JKPK (JURNAL KIMIA DAN PENDIDIKAN KIMIA), *Vol 2, No 1, April 2017* *Hal. 1-12*
Program Studi Pendidikan Kimia Universitas Sebelas Maret *ISSN 2503-4146*
https://jurnal.uns.ac.id/jkpk *ISSN 2503-4154 (online)*

IMPLEMENTASI *ITEM RESPONSE THEORY* SEBAGAI BASIS ANALISIS KUALITAS BUTIR SOAL DAN KEMAMPUAN KIMIA SISWA KOTA YOGYAKARTA

Implementation of Item Response Theory for Analysis of Test Items Quality and Students' Ability in Chemistry

Rizki Nor Amelia* dan Kriswantoro

Penelitian dan Evaluasi Pendidikan, Program Pascasarjana, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

* Untuk Korespondensi, Telp: 085743144516, e-mail: rizkinoramelia@gmail.com

Received: March 29, 2017 Accepted: April 26, 2017 Online Published: April 30, 2017

DOI : 10.20961/jkpk.v2i1.8512

ABSTRAK

Penelitian ini bertujuan untuk mendeskripsikan kualitas butir soal hasil pengembangan alat ukur (soal mid semester 1 mata pelajaran kimia bagi kelas XI-IPA) yang dibuat oleh guru dan mengetahui karakteristik hasil pengukuran kemampuan kimia siswa SMA. Desain penelitian yang digunakan adalah penelitian deskriptif dengan subjek penelitian sebanyak 101 pola respon siswa terhadap perangkat tes berupa soal pilihan ganda lima alternatif jawaban pada mid semester I mata pelajaran kimia kelas XI IPA Tahun Ajaran 2015/2016 yang dikumpulkan melalui teknik dokumentasi. Objek dalam penelitian ini adalah kualitas alat ukur dan prestasi belajar siswa yang dilihat dari estimasi kemampuannya. Pola respon yang diperoleh akan dianalisis secara kuantitatif menggunakan pendekatan modern (*Item Response Theory* atau IRT) dengan bantuan program BILOG MG V3.0 model 1-PL, 2-PL, dan 3-PL; dan untuk melihat apakah terdapat perbedaan kemampuan yang signifikan pada kemampuan siswa yang diestimasi menggunakan model 1-PL, 2-PL, dan 3-PL maka digunakan uji *One-Way Anova Repeated Measure* (Anova pengukuran berulang). Hasil penelitian menunjukkan bahwa rerata tingkat kesukaran (*b*) baik, daya beda (*a*) baik, dan *pseudo-guessing* (*c*) baik Alat ukur yang disusun guru cocok bagi siswa yang memiliki kemampuan kimia sedang karena hanya mampu mengukur kemampuan kimia pada kisaran interval [-1,0 sampai +1,7]. Fungsi informasi tes maksimum diperoleh sebesar 68,83 (SEM = 0,121) pada kemampuan 0,2 logit. Selain menjadi model yang paling cocok dengan data penelitian ini, model 2-PL menghasilkan estimasi kemampuan yang paling tinggi dibandingkan kedua model lainnya. Rerata kemampuan siswa kelas XI IPA sebesar -0,0185 logit termasuk dalam kategori sedang.

Kata Kunci : *kualitas butir soal, kemampuan kimia, Item Response Theory*

ABSTRACT

This first aim of this study is to describe the quality of chemistry test item made by teacher. The test was developed for 11th grade students' science class in the first semester on academic year 2015/2016. The second aim of this study is to describe the characteristic of measurement's result for students' ability in chemistry. This is descriptive research design with the 101 student's responses patterns from multiple choice test device with 5 answer alternatives. The responses patterns were collected by documentation technique and analyzed quantitatively using Item Response Theory software such as BILOG MG V3.0 with 1-PL, 2-PL, and 3-PL models. The

differences of students' ability in chemistry in model 1-PL, 2-PL, dan 3-PL were analyzed using One-Way Anova Repeated Measure. The result showed that the mean of item difficulties level (b), item differentiate (a), and pseudo-guessing (c) are good. The measurement tools arranged by teacher were suitable for students who have the ability from -1.0 to +1.7. The maximum score of item information function is 68.83 (SEM =0.121) with ability in 0.2 logit. The highest ability's estimation score was showed by Model 2-PL. The mean of students' ability for 11th grade students is -0.0185 logit and consider as moderate category.

Keyword : *test Item quality, chemistry's ability, item response theory*

PENDAHULUAN

Dalam pasal 8 dijelaskan bahwa guru sebagai pendidik profesional wajib memiliki kualifikasi akademik, kompetensi, dan sertifikat pendidik serta sehat jasmani dan rohani demi mewujudkan tujuan pendidikan nasional [1]. Berdasarkan hal tersebut, maka salah satu kompetensi yang wajib dimiliki guru adalah kompetensi pedagogi. Kompetensi pedagogi yang dimiliki khususnya adalah kemampuan dalam menyelenggarakan penilaian proses dan hasil belajar yang terdiri dari: (a) memahami prinsip-prinsip penilaian hasil belajar sesuai dengan karakteristik mata pelajaran yang diampu, (b) menentukan aspek-aspek penilaian hasil belajar yang penting untuk dinilai, (c) menentukan prosedur penilaian hasil belajar, (d) mengembangkan instrumen penilaian hasil belajar, (e) mengadministrasikan penilaian proses dan hasil belajar secara berkesinambungan dengan menggunakan berbagai instrumen, serta (f) melakukan evaluasi proses dan hasil belajar [2]. Dalam melakukan evaluasi khususnya evaluasi hasil belajar, umumnya guru menggunakan sistem ujian.

Ujian atau tes adalah prosedur evaluasi yang biasa dilakukan oleh seorang guru terhadap pengetahuan dan ketrampilan siswa untuk mengetahui kinerjanya dengan meng-gunakan instrumen tertentu [3]. Pemilihan bentuk tes yang tepat ditentukan oleh tujuan tes, jumlah peserta tes, waktu yang tersedia untuk memeriksa lembar jawaban tes, ca-kupan materi tes, dan karakteristik mata pelajaran yang diujikan [4]. Tes prestasi be-lajar (*achievement test*) merupakan salah satu bentuk tes untuk mendapatkan data yang merupakan informasi untuk melihat seberapa banyak pengetahuan yang telah dimiliki dan dikuasai oleh seseorang sebagai akibat dari pendidikan dan pelatihan [5]. Tes prestasi belajar yang digunakan dapat berupa tes yang telah distandarkan (*standardized test*) maupun tes buatan guru sendiri (*teachermade test* atau *informal test*) [6]. Tes buatan guru adalah tes hasil belajar yang disusun oleh guru sendiri untuk kepen-tingan pengukuran dan penilaian prestasi belajar siswa, baik pada setiap penyajian satu-satuan pelajaran maupun pada ujian formatif dan sumatif [7,8].

Tes pilihan ganda merupakan salah satu bentuk tes selected response yang luas penggunaannya untuk berbagai macam keperluan misalnya: ulangan umum, ulangan kenaikan kelas, ujian akhir sekolah, ujian akhir nasional, survey internasional seperti *Trends in Mathematics and Science Study* (TIMSS) maupun *Programme for International Student Assessment* (PISA), tes

bahasa Inggris yang diselenggarakan oleh lembaga *testing* di luar negeri seperti TOEFL, IELTS, TOEIC, GRE, dan bakat skolastik. Hal tersebut tidak terlepas dari keunggulan bentuk tes pilihan ganda yang efektif untuk mengukur berbagai jenis pengetahuan dan hasil belajar yang kompleks [9], sangat tepat untuk ujian yang pesertanya banyak dan hasilnya harus segera diumumkan [10], serta karena jumlah dapat banyak maka faktor reliabilitas bertambah [11]. Namun, rupanya terdapat beberapa kelemahan, yaitu: (a) siswa tidak mempunyai keleluasaan dalam menulis, mengorganisasikan, dan mengekspresikan gagasan yang mereka miliki yang dituangkan ke dalam kata atau kalimatnya sendiri; (b) tidak dapat digunakan untuk mengukur kemampuan *problem solving*; (c) sangat sensitif terhadap terkaan; (d) penyusunan tes yang baik memerlukan waktu yang relatif lama dibandingkan dengan bentuk tes yang lainnya; serta (e) sangat sukar menentukan alternatif jawaban (*distractor*) yang benar-benar homogen, logis, dan berfungsi [10].

Hasil prasurvey di beberapa SMA Negeri Kota Yogyakarta menunjukkan bahwa bentuk tes pilihan ganda merupakan bentuk tes yang paling sering digunakan guru untuk mengukur kemampuan kognitif siswa, tidak terkecuali guru kimia. Contoh instrumen tes bentuk pilihan ganda yang umumnya dibuat sendiri oleh guru untuk melakukan pengukuran kemampuan kimia siswa di sekolahnya adalah instrumen tes mid semester, baik semester ganjil maupun genap. Tes buatan guru ini tentu saja termasuk tes yang tidak standar karena tidak didahului ujicoba, butir belum terkalibrasi, serta aspek validitas dan reliabilitas yang belum diketahui.

Instrumen tes mid semester kimia yang disusun guru haruslah memenuhi kriteria sebagai alat ukur yang baik agar dapat memberikan gambaran tentang kemampuan maupun kompetensi yang dimiliki siswa. Untuk menguji setiap butir soal yang pada akhirnya digunakan untuk melaksanakan tes, maka perlu dilakukan analisis butir soal [12]. Kegiatan menganalisis butir soal merupakan suatu kegiatan yang harus dilakukan guru untuk meningkatkan mutu butir soal yang ditulis. Dari hasil analisis tersebut, pada akhirnya akan mencerminkan karakteristik yang dimiliki oleh perangkat tes itu sendiri,

Dalam pengukuran pendidikan, terdapat dua pendekatan yang sering digunakan untuk melakukan analisis butir soal, yaitu *Classical Test Theory*, *CTT* (Teori Tes Klasik) dan *Item Response Theory, IRT* (Teori Respons Butir) [13, 14, 15, 16]. Gulliksen (1950) menyatakan bahwa CTT merupakan cikal bakal berkembangnya teori pengukuran [17]. Namun apabila CTT yang digunakan, hasil pengukuran kurang merefleksikan kemampuan siswa yang sebenarnya. Hal ini disebabkan karena siswa menjawab butir soal suatu tes yang berbentuk pilihan ganda akan diberi skor 1 jika benar dan skor 0 jika salah, sehingga kemampuan siswa dinyatakan dengan skor total yang diperolehnya. Prosedur tersebut kurang memperhatikan interaksi antara setiap orang siswa dengan butir. Namun, pendekatan IRT merupakan pendekatan alternatif yang dapat digunakan dalam menganalisis suatu tes. Hal ini dikarenakan

IRT menggunakan model probabilistik. Model ini bermakna bahwa probabilitas subjek untuk menjawab butir dengan benar bergantung pada kemampuan subjek dan karakteristik butir. Artinya, peserta tes berkemampuan tinggi mempunyai probabilitas menjawab benar lebih besar dibandingkan peserta tes yang berkemampuan rendah. Selain itu, masih ada beberapa kelemahan yang dimiliki oleh CTT, yaitu: (a) tingkat kesukaran dan daya beda butir soal tergantung pada kelompok peserta yang mengerjakannya, (b) karakteristik butir tes berubah seiring waktu, (c) penggunaan metode dan teknik untuk desain dan analisis tes dengan memperbandingkan kemampuan siswa pada pembagian kelompok atas, tengah, dan bawah, (d) skor tes berada dalam fungsi linear, (e) konsep reliabilitas skor didefinisikan dari istilah tes paralel, (f) tidak ada dasar teori untuk menentukan bagaimana peserta memperoleh tes yang sesuai dengan kemampuan peserta yang bersangkutan, dan (g) *Standard Error Measurement* (SEM) berlaku pada seluruh peserta tes [18, 19]. Berdasarkan kelemahan tersebut, maka IRT muncul untuk mengatasi kelemahan yang ada pada CTT.

Salah satu program analisis butir soal yang berbasis IRT adalah BILOG-MG V3.0. Analisis menggunakan program ini melibatkan tiga model logistik yaitu model logistik satu parameter (1-PL), dua parameter (2-PL), dan tiga parameter (3-PL). Analisis dengan program BILOG menghasilkan output dalam bentuk tiga fase. Fase pertama merupakan estimasi butir berdasarkan teori tes klasik, fase kedua estimasi parameter butir berdasarkan IRT, dan fase ketiga estimasi kemampuan peserta tes [20].

Pada fase pertama diperoleh informasi tentang banyaknya *testee* yang menjawab benar, proporsi peluang menjawab benar dibagi peluang menjawab salah, serta koefisien korelasi biserial. Item yang memiliki nilai koefisien biserial negatif dapat mengganggu proses analisis, sehingga item tersebut tidak diikutkan dalam tahap analisis berikutnya. Fase kedua, estimasi parameter butir. Pada fase ini diperoleh informasi tentang parameter butir sesuai dengan model Parameter Logistik (PL) yang digunakan. Untuk model 1-PL didapatkan estimasi tingkat kesukaran, model 2-PL didapatkan estimasi tingkat kesukaran dan daya beda, serta model 3-PL didapatkan estimasi tingkat kesukaran, daya beda, dan tebakan semu atau *pseudo-guessing* [21]. Selain parameter butir, pada fase kedua juga dihasilkan statistik kecocokan suatu butir dengan model atau *goodness of fit*. Model yang digunakan untuk estimasi parameter adalah model logistik yang banyak menerima butir cocok. Kecocokan butir ini sangatlah penting mengingat penerapan IRT dapat dibenarkan hanya ketika data sudah sesuai dengan modelnya [22]

Program BILOG menggunakan statistik uji *likelihood ratio chi-square* (selanjutnya disebut *chi square*) untuk menguji kecocokan model. Secara empiris, kualitas butir ditelaah berdasarkan kecocokan data dengan model dan nilai parameter butir. Kecocokan suatu item dengan model dapat dilihat dari nilai *chi square* item dibandingkan dengan harga kritik distribusi *chi square* sesuai dengan dk item yang bersangkutan pada taraf signifikansi α.

Butir dikatakan cocok dengan model jika nilai χ^2 item lebih kecil atau sama dengan nilai distribusi χ^2; atau dikatakan cocok model jika probabilitas χ^2 ≥ 0,01. Taraf signifikansi (α) = 0,01 merupakan nilai default dari program BILOG dengan derajat bebas (*degree of freedom*, df) yang sudah ditetapkan oleh program [20]. Sementara itu, fase ketiga menampilkan estimasi parameter kemampuan (*θ*) peserta tes dan fungsi informasi tes. Estimasi parameter, baik butir maupun kemampuan peserta digunakan metode Bayessian karena metode tersebut merupakan metode default yang sudah ditetapkan oleh program [20].

Berdasarkan uraian yang telah dipaparkan di atas, maka penelitian ini bertujuan untuk: (a) mendeskripsikan kualitas butir soal hasil pengembangan alat ukur (soal mid semester 1 mata pelajaran kimia bagi kelas XI-IPA) yang dibuat oleh guru ditinjau dari rerata tingkat kesukaran, daya beda, pseudo*guessing*, model logistik yang paling fit dengan data penelitian, fungsi informasi tes maksimum, serta kesalahan pengukuran; dan (b) mengetahui hasil pengukuran kemampuan kimia siswa.

METODE PENELITIAN

Penelitian ini merupakan penelitian deskriptif yang menggambarkan karakteristik soal kimia buatan guru di salah satu SMA Negeri di Kota Yogyakarta beserta karakteristik kemampuan siswa dalam mata pelajaran kimia. Subjek dalam penelitian adalah 101 pola respon siswa terhadap perangkat tes berupa soal pilihan ganda lima alternatif jawaban pada mid semester I mata pelajaran kimia kelas XI IPA Tahun Ajaran 2015/2016 yang dikumpulkan melalui teknik dokumentasi. Sedangkan objek dalam penelitian ini adalah kualitas alat ukur dan prestasi belajar siswa yang dilihat dari estimasi kemampuan (*θ*). Pola respon yang diperoleh akan dianalisis secara kuantitatif menggunakan pendekatan modern (*Item Response Theory* atau IRT) dengan bantuan program BILOG MG V3.0 model 1-PL, 2-PL, dan 3-PL. Untuk melihat apakah terdapat perbedaan kemampuan yang signifikan pada kemampuan siswa yang diestimasi menggunakan model 1-PL, 2-PL, dan 3-PL maka digunakan uji *One-Way Anova Repeated Measure* (Anova pengukuran berulang).

HASIL DAN PEMBAHASAN

a. Kualitas Butir Soal Kimia Buatan Guru

Hasil analisis butir soal menggunakan pendekatan modern menghasilkan informasi bahwa sebanyak 28 butir fit dengan model 1-PL, 37 butir fit dengan model 2-PL, dan 36 butir fit dengan model 3-PL. Berdasarkan hal tersebut dapat disimpulkan bahwa model yang paling sesuai untuk soal mid semester I mata pelajaran kimia adalah model 2-PL, hal ini dikarenakan model tersebut menghasilkan butir fit yang paling banyak. Selanjutnya, selain merupakan model yang paling fit, model 2-PL juga menghasilkan 33 butir (84,62%) yang termasuk dalam kategori butir baik. Sementara model 1-PL dan 3-PL hanya menghasilkan berturut-turut 27 butir (67,50%) dan 28 butir (75,68%) butir baik.

Output pada phase 1 memuat informasi tentang estimasi parameter butir

berdasarkan teori tes klasik yaitu berupa indeks daya beda butir yang dapat ditafsirkan dari nilai korelasi bisernya. Meskipun diestimasi menggunakan model logistik yang berbeda, hasil output dari phase 1 tetaplah sama. Berdasarkan daya bedanya, butir soal dikatakan baik (diterima) apabila daya bedanya (r_{bis}) minimal 0,3 [23, 24, 25, 26]. Hasil analisis kuantitatif menunjukkan bahwa terdapat 80% butir soal memiliki daya beda yang baik, artinya 32 butir tersebut dapat membedakan siswa berkemampuan tinggi dengan siswa berkemampuan rendah. Sementara 20% sisanya, butir soal tidak cukup mampu dalam membedakan siswa berkemampuan tinggi dengan siswa berkemampuan rendah, bahkan diantaranya terdapat 3 butir memiliki daya beda yang negatif yaitu butir nomor 3, 6, dan 37. Daya beda negatif mengindikasikan bahwa siswa dengan kemampuan tinggi (kelompok atas) menjawab butir dengan salah, sementara siswa dengan kemampuan rendah (kelompok bawah) menjawab butir dengan benar. Selain memberikan informasi mengenai indeks daya beda, output phase 1 juga mengidentifikasi kelayakan masing-masing butir soal. Untuk model 1-PL, semua butir layak dianalisis. Untuk model 2-PL, butir nomor 37 tidak layak dianalisis; dan untuk model 3-PL butir nomor 3, 6, serta 37 tidak layak dianalisis.

Output pada phase 2 memuat informasi tentang estimasi parameter butir sesuai model logistik yang digunakan. Untuk parameter tingkat kesukaran butir, pada masing-masing model 1-PL, 2-PL, dan 3-PL diperoleh hasil berturut-turut sebagai berikut: 92,5% (37 butir); 100% (39 butir); dan 94,59% (35 butir) berkategori baik. Ditinjau dari tujuan pelaksanaan tes, perlu diperhatikan bahwa butir soal yang terlalu mudah atau terlalu sukar mungkin memang kurang memberikan informasi yang berguna bagi peserta tes pada umumnya. Hasil analisis menunjukkan informasi yang bervariasi terkait indeks kesukaran butir soal. Tingkat kesukaran butir soal yang baik berkisar antara -2 logit ≤ b_i ≤ 2 logit [18,27]. Nilai yang semakin mendekati -2 logit mengindikasikan butir semakin mudah, dan nilai yang mendekati +2 logit mengindikasikan butir semakin sukar. Tingkat kesukaran yang telah dirumuskan oleh guru memang tidak sesuai dengan tingkat kesukaran hasil empirik. Hal ini dikarenakan dalam membuat item tersebut, guru mengklasifikasikan item ke dalam tingkat kesukaran tertentu (mudah, sedang, dan sukar) hanya berdasarkan intuisinya [28]. Belum tentu item yang dianggap guru sebagai item"sulit" juga dirasakan sulit oleh siswa karena sangat sulit menentukan seberapa sulit item dalam suatu tes sebelum siswa melakukan tes [29].

Pada analisis butir berdasarkan pendekatan modern, daya beda butir (*a*) hanya akan muncul jika parameter butir diestimasi menggunakan model 2-PL dan 3-PL. Biasanya rentang daya beda berada antara 0-2 logit [18], meskipun sebenarnya batasnya adalah positif tak hingga [29]. Hasil analisis memperlihatkan bahwa rerata daya beda yang diestimasi menggunakan model 3-PL lebih tinggi daripada daya beda yang diestimasi menggunakan model 2-PL (2,03414>1,55664). Artinya, model 3-PL memberikan butir-butir yang lebih sensitif dalam membedakan kemampuan siswa. Setidaknya terdapat dua beberapa penyebab

suatu butir memiliki daya beda rendah, yaitu: (a) tingkat kesukaran butir soal yang terlalu rendah (butir soal terlalu sukar) atau terlalu tinggi (butir soal terlalu mudah), (b) pengecoh yang tidak masuk akal meskipun butir soal tersebut memiliki tingkat kesukaran yang diterima [30]. Keberadaan pengecoh yang tidak masuk akal ini akan memudahkan siswa untuk memutuskan bahwa pengecoh tersebut salah sehingga kemungkinan siswa menjawab benar dengan menebak sangat tinggi dan menyebabkan butir soal menjadi terlalu mudah. Sebaliknya, pengecoh yang terlalu dekat nilai kebenarannya dengan kunci dapat menyebabkan butir soal menjadi terlalu sulit.

Indeks tebakan semu (*pseudo-guessing*) hanya akan muncul jika parameter butir diestimasi menggunakan model 3-PL. Indeks ini merefleksikan hasil perilaku menebak jawaban, dimana besarnya indeks pada tes pilihan ganda terletak di sekitar seperbanyaknya pilihan jawaban. Misalnya pada tes dengan pilihan 4 jawaban, maka nilai c_i terletak di sekitar ¼ atau 0,25 [31, 32]. Dalam instrumen tes ini, banyaknya alternatif jawaban adalah 5, maka nilai c_i akan terletak di sekitar $^1/_5$ atau 0,20. Hasil analisis menunjukkan rerata *pseudoguessing* sebesar 0,13422 tergolong cukup baik karena dibawah 0,20 (untuk lima alternatif jawaban). Meskipun begitu, terdapat 20% (8 butir) yang memiliki indeks tebakan semu yang cukup tinggi yaitu butir nomor 2, 8, 14, 22, 27, 28, 36, dan 38. Dari kedelapan butir tersebut, butir nomor 28 adalah butir yang memiliki indeks tebakan semu tertinggi (*c*=0,271). Penyebab dari tingginya indeks tebakan semu dari butir-butir tersebut sebenarnya tidak terlepas dari peranan pengecoh seperti yang telah dijelaskan di pembahasan daya beda.

b. Kemampuan Kimia Siswa

Hasil pengukuran terhadap kemampuan kimia siswa dapat ditafsirkan dari output phase 3.

Tabel 1. Statistik Deskriptif θ pada Ketiga Model Logistik

Statistics				
		teta_1PL	teta_2PL	teta_3PL
N	Valid	101	101	101
	Missing	0	0	0
Mean		-,2521	-,0185	-,0475
Median		-,5414	-,3495	-,3036
Std. Deviation		1,10191	1,00153	1,03435
Skewness		,525	,314	,335

Mencermati hasil pada Tabel 1, tampak bahwa rerata θ berdasarkan model 2-PL hampir sama dengan 3-PL, dan rerata θ dari kedua model ini lebih tinggi dari rerata berdasarkan model 1-PL. Jika dilihat berdasarkan nilai simpangan baku, penyebaran θ dari model 2-PL dan 3-PL relatif sama. Sementara θ model 1-PL lebih menyebar dari reratanya. Meskipun begitu, distribusi θ hasil estimasi dari ketiga model logistik menunjukkan nilai skewness yang positif, artinya distribusi θ juling ke kanan yang menunjukkan bahwa sebagian besar siswa memiliki kemampuan kimia yang sedikit dibawah rata-rata, atau cenderung sedang.

Untuk melihat apakah terdapat perbedaan kemampuan yang signifikan pada siswa yang diestimasi menggunakan

model 1-PL, 2-PL, dan 3-PL maka digunakan uji *One-Way Anova Repeated Measure* (Anova pengukuran berulang). Analisis variansi dengan rancangan pengukuran berulang diterapkan karena semua subjek yang sama terlibat pada semua kondisi percobaan [33, 34]. Keakurasian uji F pada Anova dengan pengukuran tidak berulang tergantung pada asumsi bahwa teta-teta (kemampuan-kemampuan) yang diperoleh dari kondisi yang berbeda bersifat independen sedangkan Anova pengukuran berulang melanggar asumsi tersebut. Hal ini disebabkan karena teta-teta yang diperoleh dari masing-masing model logistik saling berhubungan sebagai akibat dari penggunaan subjek yang sama. Oleh karena itu, asumsi tambahan diperlukan untuk analisis lebih lanjut. Asumsi tersebut disebut dengan asumsi *Sphrecity*. *Sphrecity* mengacu pada kesamaan variansi perbedaan teta antar pelakuan [35]. Analisis dengan anova pengukuran berulang diperoleh dengan bantuan SPSS 21.

Selanjutnya, untuk menentukan model logistik manakah yang lebih baik digunakan pada analisis variansi pengukuran berulang (*repeated measures*) dengan model-model logistik dianggap sebagai perlakuan. Penerapan Anova dengan pengukuran berulang menggunakan asumsi *sphrecity*. Asumsi ini terpenuhi jika ada kesamaan "secara kasar" variansi selisih teta antar perlakuan. Tabel 2 merupakan uji Mauchly untuk menguji asumsi *sphrecity* dengan tingkat signifikansi α = 0,05. Uji ini menguji hipotesis bahwa variansi selisih teta antar perlakuan sama [35].

Tabel 2. Uji Mauchly untuk menguji asumsi *sphrecity*

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
teta	,899	10,566	2	**,005**	,908	,924	,500

Hasil pada Tabel 2 menunjukkan bahwa *p-value* sebesar 0,005 lebih kecil dari tingkat signifikansi α = 0,05 sehingga dapat disimpulkan bahwa ada perbedaan yang signifikan variansi selisih teta antar perlakuan artinya asumsi *sphrecity* telah dilanggar. Pelanggaran terhadap asumsi ini menyebabkan perlu ada koreksi terhadap derajat bebas (df) sehingga menghasilkan rasio F yang valid. SPSS menghasilkan tiga koreksi berdasarkan estimasi *Sphrecity* yang diberikan oleh Greenhose & Geisser (dinyatakan dengan $\acute{\varepsilon}$) Huynh & Feldt (dinyatakan dengan $\tilde{\varepsilon}$), dan menggunakan batas bawah [35].

Tabel. 3. Tabel Anova dengan Nilai *Sphrecity* Terkoreksi

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Teta	Sphericity Assumed	3,275	2	1,638	54,775	**,000**	,354	109,550	1,000
	Greenhouse-Geisser	3,275	1,816	1,803	54,775	**,000**	,354	99,480	1,000
	Huynh-Feldt	3,275	1,848	1,772	54,775	**,000**	,354	101,217	1,000
	Lower-bound	3,275	1,000	3,275	54,775	**,000**	,354	54,775	1,000
Error (teta)	Sphericity Assumed	5,979	200	,030					
	Greenhouse-Geisser	5,979	181,6	,033					
	Huynh-Feldt	5,979	184,7	,032					
	Lower-bound	5,979	100	,060					

Tabel 3 adalah tabel Anova dengan nilai yang sudah dikoreksi untuk masing-masing estimasi *Sphrecity.* Ketiga hasil koreksi tersebut menunjukkan nilai signifikansi sebesar 0,000 yang kurang dari α=0,05. Artinya, terdapat perbedaan yang signifikan variansi selisih teta antarperlakuan sehingga asumsi *Sphrecity* belum terpenuhi. Oleh karena itu, diperlukan analisis variansi multivariate (Manova) karena Manova tidak tergantung pada asumsi *Sphrecity* [35]. Selain itu, Anova pengukuran berulang merupakan kasus khusus dari Manova [36]. Prosedur Anova pengukuran berulang dengan SPSS secara otomatis menghasilkan uji multivariat seperti ditunjukkan pada Tabel 4.

Tabel 4. Uji Multivariate

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[c]
teta	Pillai's Trace	,464	42,819[b]	2,000	99,000	**,000**	,464	85,637	1,000
	Wilks' Lambda	,536	42,819[b]	2,000	99,000	**,000**	,464	85,637	1,000
	Hotelling's Trace	,865	42,819[b]	2,000	99,000	**,000**	,464	85,637	1,000
	Roy's Largest Root	,865	42,819[b]	2,000	99,000	**,000**	,464	85,637	1,000

Tabel 4 merupakan hasil empat uji statistik multivariat yang paling umum digunakan [35, 36] dengan masing-masing eigen value yang ekuivalen dengan nilai F-hitung pada Anova masing-masing dari keempat kriteria tersebut memiliki *p-value* 0,000 yang lebih kecil dari α=0,05. Hal ini mengindikasikan bahwa estimasi teta memiliki perbedaan yang signifikan antara model logistik. Selanjutnya, akan diuji model-model logistik manakah yang berbeda dengan menggunakan analisis univariate.

Berdasarkan hasil sebelumnya diketahui bahwa asumsi *Sphrecity* tidak terpenuhi, sehingga uji perbandingan ganda dilakukan menggunakan metode Bonferroni. Metode Bonferroni digunakan karena metode ini paling tahan terhadap pelanggaran asumsi *Sphrecity* [35]. Hasil uji perbandingan ganda ditunjukkan pada Tabel 5.

Tabel 5. Uji Perbandingan Ganda

Pairwise Comparisons

Measure: MEASURE_1

(I) teta	(J) teta	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	-,234*	,028	,000	-,301	-,166
	3	-,205*	,024	,000	-,263	-,147
2	1	,234*	,028	,000	,166	,301
	3	,029	,021	,511	-,022	,080
3	1	,205*	,024	,000	,147	,263
	2	-,029	,021	,511	-,080	,022

Berdasarkan Tabel 5 diperoleh bahwa perbandingan antar θ menunjukkan bahwa perbedaan rerata yang tidak signifikan terjadi hanya antara model 2-PL dengan 3-PL karena memiliki *p-value* sebesar 0,511 yang lebih besar dari α = 0,05. Namun jika dilihat dari *mean difference*, dapat dikatakan bahwa model 2-PL menghasilkan estimasi kemampuan yang lebih tinggi daripada model 3-PL, sehingga dapat disimpulkan bahwa model 2-PL lebih baik daripada model 3-PL maupun model 1-PL. Model 2-PL menghasilkan estimasi kemampuan yang paling tinggi dibandingkan kedua model lainnya, dikarenakan model ini merupakan model yang paling cocok bagi data respon yang dianalisis ini. Model yang cocok akan memiliki kemampuan melakukan generalisasi untuk memprediksi data berikutnya atau data yang berbeda.

Berdasarkan kesimpulan di atas, maka pembahasan yang lebih mendalam hanya dilakukan berdasarkan IRT model 2-PL. Dari model tersebut, hasil analisis kemampuan siswa menunjukkan rerata kemampuan siswa sebesar -0,0185 logit yang artinya rerata kemampuan kimia siswa kelas XI IPA SMA N Yogyakarta termasuk dalam kategori sedang. Informasi maksimum dicapai pada kemampuan 0,2 logit [kisaran interval -1,0 sampai +1,7 logit] dengan nilai fungsi informasi 66,83 dan kesalahan baku pengukuran 0,121. Dari hasil tersebut dapat disimpulkan bahwa perangkat tes yang dianalisis cocok untuk kelompok siswa yang berkemampuan sedang. Hal ini konsisten dengan hasil sebelumnya bahwa rerata tingkat kesukaran butir soal ($\bar{b}$ = -0,00182 logit) yang sedikit lebih rendah dari rerata kemampuan ($\bar{\theta}$ = -0,0185 logit).

KESIMPULAN

Analisis menggunakan pendekatan IRT untuk model 1-PL, 2-PL, dan 3-PL secara keseluruhan menyimpulkan:

1. Kualitas Butir Soal Kimia Buatan Guru
   a. Rerata tingkat kesukaran (*b*) baik, daya beda (*a*) baik, dan *guessing* (*c*) baik.
   b. Model 2-PL adalah model yang paling cocok dengan data penelitian ini.

c. Perangkat tes yang disusun guru cocok bagi siswa yang memiliki kemampuan kimia sedang karena hanya mampu mengukur kemampuan kimia pada kisaran interval [-1,0 logit sampai +1,7 logit].
d. Fungsi informasi tes maksimum diperoleh sebesar 68,83 (SEM = 0,121) pada kemampuan +0,2 logit.

2. Rerata kemampuan kimia siswa kelas XI IPA SMA N Yogyakarta tergolong dalam kategori sedang (-0,0185 logit).

UCAPAN TERIMAKASIH

1. Universitas Negeri Yogyakarta
2. Lembaga Pengelola Dana Pendidikan

DAFTAR RUJUKAN

[1] Republik Indonesia. (2005). *Undang-Undang RI Nomor 14, Tahun 2005, tentang Guru dan Dosen.*

[2] Kementrian Pendidikan dan Kebudayaan. (2007). *Permendiknas No.16, Tahun 2007, tentang Standar Kualifikasi Akademik dan Kompetensi Guru.*

[3] Bambang Sumintono. (Maret 2016). *Aplikasi Permodelan Rasch pada Asesmen Pendidikan: Implementasi Penilaian Formatif (Assessment for Learning)*. Makalah disajikan dalam Kuliah Umum pada Jurusan Statistika Institut Teknologi Sepuluh November, di Surabaya.

[4] Djemari Mardapi. (2012). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Yogyakarta: Nuha Litera.

[5] Anastasi, A. & Urbina, S. (2008). *Psychological Testing*. New Jersey: Prentice Hall, Inc.

[6] Grondlund, N.E. (1986). *Measurement and Evaluation in Teaching (4th Ed)*. New York: MacMillan Publishing Company.

[7] *P*opham, W.J. (1995). *Classroom Assessment: What Teachers Need To Know*. Boston: Allyn and Bacon.

[8] Cangelosi, J.S. (1995). *Merancang Tes Untuk Menilai Prestasi Siswa* (Terjemahan Lilian D. Tedjasudjana). Bandung: Penerbit ITB. (Buku asli diterbitkan tahun 1990).

[9] Miller, M.D., Linn, R.L., & Grondlund N.E. (2009). *Measurement and Assessment in Teaching (10th Ed).* New Jersey: Pearson Education, Inc.

[10] Sumarna Surapranata. (2005). *Panduan Penulisan Tes Tertulis (Penilaian Berbasis Kelas)*. Bandung: Remaja Rosdakarya.

[11] Tresna Sastrawijaya. (1988). *Proses Belajar Mengajar Kimia*. Jakarta: Depdiknas.

[12] Kaplan, R.M., & Saccuzo. (1982). Psychological Testing, Principles Applications and Issue. Monterey California: Books/Cole Publishing Company.

[13] Awopeju, O. A. & Afolabi, E. R. I. (2016). *European Scientific Journal*. 12(28). 263-284.

[14] Guler, N., Uyanik, G. K., & Teker, G. T. (2013). *European Journal of Research on Education*. 2(1). 1-6.

[15] Sharkness, J. & DeAngelo, L. (2011). *Research in Higher Education*. 52. 480-507.

[16] Fan, X. (1998). *Educational and Psychological Measurement*. 58(3). 357-673.

[17] Engruven, M. (2013). *Journal of Education*. ISSN 2298-0172. 23-30.

[18] Hambleton, R.K., & Swaminathan, H. (1985). *Items Response Theory: Principles and Application.* Boston: Kluwer-Nijjhoff Publish.

[19] Qasem, M. A. N. (2013).*Journal of Research and Method in Education*. 3(5). 77-81.

[20] Mislevy, R.J,. & Bock, R.D. (1990). *BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models (2nd Ed.)*. Mooresville: Scientific Software Inc.

[21] Kalekar, S. (2015). *Scholarly Research Journal for Humanity Science & English Language*. 2(10). 2564-2568.

[22] Kose, I. A. (2014). *Educational Research and Reviews*, 9(17). 642-649.

[23] Mardapi, D. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendekia.

[24] Talebi, G. A., Ghaffari, R., Eshandarzadeh, E., & Oskouei, A. E. (2013). *Research and Development in Medical Education*. 2(2). 20-23.

[25] Kartowagiran, B. (2012). *Penulisan Butir Soal*. Makalah disampaikan pada Pelatihan penulisan dan analisis butir soal bagi Sumber daya PNS Dik-Rekinpeg, di Hotel Kawanua Aerotel, Jakarta pada tanggal 10 Oktober 2012.

[26] Sayyah, M., Vakili, Z., Alavi, N. M., Bidgeli, M., Solemani, A., Assaian, M., & Azarbad, Z. (2012). *Nursing and Midwifery Studies*. 1(2). 83-87.

[27] Adedoyin, O.O., & Mokobi, T. (2013). *International Journal of Asian Social Sciences*. 3 (4). 992-1011.

[28] Stanley, J.C., & Wang M.D. (1968). *Differential Weighting: A Survey of Methods and Empirical Studies*. USA: Departmen of Health, Education, & Welfare.

[29] Baker, F.B. (2001). *The Basics of Item Response Theory (2nd Ed*). USA: ERIC Clearinghouse on Assessment and Evaluation.

[30] Thorndike, R.M. (2005). *Measurement and Evaluation in Psychology and Education (7th Ed).* New Jersey: Pearson Education Inc.

[31] Naga, D. S. (1992). *Teori Sekor pada Pengukuran Pendidikan*. Jakarta: Gunadarma.

[32] Huriaty, D., & Mardapi, D. (2014). *Jurnal Penelitian dan Evaluasi Pendidikan*. 18(2). 188-201.

[33] Park, E., Cho, M., & Ki, C. (2009). *Korean Journal of Laboratory Medicine*. 29(1). 1-9.

[34] Hager, W. (2007). *Psychology Science*. 49(3). 209-222.

[35] Field, A. (2009). Discovering Statistics Using SPSS (3rd Ed.) London: Sage Publication, Inc.

[36] Hair, J.F., Black, W.C., & Babin, W.J., dkk. (2006). *Multivariate Data Analysis (6th Ed.)*. New Jersey: Pearson Prentice Hall