets. Semakin rendah nilai *Variance explained by Rasch measures* keluaran software Facets akan semakin baik, demikian pula sebaiknya. Unidimensi facets memiliki arti suatu model penskoran memiliki unidmensi jika nilai keragaman yang dijelaskan oleh model pengukuran (*Variance explained by Rasch measure*) kurang dari 30%.

Tabel 2. Hasil Identifikasi Unidimensionalitas

No	Standar varitas residu	Nilai-nilai		Kategori
		Eigen	Empirical	
1.	Raw-score variance of observations	0,92	100%	Good
2.	Variance explained by Rasch measures	0,09	10,25%	Good
3.	Variance of residuals	0,83	89,75%	good

Hasil diolah menggunakan software Facets 3.70.3

Hasil identifikasi dimensionalitas dari uji keterampilan dasar kimia yang telah dilakukan berbantuan software Facet 3.70.3 dapat dilihat pada Tabel 2. Berdasarkan Tabel 2 *Variance explained by Rasch measures* memiliki nilai eigen 0,09 dengan nilai empirik 10,25%. Hasil tersebut sesuai dengan model Rasch berada di bawah 30%. Hal ini membuktikan bahwa instrumen uji Keterampilan hukum dasar kimia yang digunakan memenuhi unsur unidimensionalitas.

2. Pengujian karakteristik rater

Pengujian karakteristik rater didasari atas hasil analisis MFRM berbantuan perangkat lunak facet 3.70.3. Karakteristik rater yang dilihat yaitu: (1) ketajaman, (2) kemurahan, (3) ukuran pemusatan dan (4) efek halo.

a. efek Ketajaman rater

Jumlah rater dalam penelitian ini berjumlah 31 orang dengan kualifikasi 24 rater dalam (guru kimia dan) dan 7 orang rater luar (praktisi). Ada 8 rater yang memiliki karakter skala ketajaman atau *severity scale* yang sangat tinggi. Berdasarkan nilai yang diberikan rater (observed score) untuk 43 peserta terdapat nilai (*expected score*) yang lebih rendah dari yang seharusnya hal tersebut dapat dilihat pada Tabel 7.

Selisih nilai yang diberikan bekisar antara 0,1 sampai dengan 10 lebih rendah dari nilai yang diharapkan. Akibat dari efek ketajaman secara statistik mengakibatkan nilai infit dan outfit rater kurang dari 0,5. Hal tersebut menandakan rater tidak cocok model atau rater tersebut terlalu tajam (pelit). Hal-hal tersebut tentu saja sangat merugikan peserta uji karena dinilai terlalu rendah. Pengelompokan rater berdasarkan golongan rater menghasilkan nilai yang dapat dilihat pada Tabel 8.

Tabel 7. Bias interaksi rater dan peserta uji

Observd score	Expctd score	Kode peserta	Obs-Exp average	Bias Size	Infit Mnsq	Outfit Msq
kode rater 31PL TEX						
27	28,03	22 025P	-0,1	-0,12	0,2	0,2
27	28,46	19 022P	-0,15	-0,18	0,2	0,2
25	27,12	23 027P	-0,21	-0,25	0,3	0,3
25	27,16	29 033P	-0,22	-0,25	0,5	0,5
25	27,22	18 021L	-0,22	-0,26	1	1
25	27,32	20 023P	-0,23	-0,27	0,2	0,2
kode rater 27 PL Tex						
25	27,39	3 004L	-0,24	-0,28	0,4	0,4
25	28,06	38 042L	-0,31	-0,36	0,9	0,9
25	28,28	39 043P	-0,33	-0,39	0,4	0,4
25	28,31	25 029P	-0,33	-0,39	0,5	0,5
24	27,48	33 037P	-0,35	-0,4	0,6	0,6
24	27,55	6 007L	-0,35	-0,41	0,5	0,5
Kode rater 15 PD Tex						
23	27,32	41 045P	-0,43	-0,5	1	1
22	27,93	4 005L	-0,59	-0,69	0,1	0,1
22	28,21	2 003L	-0,62	-0,72	0,1	0,1
21	27,42	37 041P	-0,64	-0,74	0,4	0,4
18	27,42	1 001L	-0,94	-1,13	1,1	1,1
18	27,61	21 024L	-0,96	-1,15	0,5	0,5
17	26,52	32 036P	-0,95	-1,17	1,2	1,2

(data diolah dari keluaran software Facet 3.70.3)

Tabel 8. Tingkat Ketajaman rater untuk masing-masing kelompok

No	Rater s	Model		Nilai Infit		Nilai Outfit		Estim. Disc	Corr PtBis	Exact obs%	Agree exp%
		Log sv	S.E	Mnsq	Zstd	Mnsq	Zstd				
1	Internal rater s	0.56	0.06	1.04	0.8	1.03	0.5	0.95	0.37	80.5	51.9
2	Outsider rater s	0.74	0.06	0.95	-1.0	0.93	-1.4	1.07	0.40	81.5	51.9

Keterangan, log sv=logit severe atau mean measure merupakan rata-rata nilai logit pengukuran, SE=Standar eror, estim disk=estimasi diskriminasi. (Data diolah menggunakan software Facets 3.70.3)

Berdasarkan Tabel 8 diketahui bahwa tingkat ketajaman rater dalam dan rater luar sedikit berbeda. Rater dalam memiliki tren tingkat ketajaman lebih rendah. Sedangkan rater luar lebih tinggi atau nilai Tingkat kepelitan (*harshness*) lebih tinggi.

b. Efek kemurahan penilai

Efek kemurahan dilihat dari bias interaksi antara rater dan peserta uji. Efek kemurahan rater disebabkan adanya kecenderungan memberikan nilai tinggi kepada peserta uji yang disukai. Efek kemurahan dapat dilihat pada Tabel 9.

Tabel 9. Bias interaksi rater dan peserta uji

Observd score	Expctd score	Observd count	Obs-Exp average	Bias Size	Model S.E.	Infit Mnsq	Outfit Msq
Rater 1 Pd Tex							
35	27,22	10	0,78	1,22	0,48	1	1
35	28,25	10	0,68	1,1	0,48	0,7	0,7
34	27,48	10	0,65	0,97	0,45	1,3	1,2
34	27,58	10	0,64	0,96	0,45	0,4	0,5
33	27,52	10	0,55	0,78	0,42	0,6	0,7
Rater 7 Pd Tex							
33	27,84	10	0,52	0,74	0,42	0,7	0,6
33	28,18	10	0,48	0,7	0,42	1,7	1,6
32	27,87	10	0,41	0,57	0,4	0,5	0,5
32	28,15	10	0,38	0,53	0,4	0,5	0,5
31	27,39	10	0,36	0,47	0,38	1,2	1,1
Rater 8 Pd tex							
31	27,45	10	0,35	0,47	0,38	1,1	1,1
31	27,77	10	0,32	0,43	0,38	0,7	0,7
31	27,9	10	0,31	0,41	0,38	0,3	0,3
31	28,09	10	0,29	0,39	0,38	1,4	1,4
31	28,52	10	0,25	0,33	0,38	0,3	0,3

(data diolah dari keluaran software Facet 3.70.3)

Hasil pengujian terhadap bias interaksi antara rater dan peserta uji terdapat dilihat dari nilai observed score dan expected score. Efek kemurahan terjadi jika nilai observed score > nilai expected score. selain itu juga dilihat dari ukuran bias rater. Terdapat tiga rater yang memberikan nilai yang lebih tinggi dari seharusnya. Hasil ini berakibat nilai yang diberikan tidak sewajarnya dan tidak menunjukkan abilitas sebenarnya yang dimiliki oleh peserta uji.

c. Efek Ukuran pemusatan

Efek Ukuran pemusatan terjadi jika rater memberikan penilaian mengumpul pada nilai tengah. Rater tidak menggunakan skala penilaian yang ekstrem tinggi dan ekstrem rendah dan lebih cenderung mengelompokkan seluruh penilaian di tengah-tengah skala. Hal ini dapat mengurangi keakuratan dalam penilaian. Konsekuensinya adalah, sebagian besar nilai dari penilaian performansi sistematis hilang. Penilaian tersebut gagal untuk memisahkan keseluruhan atau antar individu, dan penilaian tersebut menjadi tidak berguna sebagai alat dalam pengambilan keputusan. Bias central tendensi berdasarkan kerangka Rasch dilihat dari sebaran nilai serta kualitas dari sebaran nilai tersebut. Sebaran nilai dapat dilihat pada Tabel 10.

Tabel 10. Analisa Statistik Terhadap Central Tendency

Data Category					Quality Control		
Score	Total	Used	%	Cum.%	Avge Meas	Exp. Meas	Outfit Mnsq
1	3705	3705	28	28	-0,50	-0,52	1,0
2	4290	4290	32	60	-0,41	-0,42	1,1
3	4027	4027	30	90	-0,34	-0,29	1,1
4	1308	1308	10	100	-0,04	-0,14	0.9

(data diolah dari keluaran software Facet 3.70.3)

Sebaran nilai untuk penggunaan skor tidak mengumpul pada nilai tengah tetapi merata pada semua skor. Hal tersebut dapat dilihat pada sebaran% skor dimulai dari 28, 32,30 dan 10. Kualitas rata-rata pengukuran bergerak dengan jarak skala yang seimbang yaitu dari-0,50,-0,41,-0,34, dan-0,04 yang tidah jauh berbeda dengan nilai ekspektasinya. Berdasarkan analisis data tersebut dapat disimpulkan bahwa tidak ada pada proses pengukuran efek central tendensi pada rater tidak terbukti.

d. Efek Halo efek

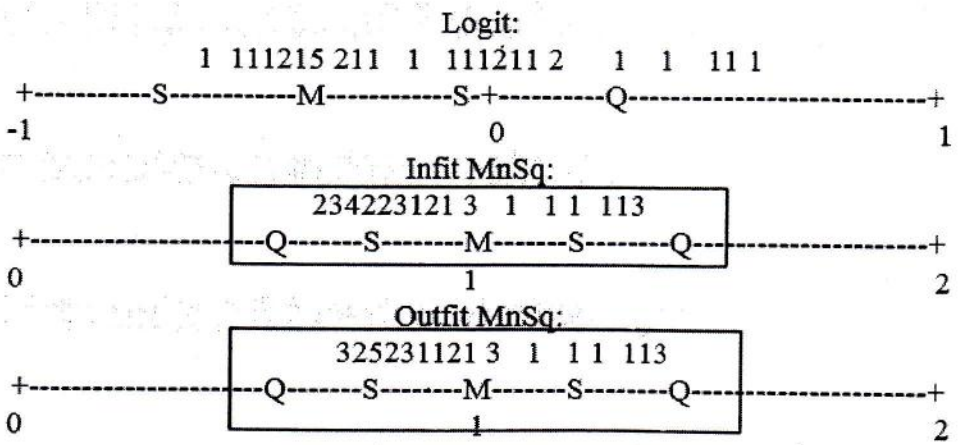
Halo Effect merupakan bias sistematik dalam penilaian terhadap suatu subjek. Bias halo terjadi karena melakukan generalisasi dari satu aspek penilaian sehingga memengaruhi seluruh aspek penilaian. *Halo Effect* biasanya terjadi pada saat pertemuan pertama atau terjadi kejadian istimewa antara peserta uji dan rater. Terjadinya *Halo Effect* dikarenakan cara berpikir individu yang cenderung membuat kategorisasi-kategorisasi mengenai sifat manusia, yaitu kategorisasi sifat-sifat baik dan sifat-sifat buruk. Adanya halo efek di dalam penelitian menggunakan MFRM dapat dilihat melalui 2 hal yakni halo efek secara kelompok dan individual. Efek halo kelompok dilihat dari nilai ratio indek separasi dan Chi Square. Hasil analisis MFRM untuk analisis bias halo kelompok rater dapat dilihat pada Tabel 11.

Tabel 11. Hasil analisis hallo effect

```
+-------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .05  Adj (True) S.D. .37  Separation 6.05  Strata 8.40  Reliability (not inter-rater ) .97
Model, Sample: RMSE .05  Adj (True) S.D. .38  Separation 6.15  Strata 8.53  Reliability (not inter-rater ) .97
Model, Fixed (all same) chi-square:  1143.5  d.f.: 30  significance (probability): .00
Model,  Random (normal) chi-square:  29.2  d.f.: 29  significance (probability): .45
Inter-Rater  agreement opportunities: 199950  Exact agreements: 57821 =  28.9%  Expected:  57139.4 =  28.6%
```

Berdasarkan Tabel 11 diketahui strata indek separasi 8,4 dengan indek separasi ratio 6,05. Hal tersebut berarti bahwa penyebaran tingkat kesulitan 8, 4 kali lebih besar dari tingkat kesulitan pengukuran. Bukti-bukti tersebut menunjukkan bahwa secara kelompok tidak terbukti halo efek. Bias halo efek secara individu terjadi jika ditemukan penyimpangan nilai infit dan outfit model. Adanya nilai outfit dan infit MNSQ yang berada di atas 2,0 rater terhadap peserta uji menunjukkan adanya intrepestasi yang berbeda antara rater terhadap peserta uji dalam kelompok rater. Petunjuk adanya efek halo dalam MFRM untuk rater dapat dilihat dari analisis interaksi antara rater dan peserta uji. Analisis awal adanya bias efek halo diidentifikasi dari titik sebaran dalam continum bias peserta uji dan rater. Gambaran sebaran nilai outfit dan infit rater dapat dilihat pada Tabel 12.

Tabel 12. Analisis interaksi rater dan peserta



Berdasarkan hasil interaksi antara rater dan peserta uji diketahui tidak ada penyimpangan (*abberant*) dari titik tengah hal ini berarti secara personal tidak cukup bukti adanya efek halo dari rater.

IV. KESIMPULAN DAN REKOMENDASI

Secara teknikal butir-butir yang digunakan dalam tes ini memiliki kesesuaian model dengan kerangka Rasch hal ini dibuktikan bahwa butir yang digunakan tidak ada yang melewati batas ambang infit dan outfit. Efek bias (ketidak objektifan dan keadilan) rater hasil penelitian ini menunjukkan bahwa, (1) tingkat ketajaman rater luar lebih tajam dibandingkan peneliti dalam. (2) Terdapat 3 rater dalam yang mengalami Bias kemurahan (3) efek pemusatan penilaian tidak ditemukan baik untuk Rater dalam atau luar hal ini membuktikan rater. (4) Halo efect tidak terjadi pada para rater hal ini menunjukkan bahwa rater tidak memiliki kesan positif atau negatif terhadap peserta uji.

Rekomendasi yang bisa diberikan adalah perlu penyebaran lebih luas penggunaan pemodelan Rasch agar rasa objektif dan adil di dapatkan oleh semua peserta uji.

V. DAFTAR PUSTAKA

Alkan, F. (2013). The effect of alternative assessment techniques on chemistry competency perceptions and chemistry success of prospective science teachers. *Journal of Baltic Science Education*, 12(6).

Ashraf, S. S., Marzouk, S. A., Shehadi, I. A., & Murphy, B. M. (2010). An integrated profesional and transferable skills course for undergraduate chemistry students. *Journal of Chemical Education*, 88(1), 44-48.

Balzer, W. K. (1986). Biases in the recording of performance-related information: The effects of initial impression and centrality of the appraisal task. *Organizational Behavior and Human Decision Processes*, 37(3), 329-347.

Barrett, G. V., & Kernan, M. C. (1987). Performance appraisal and terminations: A review of court decisions since Brito v. Zia with implications for personnel practices. *Personnel Psychology*, 40(3), 489-503.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice*, 12, (2), 167-173.

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, 17(9), 747-751.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of applied psychology*, 60(5), 556.

Breuer, K., P. Nieken, and D. Sliwka. 2013. Social Ties and Subjective Performance Evaluations: An Empirical Investigation. *Review Management Science*, 7, 141-157.

Cacciatore, K. L., & Sevian, H. (2009). Incrementally Approaching an Inquiry Lab Curriculum: Can Changing a Single Laboratory Experiment Improve Student Performance in General Chemistry?. *Journal Chemistry Education*, 86, (4), 498.

Chesterton, L. S., Sim, J., Wright, C. C., & Foster, N. E. (2007). Interrater reliability of algometry in measuring pressure pain thresholds in healthy humans, using multiple rater s. *The Clinical journal of pain*, 23(9), 760-766.

Congdon, P. J., & McQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, 37(2), 163-178.

Devine, D. J., Olafson, K. M., Jarvis, L. L., Bott, J. P., Clayton, L. D., & Wolfe, J. M. (2004). Explaining jury verdicts: Is leniency bias for real?. *Journal of Applied Social Psychology*, 34(10), 2069-2098.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-303

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Pemodelan Rasch to detect *Halo Effect* in three types of rater s. *Theory and Practice in Language Studies*, 1(11), 1531-1540.

Holroyd, C. (2000). Are assessors profesional? Student assessment and the professionalism of academics. Active learning in higher education, 1(1), 28-44.

Leggett, M., Kinnear, A., Boyce, M., & Bennett, I. (2004). Student and staff perceptions of the importance of generic skills in science. *Higher Education Research & Development*, 23(3), 295-312.

Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *physical medicine and rehabilitation*, 70(12), 857.

Lunz, M. E., & Stahl, J. A. (1993). The effect of rater severity on person ability measure: a Pemodelan Rasch analysis. *American Journal of Occupational Therapy*, 47(4), 311-317.

McDonnell, C., O'Connor, C., & Seery, M. K. (2007). Developing practical chemistry skills by means of student-driven problem *Based Learning* mini-projects. Chemistry Education Research and Practice, 8(2), 130-139.

Potgieter, M., Davidowitz, B., & Venter, E. (2008). Assessment of preparedness of first-year chemistry students: development and application of an instrument for diagnostic and placement purposes. *African Journal of Research in Mathematics, Science and Technology Education*, 12, (sup1), 1-17.

Schumacker, R. E. (1998). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of outcome measurement*, 3(4), 323-338.

Stahl, J. A. (1994). What does generalizability theory offer that many-facet Rasch measurement cannot duplicate. *Rasch Measurement Transactions*, 8(1), 342-3.