Mind Set, Desember 2010, hal. 28 - 52
ISSN 2086 - 1966

Vol. 2, No. 1

Penerapan Model *Polytomous Item Response Theory* untuk Mengevaluasi Skala *Student's Perception Of Assessment Questionnaire*

YUSUF HADI YUDHA

Fakultas Psikologi, Universitas Pancasila
Srengseng Sawah, Jagakarsa - Jakarta Selatan 12640
E-mail: yh_yudha@yahoo.com

Diterima 10 Juli 2010, Disetujui 10 Desember 2010

Abstract : The aims of this study were to evaluate reliability and validity of the Students' Perception of Assessment Questionnaire (SPAQ), to evaluate and refine the instrument to improve its capacity as a measure through IRT modelling, to detect Differential Item Functioning (DIF), to evaluate students' perception on assessment, and to evaluate gender-based, grade-based, class program-based, school location-based and school status-based differences in students' perceptions. Results showed that the validity and reliability coefficients revealed that the SPAQ was suitable for assessing students' perceptions on five assessment dimensions: Congruence With Planned Learning, Authenticity, Student Consultation, Transparency and Diversity. For the IRT models applied, only Generalized Partial Credit Model shows that 30 items of the SPAQ scale fit to the data. There are six item shows DIF in two categories: Item 14, 17, 18, 21, 24, and 25. The average scale-item mean values for all the scales were less than 3.0, which indicates a need to address these dimensions of assessment at classroom level. The perceptions of students grouped on the basis of gender, of grade level, of class program, of school location and of school status groups were comparable, but on the basis of school location groups were statistically significantly different for each sub scale. Specifically how a student thinks about their assessment will influence their response.

Key words: Perception, assessment, item response theory, partial credit model, generalized partial credit model, graded response model, differential item functioning

PENDAHULUAN

Pada dasarnya, penilaian (*assessment*) merupakan salah satu aspek penting dalam pembelajaran. Black dan William (1998) mensintesis lebih dari 250 studi yang menghubungkan antara penilaian dengan pembelajaran, dan menemukan bahwa intensitas penggunaan penilaian di kelas yang bertujuan meningkatkan pembelajaran dapat meningkatkan prestasi siswa. Meningkatkan frekuensi penilaian, bagaimanapun tidak secara langsung dapat meningkatkan pembelajaran. Akan tetapi, jika guru dapat memanfaatkan penilaian kelas untuk meningkatkan pengetahuan, ketrampilan, dan kepercayaan diri yang akan digunakan oleh para siswa dalam proses pembelajaran, menggunakan pengetahuan yang telah didapat sebagai dasar untuk meningkatkan metode pembelajaran, dan memonitor perubahan persepsi siswa dalam kegiatan belajar mengajar, penilaian kelas dapat meningkatkan pembelajaran.

Proses pembelajaran juga akan meningkat ketika para siswa terdorong untuk memikirkan apa yang telah mereka pelajari, mengingat kembali pengalaman belajar mereka dan untuk menerapkan apa yang telah mereka pelajari untuk proses belajar di masa yang akan datang. Dengan adanya penilaian, proses tersebut dapat menjadi siklus yang berkelanjutan. Ketika para siswa dan guru merasa nyaman siklus yang berkelanjutan dan melakukan berbagai penyesuaian, proses pembelajaran akan menjadi lebih efisien dan para siswa mulai

menyadari bahwa mereka dapat mempelajari lebih dari yang telah mereka peroleh di kelas dan mulai memikirkan untuk mengetahui hal-hal baru yang tidak/belum diajarkan di kelas, sehingga mereka tidak hanya mempermasalahkan tentang kualitas guru atau ketelitiannya. Dan ketika para siswa mulai terlibat dalam pengalaman berkelanjutan ini, mereka bisa memonitor proses belajar mereka, melakukan evaluasi, dan mengembangkan suatu kebiasaan berpikir untuk secara terus menerus meninjau ulang apa yang telah mereka ketahui dan mengasah rasa ingin tahu.

Dietel, Herman dan Knuth (1991) menyarankan beberapa karakateristik penting dalam membuat suatu penilaian yang baik, salah satu di antaranya adalah keterlibatan siswa. Mereka berpendapat bahwa penilaian yang baik seharusnya melibatkan siswa dalam menentukan tujuan dari penilaian tersebut dan membuat kriteria penilaian secara bersama, serta mengerjakan tugas-tugas yang dapat mengukur aktivitas-aktivitas pembelajaran sesuai dengan permasalahan di kehidupan nyata. Elliot dan Harackiewicz (1994) menambahkan bahwa keterlibatan siswa dalam proses penilaian menunjukkan sejauh mana siswa memiliki kepedulian akan hasil terbaik yang dapat dicapai dalam proses belajarnya, waktu yang dicurahkan untuk tugas-tugas yang harus dikerjakan, serta sejauh mana siswa terlibat dalam proses pembelajaran.

Dalam kenyataannya guru sangat jarang melibatkan siswa dalam mengembangkan proses penilaian di kelas. Padahal, interaksi antara guru dan murid di dalam proses pembelajaran merupakan bagian yang menentukan pembelajaran yang efektif. Efektivitas pembelajaran dicirikan atau mensyaratkan adanya peran serta aktif dari murid dalam pembelajaran. Kemampuan guru mendorong para murid aktif dalam proses pembelajaran menjadi faktor penting dalam menciptakan pembelajaran yang bermutu. Keberhasilan murid dalam memahami atau menguasai apa yang disampaikan guru dalam pembelajaran (konsep atau bahan ajar) tidak dapat dipisahkan dari kemampuan guru mengkomunikasikannya. Untuk itu diperlukan kemampuan atau keterampilan guru berkomunikasi secara efektif dalam menyelenggarakan pembelajaran yang bermutu (*quality teaching*). Selain itu, juga diperlukan kemampuan guru dalam menerapkan dan mengembangkan pendekatan-pendekatan partisipatif (*active learning*). Solomon (1996) menjelaskan bahwa keterlibatan peserta belajar yang aktif dalam pembelajaran merupakan mata rantai penting yang menghubungkan perilaku guru dan prestasi siswa, sehingga apa yang dilakukan siswa merupakan penentu yang lebih penting dalam pembelajaran daripada apa yang dilakukan guru. Dengan melibatkan siswa dalam proses pengajaran, penilaian dan hasil belajar, validitas dari proses penilaian tersebut dapat tercapai dan kesalahan pengukuran dari suatu penilaian dapat dihindari (Steinberg, 2000).

Biggs (1999, 2003), Ramsden (1992) dan Entwistle (1981) memiliki pandangan yang sama tentang prestasi belajar siswa, mereka berpendapat bahwa faktor yang berpengaruh langsung terhadap prestasi belajar adalah pendekatan belajar. Pendekatan belajar itu sendiri dipengaruhi oleh persepsi siswa terhadap penilaian yang diberikan oleh guru, dan persepsi yang terbentuk ini dipengaruhi oleh karakteristik siswa. Dengan demikian karakteristik dan persepsi siswa terhadap penilaian akan mempengaruhi pendekatan dan strategi belajar, yang kemudian akan mempengaruhi hasil belajar (Pandia, 2006).

Persepsi siswa dapat memberikan gambaran penting terhadap jalannya kegiatan belajar mengajar di kelas. Informasi mengenai siswa dan lingkungan pembelajaran di kelas dapat diobservasi oleh pihak luar, yang kemungkinan akan memberikan penilaian yang lebih objektif, namun hal ini akan menjadi sulit bagi mereka untuk mendapatkan gambaran sesungguhnya tentang situasi kelas tanpa terlibat langsung di dalamnya. Sedangkan siswa sebaliknya, mereka terlibat langsung di dalam proses tersebut, sehingga akan menciptakan kesan yang lebih mendalam (Moos, 1979).

Salah satu alat ukur untuk menilai kegiatan pembelajaran di kelas berdasarkan persepsi siswa adalah Skala *Student's Perceptions of Assessment Questionnaire* (SPAQ), di mana para siswa diminta untuk merespon 30 pernyataan yang berkaitan dengan persepsi mereka tentang apa yang dilakukan oleh guru selama proses pengajaran berlangsung, dan sejauh mana mereka mendapatkan sesuatu dari yang telah dipelajarinya.

SPAQ merupakan hasil pengembangan dan proses validasi alat ukur yang dilakukan di Essex, England (Dorman & Knightley, 2005) dan di Australia (Fisher, Waldrip, & Dorman, 2005), terdiri dari 30 item dan dikelompokkan

ke dalam 5 subskala untuk menilai kegiatan pembelajaran berdasarkan persepsi siswa, yakni *Congruence with planned learning* (Kesesuaian dengan Rencana Pembelajaran), *Authenticity* (Kesesuaian dengan Kehidupan Nyata), *Student Consultation* (Konsultasi Siswa), *Transparency* (Keterbukaan), *Diversity* (Keberagaman). Beberapa penelitian dengan menggunakan skala SPAQ (Tabel 1) menunjukkan bahwa alat ukur tersebut valid dan reliabel (Koul & Fisher, 2005; Dorman, Fisher & Waldrip, 2005; Lalor, 2006). Namun sepengetahuan peneliti, alat ukur tersebut belum pernah diujikan kepada siswa-siswi SMA di Indonesia. Dan dengan menerapkan model *polytomous IRT*, diharapkan dapat memberikan evaluasi terhadap pengembangan skala SPAQ yang sangat berguna untuk menilai kegiatan pembelajaran di kelas.

Persepsi Siswa Terhadap Penilaian. Diartikan sebagai tindakan siswa dalam memandang suatu penilaian dalam suatu proses belajar mengajar yang sedang diikuti (Vande Watering, et al., 2006). Pada awal tahun 1960an, penelitian mengenai pembelajaran di kelas seringkali dilakukan peneliti dengan cara mengobservasi perilaku guru dan siswa. Di akhir tahun 1960-an, pengukuran di lingkungan psikososial pembelajaran di kelas merupakan komponen yang menentukan arah dalam meramalkan dan selalu mencari cara terbaik untuk kesuksesan pembelajar (Anderson & Walberg, 1974). Sejak saat itu banyak penelitian yang menunjukkan bahwa persepsi siswa terhadap lingkungan (psikososial) pembelajaran di kelas dapat diukur dengan instrumen melalui survei, dan hasil penelitian mereka dijamin validitasnya (Anderson & Walberg, 1974; Fraser, 1997, 1998a, 1998b, 2002b; Moos, 1979).

Dalam suatu studi, Schaffner et al. (2000) melakukan penelitian terhadap persepsi siswa di Amerika. Sampel yang digunakan adalah siswa kelas 4 sampai dengan kelas 12, di mana mereka memberikan respon terhadap sejumlah pertanyaan yang terdapat dalam kuesioner. Instrumen yang digunakan saat itu bernama skala *Perception of Assessment Task* (PAT) (Schaffner, Bury, Stock, Cho, Boney, & Hamilton, 2000). Instrumen tersebut terdiri dari 55 item pertanyaan dan dikelompokkan dalam 6 subskala, di mana pertanyaan yang diberikan berkaitan dengan perasaan siswa tentang apa yang dilakukan oleh guru selama proses pengajaran berlangsung, dan sejauh mana mereka mendapatkan sesuatu yang telah dipelajarinya (Dorman, J. P., Fisher, D. L. & Waldrip, B. G., 2005).

Dalam perkembangannya, Schaffner et al (2000) melakukan penelitian serupa dengan menggunakan instrumen yang sama, terhadap 470 siswa di 3 sekolah Australian bagian Barat yang terdiri dari siswa kelas 8, 9 dan 10 di 20 kelas program sains yang berbeda. Untuk mendapatkan gambaran lebih jauh mengenai validitas instrumen tersebut, Schaffner et al. (2000) melakukan wawancara lebih mendalam terhadap 40 siswa yang dipilih secara acak dari keseluruhan siswa yang menjadi sampel. Dari hasil analisis *internal consistency reliability* dan exploratory *factor analysis*, dilakukan perbaikan terhadap instrumen tersebut sehingga menghasilkan sebuah skala baru yang hanya terdiri dari 30 item pertanyaan dan dikelompokkan dalam 5 subskala. Instrumen tersebut dinamakan *Student Perceptions of Assessment Questionnaire* (SPAQ) (Koul, R. & Fisher, D., 2005). SPAQ meliputi lima aspek yang diukur (Cavanagh, R., Waldrip, B., Romanoski, J., Dorman, J. & Fisher, D., 2005):

1. *Congruence with Planned Learning* : sejauh mana tugas yang dinilai, materi ujian yang diberikan sesuai sasaran, tujuan, dan kegiatan program pembelajaran.
2. *Authenticity* : sejauh mana tugas-tugas dan ujian yang diberikan sesuai dengan situasi nyata yang dialami siswa.
3. *Student Consultation* : sejauh mana para siswa dapat berkonsultasi dan mendapat informasi yang berkaitan dengan tugas-tugas yang diberikan.
4. *Transparency* : sejauh mana tujuan dan bentuk dan materi tugas, ujian dijelaskan dengan sungguh-sungguh kepada siswa.
5. *Diversity* : sejauh mana para siswa mempunyai kesempatan yang sama dalam mengerjakan tugas.

Teori Tes Klasik. Teori ini sudah dikembangkan dan diaplikasikan sejak awal abad ke-20 (Schumacker, 2005), namun penggunaannya tetap dipertahankan sampai saat ini. Teori tes klasik dikembangkan berdasarkan tiga konsep dasar dalam pengukuran, yaitu: *test score/observed score, true score dan error score*. Teori tersebut menyatakan bahwa *test score/observed score* merupakan hasil penjumlahan dari *true score* dengan *error score*,

yang dalam bentuk model matematis yang dikenal dengan model skor murni *(true score model)*.

$X = T + E$

di mana:

X = *test score / observed score*

T = *true score (expected test score)*

E = *random error score*

Asumsi-Asumsi Teori Tes Klasik. Inti dari teori tes klasik itu berupa asumsi-asumsi yang dapat dirumuskan secara matematis. Asumsi-asumsi tersebut antara lain:

1. $\hat{E} = 0, E(X) = T$
   Nilai harapan skor-skor kesalahan seorang subjek sama dengan nol. Jika seorang dites dengan suatu tes yang sama berulang-ulang (sampai tidak terhingga), maka rata-rata skor kesalahannya akan sama dengan nol.
2. $P_{T\hat{E}} = 0$
   Skor murni dan skor kesalahan yang dicapai oleh suatu populasi subjek pada suatu tes tidak berkorelasi satu sama lain.
3. $P(E_1, E_2) = 0$
   Skor-skor kesalahan pada dua tes (mengukur hal yang sama) tidak saling berkorelasi.

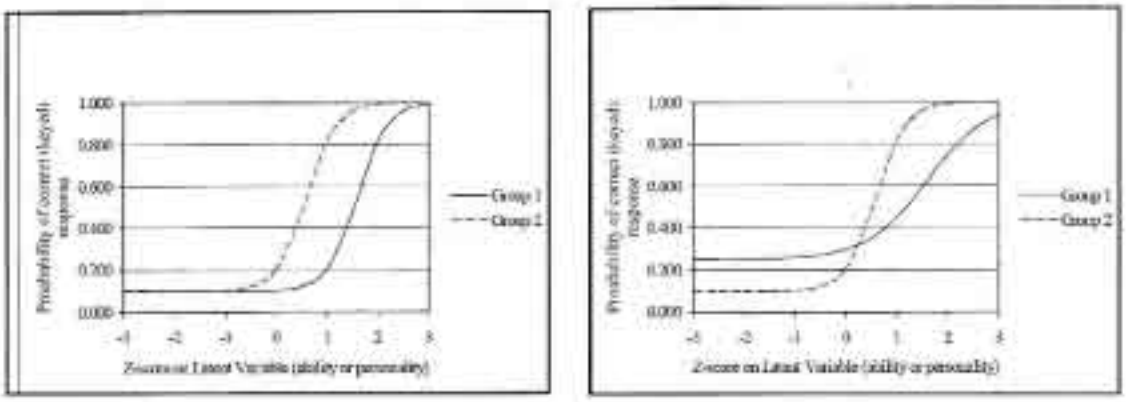
Keunggulan Teori Tes Klasik. Beberapa keunggulan yang dapat diperoleh dengan menerapkan model teori tes klasik antara lain, analisis tetap dapat dilakukan walaupun dengan jumlah sampel yang relatif kecil. Sehingga dalam melakukan uji coba atau melakukan tes dengan alat ukur (skala) tertentu dapat dilakukan dengan sampel yang kecil (seperti penggunaan analisis butir soal pada tingkat penilaian kelas). Selain itu, prosedur analisis dalam teori tes klasik relatif sederhana dan estimasi parameter model secara konseptual dapat dilakukan secara langsung. Keunggulan lain dari teori tes klasik adalah modelnya didasarkan pada asumsi yang lemah, yakni asumsi yang dapat dipenuhi dengan mudah oleh kebanyakan data tes.

Kelemahan Teori Tes Klasik. Dapat dikatakan bahwa kelemahan utama teori tes klasik adalah bahwa alat ukur yang disusun berdasarkan tes klasik itu terikat kepada sampel *(sample bound)*. Karena itu jika seperangkat tes diberikan kepada kelompok subyek yang rendah kemampuannya akan merupakan tes yang sukar, dan apabila diberikan kelompok subyek yang tinggi kemampuannya akan merupakan tes yang mudah. Demikian pula kalau dilihat dari arah subyek. Sekelompok subyek akan terlihat mempunyai kemampuan tinggi kalau mereka mengerjakan tes yang mudah, dan akan terlihat berkemampuan rendah apabila mereka mengerjakan tes yang sulit. Hal-hal seperti di atas akan menimbulkan kesukaran-kesukaran, terutama dalam kehidupan praktis. Selain itu dalam teori tes klasik juga sulit untuk menyeleksi soal-soal yang tingkat kesukarannya sesuai dengan kemampuan siswa yang akan diukur. Jika soal diberikan kepada kelompok orang-orang yang pandai, maka berdasarkan teori tes klasik, tingkat kesukaran soal tersebut akan terlihat mudah, karena sebagian besar akan menjawab dengan benar. Tetapi kalau soal-soal tersebut diberikan kepada orang-orang yang kurang pandai, maka soal tersebut akan terlihat sukar, karena kemungkinan sebagian besar orang-orang dalam kelompok tersebut tidak dapat menjawab dengan benar. Jadi dalam teori tes klasik tingkat kesukaran soal tidak tetap, tergantung kepada tingkat kemampuan sampel siswa yang menempuh tes tersebut.

Teori Tes Modern. Dasar dari teori tes modern adalah sifat-sifat atau kemampuan yang tidak dapat dilihat (latent) yang mendasari kinerja *(performance)* atau respon subjek terhadap butir soal (item) tertentu. Oleh karena itu, teori ini juga sering disebut sebagai *latent trait theory* atau lebih populer dengan istilah *item response theory* (IRT). IRT dikembangkan berdasarkan dua konsep dasar (Suryabrata, 2000), yaitu:

1. Kinerja (*performance*) seorang subyek pada suatu butir soal dapat diprediksikan atau dijelaskan dari suatu perangkat faktor-faktor yang disebut sifat-sifat, atau sifat-sifat laten, atau kemampuan (*ability*)
2. Hubungan kinerja subyek pada suatu butir soal dan perangkat sifat-sifat yang mendasari kinerja dapat dideskripsikan dengan fungsi meningkat secara monoton yang disebut fungsi *Item Characteristic Function (ICC)*. Fungsi ini menyatakan bahwa apabila taraf sifat (kemampuan) meningkat, maka probabilitas suatu respons yang benar terhadap suatu butir soal juga meningkat.

Keunggulan IRT. IRT memiliki banyak keunggulan yang disebut dengan sifat parameter item dan parameter kemampuan yang invarian (*invariance property*), yaitu karakteristik item



Gambar 1. Contoh ICC dari Item yang Mengandung *Uniform* DIF (kiri) dan *Non Uniform* DIF (kanan)

kedua kelompok peserta tes adalah tidak sama (*nonuniform*) untuk setiap tingkat kemampuan (*ability levels*).

Masalah Penelitian. Persepsi siswa mengenai apa yang dianggap bernilai dan tidak bernilai oleh guru dalam penilaian berdampak penting terhadap perilaku belajar siswa, dan hal ini akan mempengaruhi prestasi belajarnya (Fry, Ketteridge & Marshal, 1999). Berkenaan dengan hal tersebut, maka permasalahan yang akan dikaji dalam penelitian ini adalah:

- Apakah item-item dalam SPAQ dapat mengukur persepsi siswa terhadap penilaian?
- Model IRT manakah yang dapat digunakan untuk menjelaskan dan mengevaluasi item-item pada skala SPAQ?
- Dalam karakteristik siswa yang sama, apakah siswa cenderung untuk memberikan respon yang sama untuk setiap item?

METODE

Responden Penelitian. Sesuai dengan tujuan yang hendak dicapai, maka sampel dalam

Tabel 1. Karakteristik Responden

Nama Sekolah	Program	Kelas	Jenis Kelamin		Total
			Laki-Laki	Perempuan	
SMAN 8	Reguler	Kelas X	17	21	38
		Kelas XI	11	21	32
	Internasional	Kelas X	13	32	45
SMAN 21	Reguler	Kelas X	66	68	134
		Kelas XI	0	1	1
		Kelas XII	0	1	1
	Internasional	Kelas XI	8	12	20
		Kelas XII	2	4	6
SMAN 68	Reguler	Kelas X	17	17	34
	Internasional	Kelas X	3	7	10
		Kelas XI	10	4	14
SMAN 70	Internasional	Kelas X	6	11	17
		Kelas XI	9	11	20
SMA Tarakanita 1	Reguler	Kelas X	0	148	148
		Kelas XI	0	73	73
		Kelas XII	0	46	46
		Total Sampel	**162**	**477**	**639**

penelitian ini adalah siswa SMA di Jakarta kelas X, XI dan XII IPA, baik siswa program reguler maupun program kelas internasional. Sedangkan sekolah-sekolah yang dipilih adalah: SMAN 8, SMAN 21, SMAN 68, SMAN 70, dan SMA Tarakanita 1 dengan rincian seperti pada, Tabel 1.

Dengan kriteria tersebut diharapkan subyek memiliki pengetahuan terhadap objek persepsinya, dalam hal ini adalah penilaian pelajaran IPA (Fisika, Kimia dan Biologi) yang diberikan oleh guru di kelas. Penilaian tersebut termasuk ujian, tugas-tugas, projek, portfolio serta jenis penilaian-penilaian lainnya.

Teknik Pengambilan Sampel. Metode yang digunakan dalam penelitian ini adalah *non probability*, di mana tidak semua subyek memiliki kesempatan yang sama untuk dipilih sebagai sampel (Kerlinger, 1986). Sementara teknik yang digunakan untuk mendapatkan sampel adalah *accidental sampling technique*, yaitu cara pengambilan subyek berdasarkan yang paling mudah untuk ditemui dan bersedia menjadi subyek penelitian tetapi masih memenuhi kriteria menjadi subyek penelitian, yaitu siswa SMA yang memilih kelas IPA.

Instrumen Penelitian. Skala *Student Perception of Assessment Questionnaire* (SPAQ) dikembangkan di Essex, Inggris (Dorman & Knightley, 2005) dan di Australia (Fisher, Waldrip, & Dorman, 2005), terdiri dari 30 item dan dikelompokkan ke dalam 5 subskala untuk menilai kegiatan pembelajaran berdasarkan persepsi siswa, yakni : *Congruence with planned learning* (Kesesuaian dengan Rencana Pembelajaran), *Authencity* (Kesesuaian dengan Kehidupan Nyata), *Student Consultation* (Konsultasi Siswa), *Transparency* (Keterbukaan), *Diversity* (Keberagaman), dan mempunyai hubungan dengan lingkungan pembelajaran di kelas *(learning environment)*. Format skala SPAQ menggunakan model skala Likert 4 pilihan (1 = hampir tidak pernah, 2 = kadang-kadang, 3 = sering, 4 = hampir selalu / sering sekali) dan terdiri dari 30 item untuk mengukur respon siswa terhadap kegiatan proses pembelajaran.

Prosedur Penelitian. Berkaitan dengan jalannya penelitian ini, peneliti merencanakan langkah-langkah prosedur penelitian yang diharapkan dapat menunjang kelancaran serta keberhasilan penelitian, yaitu sebagai berikut:
Tahapan pemilihan dan penerjemahan alat ukur yang digunakan antara lain:

- Penelusuran latar belakang pentingnya penelitian ini dilakukan
- Pencarian alat ukur (beserta dukungan teori) yang dianggap sesuai dengan latar belakang dan tujuan penelitian.
- Setelah dipilih alat ukur *SPAQ Scale*, dilakukan penterjemahan terhadap item-item *SPAQ Scale* ke dalam Bahasa Indonesia, serta dilakukan beberapa penyesuaian item agar mudah dimengerti

Tabel 2. Definisi Operasional Skala SPAQ

Student Perception of Assessment Questionnaire (SPAQ) Scale		
Sub Skala	**Definisi Operasional**	**No. Item**
Congruence with planned learning	Sejauh mana tugas yang dinilai, materi ujian yang diberikan sesuai sasaran, tujuan, dan kegiatan program pembelajaran.	1, 2, 3 4, 5, 6
Authenticity	Sejauh mana tugas-tugas dan ujian yang diberikan sesuai dengan situasi nyata yang dialami siswa	7, 8, 9 10, 11, 12
Student Consultation	Sejauh mana para siswa dapat berkonsultasi dan mendapat informasi yang berkaitan dengan tugas-tugas yang diberikan	13, 14, 15, 16, 17, 18
Transparency	Sejauh mana tujuan dan bentuk dan materi tugas, ujian dijelaskan dengan sungguh-sungguh kepada siswa	19, 20, 21, 22, 23, 24
Diversity	Sejauh mana para siswa mempunyai kesempatan yang sama dalam mengerjakan tugas	25, 26, 27, 28, 29, 30

Tabel 3. *Item Scale Correlation Skala SPAQ*

No Item	*Item-Scale Correlation*	No Item	*Item-Scale Correlation*	No Item	*Item-Scale Correlation*
1	0,632	11	0,684	21	0,720
2	0,694	12	0,673	22	0,612
3	0,706	13	0,597	23	0,673
4	0,677	14	0,638	24	0,647
5	0,687	15	0,648	25	0,429
6	0,643	16	0,585	26	0,593
7	0,634	17	0,579	27	0,529
8	0,756	18	0,661	28	0,626
9	0,712	19	0,667	29	0,654
10	0,743	20	0,710	30	0,654

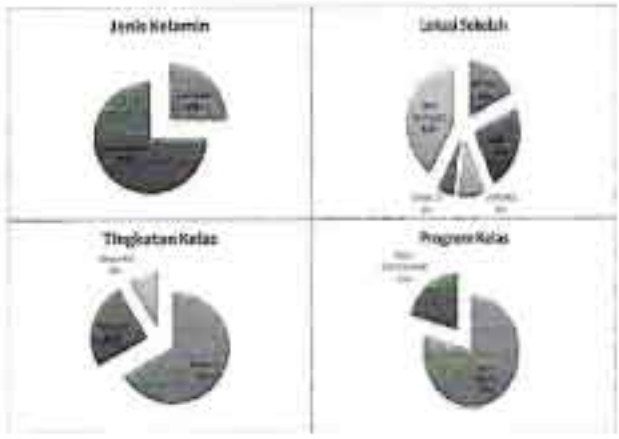
oleh siswa di Indonesia. Sebelum melakukan penterjemahan dan beberapa penyesuaian, peneliti meminta izin kepada pengembang *SPAQ Scale* dan yang mempublikasikannya, yaitu Prof. Bruce Waldrip (*University of Southern Queensland*) dan Dr. Rob Cavanagh (*Curtin University of Technology*) dengan cara mengirimkan *email* kepada yang bersangkutan.

- Selanjutnya, hasil terjemahan tersebut dikonsultasikan kepada seorang ahli bahasa (bilingualis) yang menguasai Bahasa Indonesia dan Bahasa Inggris.

Analisis Data. Analisis data hasil respon siswa terhadap skala SPAQ dilakukan dalam empat tahap. Tahap pertama adalah analisis deskriptif dan uji reliabilitas *(alpha cronbach* dan *internal consistency)*. Analisis ini dilakukan dengan menggunakan bantuan program SPSS versi 13.0. Tahap kedua analisis adalah pengujian asumsi *unidimensionality* sekaligus menguji validitas konstruk skala SPAQ, yaitu dengan metode *Confirmatory Factor Analysis*. Pada tahap ini, analisis dilakukan dengan menggunakan program LISREL versi 8.3 (Joreskog & Sorbom, 1996). Tahap ketiga analisis dilakukan menguji model *polytomous* IRT, apakah dapat fit dengan data. Model yang diujikan yaitu: *Partial Credit Model* (PCM), *Generalized Partial Credit Model* (GPCM) dan *Graded Response Model* (GRM). Analisis dilakukan dengan menggunakan program PARSCALE for Windows versi 4.1 (Muraki, 2003). Tahap keempat analisis adalah pengujian terhadap *Differential Item Functioning* (DIF) dengan membagi kelompok sampel berdasarkan jenis kelamin, program kelas, serta status sekolah. Pada tahap ini, analisis dilakukan dengan menggunakan program QUEST versi 2.1 (Adams & Khoo, 1996).

HASIL

Karakteristik Responden. Responden dalam penelitian ini adalah siswa-siswi SMA unggulan yang tersebar di wilayah DKI Jakarta. Berikut ini merupakan gambaran umum mengenai responden berdasarkan jenis kelamin, lokasi sekolah, tingkatan kelas, dan program kelas.



Gambar 2. Karakteristik Responden Berdasarkan Jenis Kelamin, Lokasi Sekolah, Tingkatan Kelas dan Program

Berdasarkan Gambar 2 dapat diketahui bahwa mayoritas responden dalam penelitian ini adalah perempuan. Hal ini dikarenakan jumlah sampel terbanyak adalah dari SMA Tarakanita, yaitu sebesar 42%. Sebagaimana diketahui bahwa di sekolah tersebut, semua siswanya adalah perempuan. Berdasarkan tingkatan kelas, diketahui bahwa responden terbanyak adalah siswa kelas X (67%), serta sebagian besar dari responden merupakan siswa Program Kelas Reguler. Hal ini disebabkan sedikitnya jumlah siswa Program Kelas Internasional di masing-masing sekolah, bahkan untuk di SMA Tarakanita 1, hanya terdapat Program Kelas Reguler.

Reliabilitas Alat Ukur. Pengujian reliabilitas skala SPAQ dilakukan dengan mengukur *internal consistency* dari skala tersebut, yaitu dengan melihat *Item-Scale Correlation, Inter-Scale Correlation* dan Reliabilitas Cronbach Alpha. Dalam pengujian ini digunakan program ITEMAN dan SPSS versi 13.0, hasil yang diperoleh dapat dilihat pada tabel Tabel 3.

Tabel 4 menunjukkan keseluruhan item memiliki korelasi yang cukup tinggi (di atas 0,5) dengan skala SPAQ. Hal ini merupakan salah satu bukti bahwa item-item tersebut mengukur aspek yang sama, yaitu persepsi siswa terhadap penilaian di kelas.

Tahap selanjutnya dalam menguji reliabilitas adalah dengan melihat nilai *Inter-Scale Correlation*. Hasil pengujian dengan progam SPSS versi 13.0 menunjukkan kelima sub skala SPAQ (*Congruence with planned learning, Authenticity, Student Consultation, Transparency* dan *Diversity*) saling berkorelasi positif satu sama lain. Walaupun korelasi antar sub skala

Tabel 4. *Inter-Scale Correlation* Skala SPAQ

Skala SPAQ	*Congruence With Planned Learning*	*Authenticity*	*Student Consultation*	*Transparency*	*Diversity*
Congruence With Planned Learning	1,000	0,343	0,298	0,355	0,305
Authenticity	0,343	1,000	0,387	0,333	0,331
Student Consultation	0,298	0,387	1,000	0,431	0,368
Transparency	0,355	0,333	0,431	1,000	0,469
Diversity	0,305	0,331	0,368	0,469	1,000

tersebut tidak cukup kuat, namun seluruhnya signifikan pada taraf 0,01. Hal ini merupakan salah satu bukti, bahwa kelima sub skala SPAQ juga mengukur aspek yang sama, yaitu persepsi siswa terhadap penilaian di kelas.

Pengujian terhadap reliabilitas *Cronbach Alpha* menunjukkan nilai reliabilitas tertinggi adalah sebesar 0,791, yaitu untuk sub skala *Authenticity*. Sedangkan yang terendah sebesar 0,609, yaitu untuk sub skala *Diversity*. Sub skala *Congruence with planned learning* memiliki nilai reliabilitas sebesar 0,759, sub skala *Student consultation* sebesar 0,675, dan sub skala *Transparency* sebesar 0,751. Dengan melihat nilai reliabilitas untuk kelima sub skala berada di atas 0,6, maka dapat dikatakan bahwa item-item pada skala SPAQ cukup dapat diandalkan (reliabel).

Berdasarkan Tabel 5 terlihat bahwa rata-rata skor siswa untuk masing-masing sub skala secara keseluruhan berada di atas 2,0. Hal ini menunjukkan

Tabel 5. Reliabilitas *Cronbach Alpha* Skala SPAQ

Skala SPAQ	Rata-rata	Standar Deviasi	Cronbach Alpha
Congruence With Planned Learning	2,759	0,510	0,759
Authenticity	2,476	0,554	0,791
Student Consultation	2,103	0,467	0,675
Transparency	2,736	0,498	0,751
Diversity	2,597	0,457	0,609

bahwa, secara umum siswa memiliki persepsi yang positif terhadap penilaian di kelas.

Unidimensionality. Untuk menguji apakah item-item dalam skala SPAQ dapat *fit* dengan model IRT, maka perlu dilakukan pengujian terhadap item-item tersebut apakah benar unidimensi *(unidimensionality)*, karena pengujian tersebut merupakan salah satu asumsi yang harus terpenuhi dalam menerapkan model IRT (Hambleton, 1991). Salah satu cara yang dapat dilakukan untuk menguji asumsi tersebut adalah dengan *Confirmatory Factor Analysis* (CFA). CFA dilakukan dengan membuat suatu model pengukuran *(measurement model)* untuk menggambarkan sebaik apa indikator-indikator tersebut dapat digunakan sebagai instrumen pengukuran variabel laten. Selain untuk menguji asumsi *unidimensionality*, CFA juga digunakan untuk menguji validitas konstruk dari suatu alat ukur.

Dalam menguji asumsi *unidimensionality*, peneliti membuat beberapa alternatif model pengukuran, yaitu:

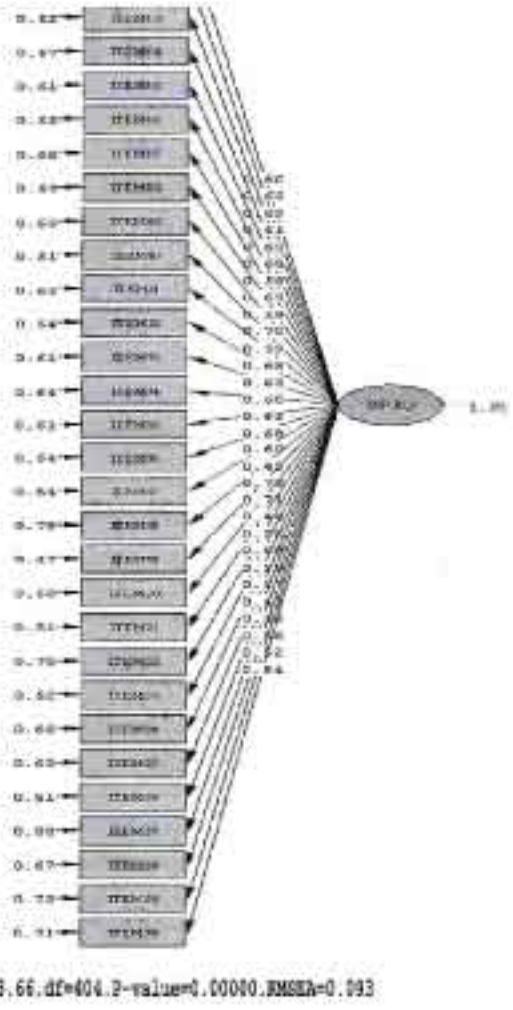
1. Model pengukuran dengan satu variabel laten (dalam hal ini skala SPAQ) dan 30 item-itemnya dijadikan sebagai indikator-indikator yang dapat diukur secara langsung (*first order factor*), maupun secara tidak langsung atau melalui lima sub skala SPAQ (*second order faktor*)
2. Model pengukuran dengan satu variabel laten (dalam hal ini skala SPAQ) dan 5 sub skala SPAQ (*Congruence with planned learning, Authenticity, Student Consultation, Transparency* dan *Diversity*) dijadikan sebagai indikator-indikator dengan cara menjumlahkan masing-masing skor dari setiap item (skor total) untuk setiap sub skala SPAQ.

Untuk menilai apakah model pengukuran tersebut benar-benar fit dengan data, perlu memperhatikan nilai indeks fit. Suatu indeks yang

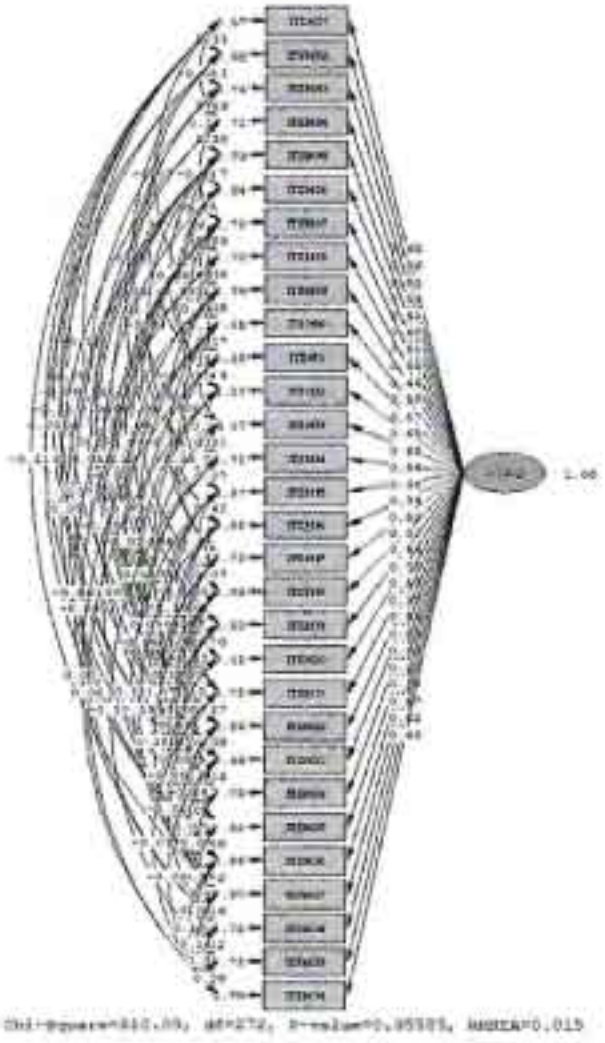
menunjukkan model tersebut fit, tidak memberikan jaminan bahwa model memang benar-benar fit, begitu pula sebaliknya. Oleh karena itu peneliti tidak hanya bergantung pada salah satu indeks fit untuk menguji model tersebut, akan tetapi menggunakan 3 kriteria fit yang umum digunakan, yaitu:

1. *P-value*
   Jika nilai P-value lebih besar dari 0,05, maka model dapat dikatakan fit dengan data
2. *Root Mean Square Error of Approximation* (RMSEA) < 0,10
   Jika nilai RMSEA lebih kecil dari 0,10, maka model dapat dikatakan fit dengan data
3. *Goodness of Fit Index* (GFI) ≥ 0,90
   Jika nilai GFI lebih besar atau sama dengan 0,90, maka model dapat dikatakan fit dengan data.

Sehingga jika memenuhi 3 kriteria tersebut di atas, maka dapat dikatakan model pengukuran tersebut fit dengan data.



Gambar 3. *Path Diagram First Order Factor* Skala SPAQ



Gambar 4. *Path Diagram* Modifikasi *First Order Factor* Skala SPAQ

***First Order Factor* Skala SPAQ**

Model pengukuran dengan *first order factor* ini dilakukan guna menguji *unidimensionality* dari skala SPAQ yang diukur secara langsung oleh ke-30 item-itemnya (indikator). Pengujian terhadap model tersebut menghasilkan nilai P-value = 0,00 (lebih kecil dari 0,05), RMSEA = 0,093 (lebih kecil dari 0,10) dan GFI = 0,90 (lebih besar atau sama dengan 0,90). Berdasarkan kriteria yang telah disebutkan sebelumnya di awal pembahasan, maka dapat disimpulkan model ini tidak benar-benar fit dengan data, karena dari ketiga kriteria, hanya 2 yang terpenuhi yaitu RMSEA dan GFI.

Oleh karena itu, peneliti melakukan modifikasi berdasarkan *modification index* dari *output* yang dihasilkan, antara lain dengan mengestimasi korelasi antar kesalahan pengukuran *(error)* hingga diperoleh model yang fit dengan data.

Hasil yang diperoleh dari ke-30 indikator item, seluruhnya dapat mengukur variabel laten (dalam hal ini skala SPAQ) dengan baik, hal ini terlihat dari nilai T-value yang seluruhnya lebih besar dari 1,96 (Tabel 6). Pengujian terhadap model pengukuran juga menunjukkan nilai P-value sebesar 0,05585 (lebih besar dari 0,05), nilai RMSEA sebesar 0,015 (lebih kecil dari 0,10) dan nilai GFI sebesar 0,99 (lebih besar dari 0,90). Sehingga dapat disimpulkan bahwa model tersebut fit dengan data, atau dengan kata lain ke-30 item tersebut merupakan indikator yang

Tabel 6. Hasil CFA First Order Factor Skala SPAQ

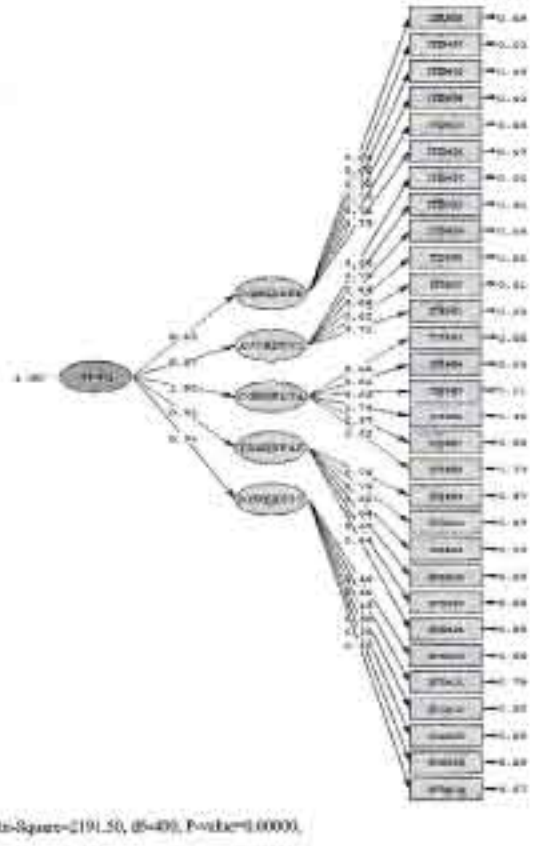
Variabel Laten	Indikator	Loading Factor	T-value	SEM
Skala SPAQ	Item 1	0,65	26,53	-0,02
	Item 2	0,56	23,46	-0,02
	Item 3	0,51	20,98	-0,02
	Item 4	0,54	19,33	-0,03
	Item 5	0,52	18,88	-0,03
	Item 6	0,60	25,54	-0,02
	Item 7	0,50	19,74	-0,03
	Item 8	0,55	22,96	-0,02
	Item 9	0,46	16,31	-0,03
	Item 10	0,59	22,13	-0,03
	Item 11	0,57	23,34	-0,02
	Item 12	0,65	31,34	-0,02
	Item 13	0,65	33,27	-0,02
	Item 14	0,54	22,55	-0,02
	Item 15	0,36	11,86	-0,03
	Item 16	0,34	9,94	-0,03
	Item 17	0,52	21,36	-0,02
	Item 18	0,32	10,94	-0,03
	Item 19	0,64	30,41	-0,02
	Item 20	0,70	31,28	-0,02
	Item 21	0,50	20,10	-0,02
	Item 22	0,37	13,62	-0,03
	Item 23	0,59	25,14	-0,02
	Item 24	0,50	20,05	-0,03
	Item 25	0,43	15,19	-0,03
	Item 26	0,37	12,68	-0,03
	Item 27	0,18	5,72	-0,03
	Item 28	0,47	17,25	-0,03
	Item 29	0,52	19,77	-0,03
	Item 30	0,46	16,32	-0,03

valid bagi pengukuran konstruk SPAQ.

Berdasarkan Tabel 6 dapat diketahui bahwa item indikator yang memberikan kontribusi terbesar untuk skala SPAQ Item 20 (*faktor loading* sebesar 0,70), sedangkan item indikator yang memberikan kontribusi terkecil adalah Item 27 (*faktor loading sebesar* 0,18).

***Second Order Factor* Skala SPAQ**

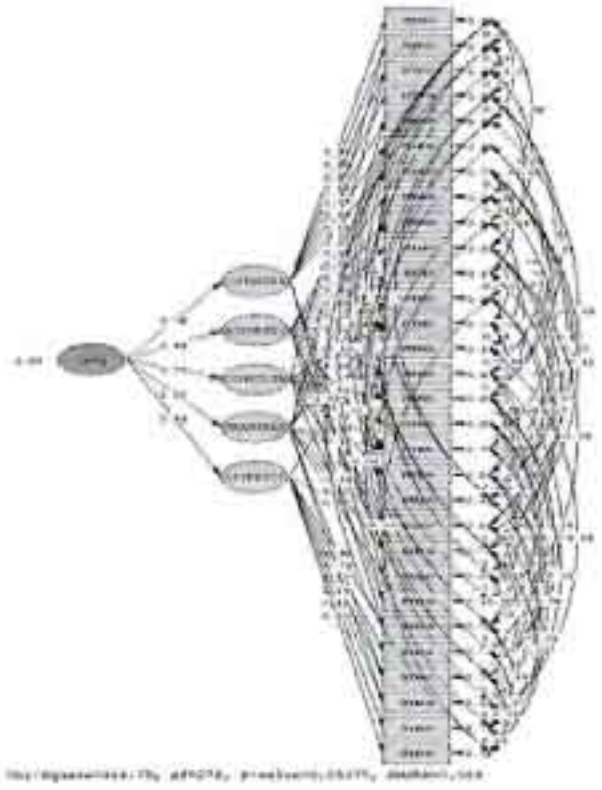
Model pengukuran dengan *second order factor* ini dilakukan guna menguji *unidimensionality* dari skala SPAQ yang diukur secara tidak langsung oleh ke-30 item-itemnya (indikator), tetapi melalui lima sub skala SPAQ, yaitu *Congruence with planned learning, Authenticity, Student Consultation, Transparency* dan *Diversity*. Pengujian terhadap model tersebut menghasilkan nilai P-value = 0,00 (lebih kecil dari 0,05), RMSEA = 0,084 (lebih kecil dari 0,10) dan GFI = 0,92 (lebih besar dari 0.90). Berdasarkan kriteria yang telah disebutkan sebelumnya di awal pembahasan, maka dapat disimpulkan model ini tidak benar-benar fit dengan data, karena dari ketiga kriteria, hanya 2 yang



Gambar 5. Path Diagram *Second Order Factor* Skala SPAQ

terpenuhi yaitu RMSEA dan GFI.

Oleh karena itu, peneliti melakukan modifikasi berdasarkan *modification index* dari output yang dihasilkan, antara lain dengan mengestimasi korelasi antar kesalahan pengukuran *(error)* hingga diperoleh model yang fit dengan data. Hasil yang diperoleh dari kelima indikator (sub skala SPAQ) dan ke-30 item indikator, seluruhnya dapat mengukur variabel laten (dalam hal ini skala SPAQ) dengan baik, hal ini terlihat dari nilai T-value yang seluruhnya lebih besar dari 1,96 (Tabel 7). Pengujian terhadap model pengukuran juga menunjukkan nilai



Gambar 6. Path Diagram Modifikasi *Second Order Factor* Skala SPAQ

P-value sebesar 0,06375 (lebih besar dari 0,05), nilai RMSEA sebesar 0,014 (lebih kecil dari 0,10) dan nilai GFI sebesar 0,97 (lebih besar dari 0,90). Sehingga dapat disimpulkan bahwa model tersebut fit dengan data, atau dengan kata lain ke-30 item tersebut merupakan indikator yang valid bagi pengukuran konstruk SPAQ

Dari Tabel 7 di atas juga dapat dilihat koefisien muatan faktor (*factor loading*) dari setiap item yang ada, semua bernilai positif. Item-item yang memberikan kontribusi terbesar untuk masing-masing sub skala *Congruence with planned learning, Authenticity, Student Consultation, Transparency* dan *Diversity* secara berturut-turut adalah item 5, item 10, item 17, item 21 dan item 28.

Model Fit dan Estimasi Parameter Item. Model-model pada *Item Response Theory* (IRT) tergantung pada bentuk matematik fungsi karakteristik item dan banyaknya parameter yang dilibatkan dalam model. Model yang sesuai (fit) dengan alat ukur (skala) tertentu, belum tentu fit dengan skala yang lain. Oleh karena itu, setelah dilakukan pengujian dimensi *(unidimensionality)*, maka perlu dilakukan pengujian terhadap kesesuaian antara model dengan skala yang akan dianalisis.

Tabel 7. Hasil CFA *Second Order Factor* Skala SPAQ

Sub Variabel Laten	Indikator	Loading Factor	T-value	SEM
Congruence With Planned Learning	Item 1	0,63	11,45	0,06
	Item 2	0,62	9,96	0,06
	Item 3	0,87	11,96	0,07
	Item 4	0,50	9,42	0,05
	Item 5	0,97	12,74	0,08
	Item 6	0,64	12,11	0,05
Authenticity	Item 7	0,39	7,89	0,05
	Item 8	0,66	15,96	0,04
	Item 9	0,73	15,20	0,05
	Item 10	0,81	18,44	0,04
	Item 11	0,58	12,24	0,05
	Item 12	0,50	10,90	0,05
Student Consultation	Item 13	0,26	3,97	0,07
	Item 14	0,60	11,83	0,05
	Item 15	0,17	3,67	0,05
	Item 16	0,33	6,98	0,05
	Item 17	0,62	12,13	0,05
	Item 18	0,60	8,61	0,07
Transparency	Item 19	0,18	2,32	0,08
	Item 20	0,74	15,20	0,05
	Item 21	0,92	4,19	0,22
	Item 22	0,45	5,90	0,08
	Item 23	0,57	13,90	0,04
	Item 24	0,82	13,88	0,06
Diversity	Item 25	0,59	3,98	0,15
	Item 26	0,50	4,04	0,12
	Item 27	1,17	3,24	0,36
	Item 28	1,19	3,56	0,33
	Item 29	0,46	3,91	0,12
	Item 30	0,50	4,11	0,12

Untuk skala *Student Perceptions of Assessment Questionnaire* (SPAQ) yang menggunakan metode rating (skala Likert), peneliti menerapkan 3 model *Polytomous IRT* yaitu: *Graded Response Model* (Samejima, 1969), *Partial Credit Model* (Masters, 1982), *Generalized Partial Credit Model* (Muraki, 1992). Pemilihan model yang sesuai tergantung pada asumsi yang sesuai bagi perangkat data yang akan dianalisis. Kesesuaian ini dapat dibuktikan kemudian dengan menunjukkan seberapa baik model tersebut dapat menjelaskan hasil tes yang diperoleh. Kriteria yang digunakan dalam pengujian model statistik *chi-square*. Jika nilai *probability* statistik *chi-square* lebih besar dari 0,05, maka dapat dikatakan model tersebut fit dengan data.

***Partial Credit Model* (PCM).** Tabel 8 menunjukkan hasil estimasi parameter item (*estimated item parameters*) untuk ke-30 item skala *Student Perceptions of Assessment Questionnaire* (SPAQ) dengan menggunakan PARSCALE for Windows versi 4.1 (Muraki, 2003).

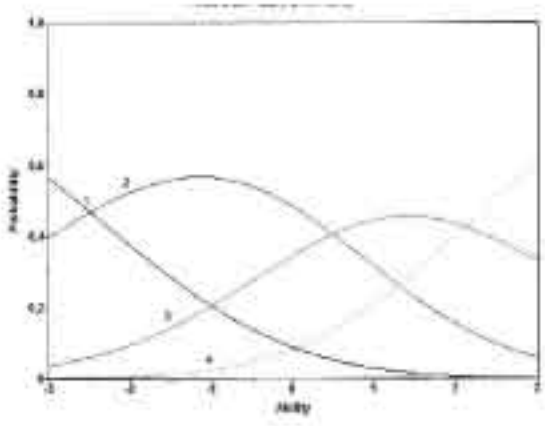
Tingkat kesulitan (*threshold*) untuk setiap kategori masing-masing item, seluruhnya

Tabel 8. ***Estimated Item Parameters*** **dan** ***Item-Fit Statistics*** **untuk** ***Partial Credit Model*** **dengan Program PARSCALE**

Item	b1	b2	b3	Chi-Square	df	Prob.	Keterangan
Item 1	-4,886	-0,718	1,471	39,929	38	0,384	Item Fit
Item 2	-3,281	-0,535	2,095	40,712	41	0,483	Item Fit
Item 3	-2,924	-1,061	1,709	33,225	38	0,690	Item Fit
Item 4	-3,262	-0,133	3,114	22,812	43	0,995	Item Fit
Item 5	-3,895	-0,302	2,678	42,973	42	0,429	Item Fit
Item 6	-4,133	-1,790	1,348	20,017	34	0,973	Item Fit
Item 7	-2,488	0,484	2,096	55,749	46	0,154	Item Fit
Item 8	-2,739	0,158	2,027	42,813	41	0,393	Item Fit
Item 9	-2,213	1,419	3,054	37,471	46	0,811	Item Fit
Item 10	-2,932	0,179	2,505	59,946	42	0,036	Item Tidak Fit
Item 11	-3,165	-0,164	2,254	62,827	43	0,026	Item Tidak Fit
Item 12	-3,281	-0,336	2,727	67,770	43	0,009	Item Tidak Fit
Item 13	-3,595	0,415	3,115	54,327	42	0,096	Item Fit
Item 14	-2,467	-0,108	2,220	43,964	44	0,473	Item Fit
Item 15	-0,043	2,497	3,524	63,932	48	0,062	Item Fit
Item 16	1,294	2,810	3,944	49,105	41	0,180	Item Fit
Item 17	-3,474	-0,506	1,896	44,208	41	0,338	Item Fit
Item 18	0,077	1,531	4,045	57,484	48	0,164	Item Fit
Item 19	-4,035	0,025	2,751	66,862	42	0,009	Item Tidak Fit
Item 20	-4,977	-0,626	2,005	58,640	40	0,029	Item Tidak Fit
Item 21	-2,902	-0,513	1,524	39,377	38	0,408	Item Fit
Item 22	-4,678	-2,248	0,696	22,419	29	0,803	Item Fit
Item 23	-4,555	-0,232	2,434	58,867	42	0,044	Item Tidak Fit
Item 24	-3,315	0,225	2,833	57,322	42	0,058	Item Fit
Item 25	-4,016	-2,638	-0,087	26,600	25	0,376	Item Fit
Item 26	-4,052	-0,779	1,416	30,255	37	0,776	Item Fit
Item 27	-0,617	1,234	2,857	81,466	51	0,004	Item Tidak Fit
Item 28	-2,053	0,910	3,083	62,099	48	0,083	Item Fit
Item 29	-2,251	0,400	3,138	63,937	49	0,074	Item Fit
Item 30	-3,696	-0,728	1,815	43,478	40	0,325	Item Fit

menunjukkan nilai yang berurutan (dari kecil sampai besar) dengan posisi menyebar sepanjang rentang ability (*trait range*). Hal ini menunjukkan bahwa, paling tidak terdapat satu posisi tingkat kesulitan *(trait level)* tertentu di mana setiap pilihan jawaban lebih disukai. Sebagai ilustrasi dari *threshold* yang berurutan tersebut, kurva probabilitas kategori (*category probability curve*) untuk Item 7 dapat dilihat pada Gambar 7.

Threshold ditunjukkan oleh nilai yang terletak pada sumbu axis (*ability*) sejajar dengan titik potong antara dua kurva kategori respon. Pada Gambar 7, terlihat ketiga *threshold* berurutan dari yang kecil sampai besar. Kurva untuk kategori "hampir tidak pernah" (kurva 1) menunjukkan, siswa yang relatif memiliki penilaian negatif dalam memandang kegiatan penilaian (*assessment*) di kelas, memiliki kemungkinan (*probability*) yang relatif lebih



Gambar 7. ***Category Probability Curve*** **Item 7 -** *Partial Credit Model*

tinggi untuk memilih kategori "hampir tidak pernah" (*option* 1) dibandingkan memilih kategori "kadang-kadang" (*option* 2) atau pilihan lainnya pada item Nomor 7.

Pada kolom-kolom bagian akhir dari Tabel 8, sebagai bagian dari output PARSCALE, terdapat nilai statistik yang menunjukkan item fit, yaitu *likelihood-ratiochi-square*. Statistik tersebut menunjukkan item 10, item 11, item 12, item 19, item 20, item 23 dan item 27 tidak fit dengan estimasi parameter item model PCM (nilai *probability* lebih kecil dari 0,05). Selain itu, dengan melihat nilai *probability* total sebesar 0,000 (lebih kecil dari 0,05), dapat dijadikan salah satu indikasi bahwa model tersebut tidak fit dengan data.

***Generalized Partial Credit Model* (GPCM).** Hasil estimasi parameter item (*estimated item parameters*) untuk ke-30 item skala *Student Perceptions of Assessment Questionnaire* (SPAQ) dengan menggunakan PARSCALE for Windows versi 4.1 (Muraki, 2003) ditunjukkan Tabel 9.

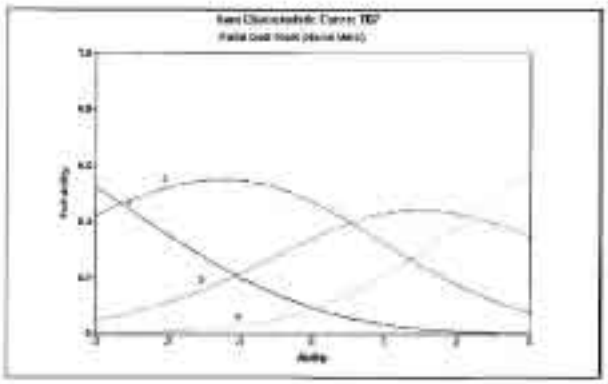
Tingkat kesulitan (*threshold*) untuk setiap kategori masing-masing item, seluruhnya menunjukkan nilai yang berurutan (dari kecil sampai besar) dengan posisi menyebar sepanjang rentang ability (*trait range*). Hasil ini tidak jauh berbeda dengan model PCM, di mana *threshold* setiap kategori untuk keseluruhan item berurutan. Sebagai ilustrasi dari *threshold* yang berurutan tersebut, kurva probabilitas kategori (*category probability curve*) untuk item 7 dapat dilihat pada Gambar 8.

Pada Gambar 8, terlihat ketiga *threshold* berurutan dari yang kecil sampai besar. Kurva untuk kategori "hampir selalu" (kurva 4) menunjukkan, siswa yang relatif memiliki penilaian positif dalam memandang kegiatan penilaian (*assessment*) di kelas, memiliki kemungkinan (*probability*) yang relatif lebih tinggi untuk memilih kategori "hampir selalu" (*option* 4) dibandingkan memilih kategori

Tabel 9. *Estimated Item Parameters* dan *Item-Fit Statistics* untuk *GeneralizedPartial Credit Model*

Item	a	b1	b2	b3	Chi-Square	df	Prob.	Keterangan
Item 1	0,525	-3,942	-0,609	1,214	39,118	41	0,555	Item Fit
Item 2	0,403	-3,186	-0,529	2,030	39,916	43	0,606	Item Fit
Item 3	0,335	-3,193	-1,206	1,885	46,178	44	0,382	Item Fit
Item 4	0,450	-2,939	-0,128	2,806	51,762	42	0,144	Item Fit
Item 5	0,404	-3,796	-0,300	2,604	49,997	43	0,215	Item Fit
Item 6	0,540	-3,338	-1,429	1,078	31,319	38	0,770	Item Fit
Item 7	0,357	-2,641	0,524	2,180	39,204	44	0,677	Item Fit
Item 8	0,482	-2,365	0,120	1,761	41,122	42	0,509	Item Fit
Item 9	0,416	-2,115	1,350	2,905	51,995	51	0,435	Item Fit
Item 10	0,576	-2,264	0,118	1,959	41,960	41	0,429	Item Fit
Item 11	0,565	-2,475	-0,144	1,778	56,524	42	0,066	Item Fit
Item 12	0,675	-2,323	-0,239	1,918	24,503	42	0,986	Item Fit
Item 13	0,791	-2,267	0,229	2,048	32,268	39	0,769	Item Fit
Item 14	0,494	-2,108	-0,101	1,894	50,545	44	0,231	Item Fit
Item 15	0,300	-0,003	3,054	4,082	46,038	46	0,471	Item Fit
Item 16	0,285	1,773	3,503	4,670	59,865	46	0,082	Item Fit
Item 17	0,521	-2,852	-0,429	1,565	56,810	42	0,063	Item Fit
Item 18	0,264	0,230	1,980	5,329	55,536	54	0,417	Item Fit
Item 19	0,775	-2,570	-0,019	1,823	46,358	41	0,261	Item Fit
Item 20	0,791	-3,114	-0,443	1,326	39,407	39	0,452	Item Fit
Item 21	0,429	-2,700	-0,486	1,412	46,815	43	0,319	Item Fit
Item 22	0,387	-4,663	-2,265	0,683	22,334	34	0,938	Item Fit
Item 23	0,667	-3,140	-0,196	1,740	37,331	41	0,635	Item Fit
Item 24	0,563	-2,572	0,156	2,226	51,856	41	0,119	Item Fit
Item 25	0,428	-3,726	-2,466	-0,092	31,180	26	0,221	Item Fit
Item 26	0,402	-3,946	-0,761	1,365	42,976	42	0,429	Item Fit
Item 27	0,173	-0,978	2,358	5,163	64,127	59	0,301	Item Fit
Item 28	0,390	-2,048	0,904	3,061	49,555	51	0,531	Item Fit
Item 29	0,419	-2,132	0,370	2,973	52,021	46	0,251	Item Fit
Item 30	0,455	-3,307	-0,661	1,623	33,341	42	0,828	Item Fit
Total					**1331,964**	**1289**	**0,198**	**Model Fit**

Catatan: a = slope, b1 – b3 = step parameter (threshold), df = degree of freedom, Prob. = probability



Gambar 8. *Category Probability Curve* Item 7 – *Generalized Partial Credit Model*

"sering" (*option* 3) atau pilihan lainnya pada item Nomor 7.

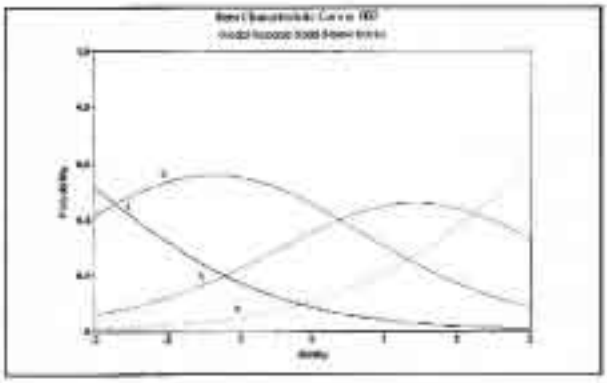
Pada kolom-kolom bagian akhir dari Tabel 9, sebagai bagian dari *output* PARSCALE, terdapat nilai statistik yang menunjukkan item *fit*, yaitu *likelihood-ratiochi-square*. Statistik tersebut menunjukkan seluruh item *fit* dengan estimasi parameter item model GPCM (nilai *probability* lebih besar dari 0,05). Selain itu, dengan melihat nilai *probability* total sebesar 0,198 (lebih besar dari 0,05), dapat dijadikan salah satu indikasi bahwa model tersebut fit dengan data.

***Graded Response Model* (GRM).** Tabel 10 menunjukkan hasil estimasi parameter item (*estimated item parameters*) untuk ke-30 item skala *Student Perceptions of Assessment Questionnaire* (SPAQ) dengan menggunakan PARSCALE for Windows versi 4.1 (Muraki, 2003).

Tingkat kesulitan (*threshold*) model GRM untuk setiap kategori masing-masing item dapat dipastikan seluruhnya menunjukkan nilai yang berurutan (dari kecil sampai besar) dengan posisi menyebar sepanjang rentang *ability* (*trait range*). Hal ini menunjukkan bahwa, paling tidak terdapat satu posisi tingkat kesulitan *(trait level)* tertentu di mana setiap pilihan jawaban lebih disukai. Sebagai ilustrasi dari *threshold* yang berurutan tersebut, kurva probabilitas kategori (*category probability*

Tabel 10. *Estimated Item Parameters* dan *Item-Fit Statistics* untuk *Graded Response Model*

Item	a	b1	b2	b3	Chi-Square	df	Prob.	Keterangan
Item 1	0,684	-4,029	-0,879	1,230	41,103	40	0,422	Item Fit
Item 2	0,534	-3,584	-0,714	2,019	55,024	42	0,086	Item Fit
Item 3	0,471	-3,917	-1,204	1,805	51,394	42	0,152	Item Fit
Item 4	0,571	-3,240	-0,281	2,750	50,518	44	0,231	Item Fit
Item 5	0,512	-4,056	-0,497	2,606	49,227	42	0,206	Item Fit
Item 6	0,669	-4,011	-1,552	0,998	42,617	37	0,242	Item Fit
Item 7	0,476	-2,924	0,191	2,624	43,558	43	0,448	Item Fit
Item 8	0,638	-2,607	-0,128	1,909	35,653	42	0,745	Item Fit
Item 9	0,536	-2,232	0,987	3,459	69,145	47	0,019	Item Tidak Fit
Item 10	0,710	-2,525	-0,095	2,038	48,616	41	0,193	Item Fit
Item 11	0,712	-2,731	-0,344	1,800	56,408	43	0,082	Item Fit
Item 12	0,816	-2,601	-0,391	1,845	25,707	41	0,970	Item Fit
Item 13	0,936	-2,409	0,047	2,017	37,999	41	0,605	Item Fit
Item 14	0,670	-2,411	-0,237	1,900	59,504	43	0,048	Item Tidak Fit
Item 15	0,392	-0,504	2,985	6,058	62,192	47	0,068	Item Fit
Item 16	0,192	-2,591	0,367	1,401	599,771	43	0,000	Item Tidak Fit
Item 17	0,674	-3,163	-0,635	1,568	52,156	41	0,114	Item Fit
Item 18	0,334	-0,694	2,467	6,877	52,429	59	0,715	Item Fit
Item 19	0,907	-2,737	-0,192	1,812	52,870	40	0,084	Item Fit
Item 20	0,945	-3,257	-0,631	1,260	54,457	37	0,032	Item Tidak Fit
Item 21	0,608	-3,030	-0,676	1,463	39,988	42	0,560	Item Fit
Item 22	0,494	-5,863	-2,456	0,678	19,794	29	0,899	Item Fit
Item 23	0,806	-3,271	-0,413	1,712	53,736	40	0,072	Item Fit
Item 24	0,708	-2,734	-0,065	2,240	46,777	41	0,247	Item Fit
Item 25	0,547	-5,227	-2,698	-0,007	22,422	24	0,554	Item Fit
Item 26	0,550	-4,171	-1,022	1,413	66,276	41	0,008	Item Tidak Fit
Item 27	0,240	-2,054	2,390	7,084	58,068	63	0,652	Item Fit
Item 28	0,528	-2,267	0,630	3,318	34,756	47	0,907	Item Fit
Item 29	0,558	-2,409	0,229	2,994	49,666	43	0,225	Item Fit
Item 30	0,618	-3,572	-0,816	1,575	46,108	41	0,269	Item Fit
Total					**1977.938**	**1266**	**0.000**	**Model Tidak Fit**



Gambar 9. *Category Probability Curve* Item 7 – *Graded Response Model*

curve) untuk item 7 dapat dilihat pada Gambar 9.

Pada kolom-kolom bagian akhir dari Tabel 10, sebagai bagian dari output PARSCALE, terdapat nilai statistik yang menunjukkan item *fit*, yaitu *likelihood-ratiochi-square*. Statistik tersebut menunjukkan item 9, item 14, item 16, item 20 dan item 26 tidak fit dengan estimasi parameter item model GRM (nilai *probability* lebih kecil dari 0,05). Selain itu, dengan melihat nilai *probability* total sebesar 0,000 (lebih kecil dari 0,05), dapat dijadikan salah satu indikasi bahwa model tersebut tidak fit dengan data.

Sub Skala SPAQ. Selain melakukan pengujian model fit secara keseluruhan untuk 30 item (simultan), peneliti juga melakukan pengujian model fit untuk kelima sub skala SPAQ (*Congruence with planned learning, Authenticity, Student Consultation, Transparency* dan *Diversity*) dengan mengkalibrasinya secara terpisah.

Hasil kalibrasi secara terpisah untuk masing-masing sub skala SPAQ sangat jauh berbeda dengan hasil kalibrasi secara simultan untuk ke-30 item SPAQ. Pada model PCM, hanya 10 item yang fit dengan model, dengan rincian sebagai berikut:

1. Sub skala *Congruence with planned learning*, hanya item 2 yang fit.
2. Sub skala *Authenticity*, item yang fit adalah: item 7, item 9 dan item 11.
3. Sub skala *Student consultation*, hanya item 17 yang fit.

Tabel 11. Statistik Item-Fit Sub Skala SPAQ

Sub Skala SPAQ	Item	MODEL POLYTOMOUS IRT					
		PCM		GPCM		GRM	
		Prob	Item Fit	Prob	Item Fit	Prob	Item Fit
Congruence With Planned Learning	Item 1	0,005	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 2	0,105	Fit	0,000	Not Fit	0,000	Not Fit
	Item 3	0,006	Not Fit	0,005	Not Fit	0,000	Not Fit
	Item 4	0,004	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 5	0,004	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 6	0,022	Not Fit	0,000	Not Fit	0,001	Not Fit
Authenticity	Item 7	0,523	Fit	0,000	Not Fit	0,000	Not Fit
	Item 8	0,002	Not Fit	0,038	Not Fit	0,000	Not Fit
	Item 9	0,102	Fit	0,005	Not Fit	0,000	Not Fit
	Item 10	0,041	Not Fit	0,001	Not Fit	0,009	Not Fit
	Item 11	0,926	Fit	0,186	Fit	0,002	Not Fit
	Item 12	0,007	Not Fit	0,155	Fit	0,004	Not Fit
Student Consultation	Item 13	0,000	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 14	0,007	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 15	0,044	Not Fit	0,015	Not Fit	0,000	Not Fit
	Item 16	0,004	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 17	0,476	Fit	0,000	Not Fit	0,001	Not Fit
	Item 18	0,006	Not Fit	0,000	Not Fit	0,000	Not Fit
Transparency	Item 19	0,300	Fit	0,017	Not Fit	0,000	Not Fit
	Item 20	0,022	Not Fit	0,001	Not Fit	0,000	Not Fit
	Item 21	0,067	Fit	0,000	Not Fit	0,000	Not Fit
	Item 22	0,985	Fit	0,000	Not Fit	0,002	Not Fit
	Item 23	0,464	Fit	0,000	Not Fit	0,001	Not Fit
	Item 24	0,611	Fit	0,016	Not Fit	0,002	Not Fit
Diversity	Item 25	0,000	Not Fit	0,001	Not Fit	0,000	Not Fit
	Item 26	0,000	Not Fit	0,004	Not Fit	0,002	Not Fit
	Item 27	0,000	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 28	0,000	Not Fit	0,001	Not Fit	0,002	Not Fit
	Item 29	0,000	Not Fit	0,000	Not Fit	0,000	Not Fit
	Item 30	0,000	Not Fit	0,000	Not Fit	0,000	Not Fit

4. Sub skala *Transparency*, item yang fit adalah: item 19, item 21, item 22, item 23 dan item 24.
5. Sub skala *Diversity*, keseluruhan item tidak fit.

Untuk model GPCM, hasil estimasi dengan program PARSCALE menunjukkan hanya item 11 dan item 12 yang dapat fit dengan model. Sedangkan untuk model GRM, keseluruhan item tidak fit.

***Analisis Differential Item Functioning* (DIF).** Untuk mendeteksi DIF pada skala SPAQ, peneliti menggunakan Metode DIF untuk Model Rasch dengan bantuan program QUEST versi 2.1 (Adams & Khoo, 1996). Dalam metode ini, suatu item dikatakan terdeteksi DIF apabila item tersebut memiliki tingkat kesukaran yang berbeda secara signifikan antara kelompok yang diperbandingkan. Statistik yang digunakan dalam pengujian ini adalah *chi-square* dengan hipotesis sebagai berikut:

H_0: Tidak terdapat perbedaan tingkat kesukaran antara kelompok yang diperbandingkan.

H_1: Terdapat perbedaan tingkat kesukaran antara kelompok yang diperbandingkan.

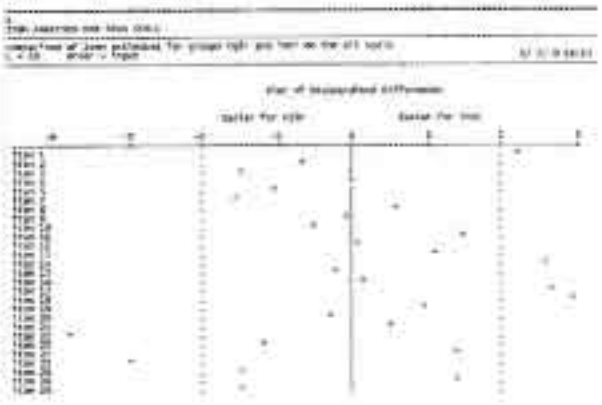
Pembagian kelompok yang dibandingkan berdasarkan 3 karakteristik responden, yaitu: jenis kelamin (laki-laki dan perempuan), program kelas (kelas reguler dan kelas internasional), serta status sekolah (sekolah negeri dan sekolah swasta).

DIF Berdasarkan Jenis Kelamin. Perbandingan estimasi tingkat kesukaran untuk kelompok responden laki-laki dan perempuan disajikan dalam tabel 12

Statistik *chi-square* pada Tabel 12 menunjukkan terdapat 3 item yang terdeteksi DIF, yaitu: item 21, 24 dan 26. Ketiga item tersebut memiliki nilai p-value kurang dari 0,05 (signifikan), sehingga hipotesis yang menyatakan tidak terdapat perbedaan tingkat kesukaran antara kelompok laki-laki dan perempuan ditolak. Dengan kata lain, ketiga item tersebut memiliki tingkat kesukaran yang berbeda untuk kelompok laki-laki dan perempuan. Untuk

Tabel 12. Perbandingan Estimasi Item Berdasarkan Jenis Kelamin

Nomor Item	Delta		Adjusted Delta		Difference		Chi-Sq	p-value	Ket.
	Laki-laki	Perempuan	Laki-laki	Perempuan	L - P	(L - P) Std'ized			
1	-0,83	-0,74	-0,83	-0,74	-0,10	-0,69	0,48	0,49	*Not DIF*
2	-0,21	-0,22	-0,21	-0,22	0,01	0,07	0,01	0,94	*Not DIF*
3	-0,34	-0,35	-0,34	-0,35	0,01	0,06	0,00	0,95	*Not DIF*
4	-0,07	0,17	-0,07	0,17	-0,24	-1,63	2,65	0,10	*Not DIF*
5	-0,21	-0,15	-0,21	-0,15	-0,6	-0,40	0,16	0,69	*Not DIF*
6	-0,93	-0,85	-0,93	-0,85	-0,08	-0,55	0,30	0,58	*Not DIF*
7	0,33	0,17	0,33	0,17	0,16	1,25	1,56	0,21	*Not DIF*
8	0,14	0,02	0,14	0,02	0,12	0,94	0,88	0,35	*Not DIF*
9	0,77	0,69	0,77	0,70	0,07	0,51	0,26	0,61	*Not DIF*
10	0,22	0,10	0,22	0,10	0,12	0,88	0,78	0,38	*Not DIF*
11	-0,08	-0,06	-0,08	-0,06	-0,01	-0,10	0,01	0,92	*Not DIF*
12	0,07	-0,07	0,07	-0,07	0,14	1,02	1,05	0,31	*Not DIF*
13	0,34	0,12	0,34	0,12	0,22	1,54	2,36	0,12	*Not DIF*
14	0,04	0,12	0,04	0,12	-0,08	-0,64	0,41	0,52	*Not DIF*
15	1,55	1,59	1,55	1,59	-0,04	-0,29	0,08	0,77	*Not DIF*
16	2,21	2,04	2,22	2,04	0,17	1,05	1,11	0,29	*Not DIF*
17	-0,49	-0,24	-0,49	-0,24	-0,25	-1,82	3,31	0,07	*Not DIF*
18	1,62	1,47	1,62	1,47	0,16	1,18	1,39	0,24	*Not DIF*
19	0,04	-0,14	0,04	-0,14	0,18	1,22	1,49	0,22	*Not DIF*
20	-0,51	-0,71	-0,51	-0,71	0,19	1,38	1,90	0,17	*Not DIF*
21	-0,57	-0,17	-0,57	-0,16	-0,40	-3,13	9,80	0,00	*DIF*
22	-1,28	-1,28	-1,28	-1,28	0,00	0,02	0,00	0,98	*Not DIF*
23	-0,32	-0,38	-0,32	-0,38	0,07	0,48	0,23	0,63	*Not DIF*
24	-0,14	0,20	-0,14	0,20	-0,34	-2,37	5,61	0,02	*DIF*
25	-1,35	-1,41	-1,34	-1,41	0,07	0,46	0,21	0,64	*Not DIF*
26	-0,92	-0,52	-0,92	-0,52	-0,40	-2,88	8,28	0,00	*DIF*
28	0,83	0,61	0,83	0,61	0,21	1,50	2,26	0,13	*Not DIF*
29	0,47	0,47	0,47	0,47	0,00	-0,01	0,00	0,99	*Not DIF*
30	-0,38	-0,47	-0,38	-0,47	0,09	0,72	0,52	0,47	*Not DIF*



Gambar 10. Plot Perbedaan Estimasi Item Berdasarkan Jenis Kelamin

mendapatkan gambaran lebih jelas mengenai item-item pada skala SPAQ yang terdeteksi DIF berdasarkan Jenis Kelamin, Gambar 10 menampilkan plot perbedaan tingkat kesukaran antara kedua kelompok tersebut.

Berdasarkan gambar 10, dapat diketahui secara jelas bahwa item 21, 24 dan 26 memiliki tingkat kesukaran yang lebih rendah untuk kelompok laki-laki, atau dengan kata lain item-item tersebut lebih mudah bagi kelompok laki-laki. Perbedaan tingkat kesukaran di sini dapat diartikan perbedaan antara kelompok laki-laki dengan perempuan dalam mempersepsikan penilaian pelajaran IPA di kelas, di mana persepsi seseorang dapat dipengaruhi oleh faktor psikologis (seperti emosi dan pengalaman masa lalu).

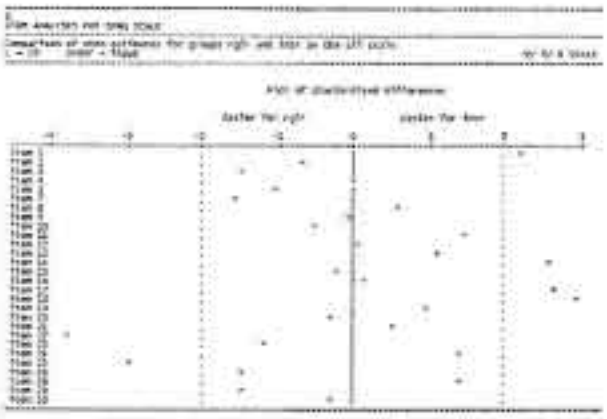
DIF Berdasarkan Program Kelas. Tabel 13 menunjukkan perbandingan estimasi item skala SPAQ berdasarkan perbedaan program kelas yang dipilih para siswa. Program kelas yang dimaksud di sini adalah program kelas reguler dan program kelas internasional.

Statistik *chi-square* pada Tabel 13 menunjukkan terdapat 6 item yang terdeteksi DIF, yaitu: Item 1, 14, 17, 18, 22 dan 25. Keenam item tersebut memiliki nilai p-value kurang dari 0,05 (signifikan), sehingga dapat disimpulkan 6 item tersebut memiliki tingkat kesukaran yang berbeda untuk siswa di program kelas reguler dan program kelas internasional. Gambar 11 menampilkan plot perbedaan tingkat kesukaran antara kedua kelompok tersebut.

Berdasarkan gambar 11, dapat diketahui secara jelas bahwa Item 1, 14, 17, 18 lebih mudah bagi siswa program kelas internasional, sedangkan Item

Tabel 13. Perbandingan Estimasi Item Berdasarkan Program Kelas

Nomor Item	Delta		Adjusted Delta		Difference		Chi-Sq	p-value	Ket.
	Laki-laki	Perempuan	Laki-laki	Perempuan	L - P	(L - P) Std'ized			
1	-0,77	-0,80	-0,82	-0,81	-0,01	-0,08	0,01	0,94	*Not DIF*
2	-0,28	-0,17	-0,32	-0,18	-0,15	-1,26	1,58	0,21	*Not DIF*
3	-0,32	-0,43	-0,36	-0,44	0,08	0,70	0,49	0,49	*Not DIF*
4	0,01	0,21	-0,04	0,20	-0,24	-1,97	3,86	0,05	*DIF*
5	-0,18	-0,18	-0,22	-0,19	-0,04	-0,29	0,08	0,77	*Not DIF*
6	-1,10	-0,68	-1,14	-0,69	-0,45	-3,70	13,70	0,00	*DIF*
7	0,45	-0,12	0,41	-0,13	0,54	4,86	23,64	0,00	*DIF*
8	0,21	-0,24	0,17	-0,24	0,41	3,70	13,72	0,00	*DIF*
9	0,79	0,57	0,75	0,56	0,19	1,54	2,37	0,12	*Not DIF*
10	0,18	-0,02	0,13	-0,02	0,16	1,33	1,76	0,18	*Not DIF*
11	-0,04	-0,17	-0,08	-0,17	0,10	0,82	0,67	0,41	*Not DIF*
12	-0,03	-0,06	-0,08	-0,06	-0,01	-0,10	0,01	0,92	*Not DIF*
13	0,13	0,19	0,09	0,18	-0,09	-0,70	0,50	0,48	*Not DIF*
14	0,00	0,18	-0,04	0,18	-0,21	-1,95	3,81	0,05	*DIF*
15	1,61	1,52	1,57	1,51	0,06	0,50	0,25	0,62	*Not DIF*
16	2,32	1,82	2,27	1,81	0,46	3,38	11,46	0,00	*DIF*
17	-0,49	-0,09	-0,53	-0,10	-0,43	-3,75	14,07	0,00	*DIF*
18	1,41	1,60	1,37	1,59	-0,22	-1,95	3,81	0,05	*DIF*
19	-0,13	-0,10	-0,17	-0,11	-0,06	-0,49	0,24	0,63	*Not DIF*
20	-0,62	-0,72	-0,66	-0,73	0,07	0,54	0,30	0,59	*Not DIF*
21	-0,47	0,00	-0,51	0,00	-0,51	-4,65	21,61	0,00	*DIF*
23	-0,38	-0,37	-0,42	-0,38	-0,04	-0,35	0,12	0,73	*Not DIF*
24	-0,07	0,32	-0,12	0,32	-0,43	-3,57	12,75	0,00	*DIF*
25	-1,30	-1,76	-1,34	-1,77	0,43	3,33	11,12	0,00	*DIF*
26	-0,64	-0,60	-0,68	-0,60	-0,08	-0,71	0,50	0,48	*Not DIF*
28	0,66	0,58	0,62	0,57	0,04	0,39	0,15	0,70	*Not DIF*
29	0,58	0,28	0,54	0,28	0,26	2,21	4,89	0,03	*DIF*
30	-0,35	-0,59	-0,40	-0,60	0,20	1,72	2,95	0,09	*Not DIF*



Gambar 11. Plot Perbedaan Estimasi Item Berdasarkan Program Kelas

22 dan 25 lebih mudah bagi siswa program kelas reguler.

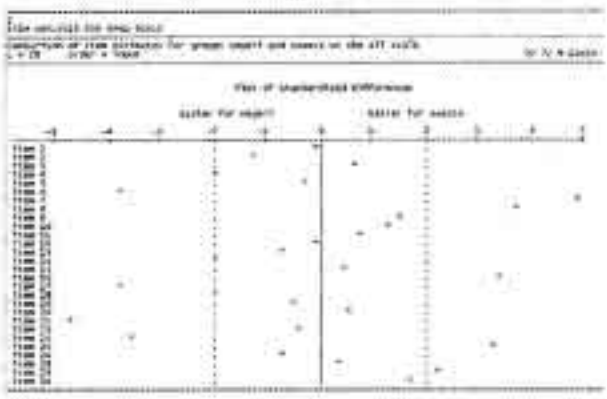
Perbedaan persepsi terhadap penilaian pelajaran IPA di kelas untuk siswa program kelas reguler dan program kelas internasional dapat disebabkan berbagai hal, salah satunya karena perbedaan kurikulum yang diajarkan kepada siswa tersebut. Program kelas reguler mengacu pada kurikulum nasional Indonesia (KTSP), sedangkan program kelas internasional mengacu pada kurikulum Cambridge IGCSE dan A/AS Level. Selain itu, tenaga pengajar program kelas reguler yang merupakan guru tetap di sekolah, mayoritas lulusan Institut Keguruan (IKIP), sedangkan tenaga pengajar program kelas internasional lebih banyak yang berasal dari luar sekolah (*out source*), dan mayoritas latar belakang pendidikan mereka adalah ilmu murni (MIPA) di Univeristas Negeri non keguruan, sehingga sangat dimungkinkan terdapat perbedaan cara mengajar dan juga termasuk dalam memberikan penilaian di kelas.

DIF Berdasarkan Status Sekolah. Pembagian kelompok berdasarkan status sekolah (negeri dan swasta) mendeteksi cukup banyak item-item skala SPAQ yang mengandung DIF. Tabel 14 menunjukkan secara rinci item-item mana yang terdeteksi DIF tersebut.

Statistik *chi-square* pada Tabel 14 menunjukkan terdapat 12 item yang terdeteksi DIF, yaitu: item 4, 6, 7, 8, 14, 16, 17, 18, 21, 24, 25 dan 29. Kedua belas item tersebut memilki nilai p-value kurang dari 0,05 (signifikan), sehingga dapat disimpulkan 12 item tersebut memiliki tingkat kesukaran yang berbeda untuk siswa di sekolah negeri dan swasta. Gambar 12 menampilkan plot perbedaan tingkat

Tabel 14. Perbandingan Estimasi Item Berdasarkan Program Kelas

Nomor Item	Delta		Adjusted Delta		Difference		Chi-Sq	p-value	Ket.
	Laki-laki	Perempuan	Laki-laki	Perempuan	L - P	(L - P) Std'ized			
1	-0,77	-0,80	-0,82	-0,81	-0,01	-0,08	0,01	0,94	*Not DIF*
2	-0,28	-0,17	-0,32	-0,18	-0,15	-1,26	1,58	0,21	*Not DIF*
3	-0,32	-0,43	-0,36	-0,44	0,08	0,70	0,49	0,49	*Not DIF*
4	0,01	0,21	-0,04	0,20	-0,24	-1,97	3,86	0,05	*DIF*
5	-0,18	-0,18	-0,22	-0,19	-0,04	-0,29	0,08	0,77	*Not DIF*
6	-1,10	-0,68	-1,14	-0,69	-0,45	-3,70	13,70	0,00	*DIF*
7	0,45	-0,12	0,41	-0,13	0,54	4,86	23,64	0,00	*DIF*
8	0,21	-0,24	0,17	-0,24	0,41	3,70	13,72	0,00	*DIF*
9	0,79	0,57	0,75	0,56	0,19	1,54	2,37	0,12	*Not DIF*
10	0,18	-0,02	0,13	-0,02	0,16	1,33	1,76	0,18	*Not DIF*
11	-0,04	-0,17	-0,08	-0,17	0,10	0,82	0,67	0,41	*Not DIF*
12	-0,03	-0,06	-0,08	-0,06	-0,01	-0,10	0,01	0,92	*Not DIF*
13	0,13	0,19	0,09	0,18	-0,09	-0,70	0,50	0,48	*Not DIF*
14	0,00	0,18	-0,04	0,18	-0,21	-1,95	3,81	0,05	*DIF*
15	1,61	1,52	1,57	1,51	0,06	0,50	0,25	0,62	*Not DIF*
16	2,32	1,82	2,27	1,81	0,46	3,38	11,46	0,00	*DIF*
17	-0,49	-0,09	-0,53	-0,10	-0,43	-3,75	14,07	0,00	*DIF*
18	1,41	1,60	1,37	1,59	-0,22	-1,95	3,81	0,05	*DIF*
19	-0,13	-0,10	-0,17	-0,11	-0,06	-0,49	0,24	0,63	*Not DIF*
20	-0,62	-0,72	-0,66	-0,73	0,07	0,54	0,30	0,59	*Not DIF*
21	-0,47	0,00	-0,51	0,00	-0,51	-4,65	21,61	0,00	*DIF*
23	-0,38	-0,37	-0,42	-0,38	-0,04	-0,35	0,12	0,73	*Not DIF*
24	-0,07	0,32	-0,12	0,32	-0,43	-3,57	12,75	0,00	*DIF*
25	-1,30	-1,76	-1,34	-1,77	0,43	3,33	11,12	0,00	*DIF*
26	-0,64	-0,60	-0,68	-0,60	-0,08	-0,71	0,50	0,48	*Not DIF*
28	0,66	0,58	0,62	0,57	0,04	0,39	0,15	0,70	*Not DIF*
29	0,58	0,28	0,54	0,28	0,26	2,21	4,89	0,03	*DIF*
30	-0,35	-0,59	-0,40	-0,60	0,20	1,72	2,95	0,09	*Not DIF*



Gambar 12. Plot Perbedaan Estimasi Item Berdasarkan Status Sekolah

kesukaran antara kedua kelompok tersebut.

Gambar 12 menunjukkan bahwa item 4, 6, 14, 17, 18, 21 dan 24 lebih mudah bagi kelompok siswa di sekolah negeri. Sedangkan item 7, 8, 16, 25 dan 29 lebih mudah bagi kelompok siswa di sekolah swasta.

Pengujian terhadap DIF memberikan gambaran bahwa item-item skala SPAQ masih banyak yang mengandung DIF, sehingga responden dengan tingkat persepsi yang sama terhadap penilaian di kelas akan memberikan respon yang berbeda-beda terhadap item-item tersebut.

Dari 30 item skala SPAQ, terdapat 6 item yang terdeteksi DIF di 2 kategori kelompok yang berbeda. Untuk pengembangan skala ini, diharapkan item-item tersebut dapat direvisi atau dikeluarkan dari pengujian dengan menggunakan skala SPAQ. Item-item tersebut antara lain:

1. Item 14, yang berbunyi: "Saya tahu bagaimana ujian saya akan dinilai (*I am aware how my assessment will be marked*)". Item ini terdeteksi DIF pada pengujian berdasarkan Program Kelas dan Status Sekolah.
2. Item 17, yang berbunyi: "Guru saya telah menjelaskan kepada saya bagaimana masing-masing jenis penilaian akan digunakan (*My teacher has explained to me how each type of assessment is to be used*)". Item ini terdeteksi DIF pada pengujian berdasarkan Program Kelas dan Status Sekolah.
3. Item 18, yang berbunyi: "Saya dapat ikut menentukan bagaimana saya akan dinilai dalam pelajaran IPA (*I can have a say in how I will be assessed in science*)". Item ini terdeteksi DIF pada pengujian berdasarkan Program Kelas dan Status Sekolah.
4. Item 21, yang berbunyi: "Guru memberi tahu saya sebelumnya kapan saya akan dinilai (*I am told in advance when I am being assessed*)". Item ini terdeteksi DIF pada pengujian berdasarkan Jenis Kelamin dan Status Sekolah.
5. Item 24, yang berbunyi: "Saya tahu bagaimana tugas tertentu akan dinilai (*I know how a particular assessment tasks will be marked*)". Item ini terdeteksi DIF pada pengujian berdasarkan

Tabel 15. Item-Item Skala SPAQ yang Terdeteksi DIF Berdasarkan Ketiga Karakteristik Responden (Jenis Kelamin, Program Kelas dan Status Sekolah)

Nomer Item	Item Terdeteksi DIF (√)		
	Jenis Kelamin	Program Kelas	Status Sekolah
1		√	
2			
3			
4			√
5			
6			√
7			√
8			√
9			
10			
11			
12			
13			
14		√	√
15			
16			√
17		√	√
18		√	√
19			
20			
21	√		√
22		√	
23			
24	√		√
25		√	√
26	√		
27			
28			
29			√
30			

Jenis Kelamin dan Status Sekolah.

6. Item 25, yang berbunyi: "Saya memiliki kesempatan yang sama seperti siswa-siswi lain dalam menyelesaikan tugas (*I have as much chance as any other student at completing assessment tasks*)". Item ini terdeteksi DIF pada pengujian berdasarkan Program Kelas dan Status Sekolah.

SIMPULAN

Keunggulan utama dari penerapan model IRT adalah apa yang disebut sifat *invariance property*, di mana estimasi kemampuan peserta tes tidak tergantung pada tes yang diberikan, dan estimasi parameter item tidak bergantung pada kelompok peserta tes. Sifat tersebut hanya dapat terpenuhi jika asumsi-asumsi dari model IRT juga terpenuhi. Untuk menguji asumsi-asumsi tersebut, peneliti melakukan pengujian terhadap dua asumsi utama dari model IRT yaitu *unidimensionality* dan model fit.

Asumsi pertama dari model IRT adalah *unidimensionality*. Pengujian terhadap asumsi ini dilakukan dengan menggunakan *Confirmatory Factor Analysis* (CFA) secara bertahap. Tahap pertama CFA dilakukan terhadap 30 item skala SPAQ yang dijadikan sebagai indikator-indikator yang dapat diukur secara langsung (*first order factor*) dan secara tidak langsung (*second order factor*). Hasilnya, kedua model pengukuran tersebut memenuhi dua dari ketiga kriteria fit yang digunakan. Sehingga dapat disimpulkan model tersebut tidak benar-benar fit dengan data. Namun setelah dilakukan beberapa kali modifikasi terhadap model, yaitu dengan mengestimasi korelasi antar kesalahan pengukuran (*error*), diperoleh dua model pengukuran yang fit (*first order factor* dan *second order factor*) untuk skala SPAQ.

Tahap kedua dari CFA dilakukan terhadap model pengukuran, di mana 5 sub skala SPAQ dijadikan sebagai indikator yang mengukur secara langsung skala SPAQ. Setelah dilakukan pengujian, 3 kriteria fit dapat terpenuhi, sehingga dapat disimpulkan model pengukuran tersebut dapat fit dengan data.

Dari kedua tahapan pengujian asumsi *unidimensionality*, dapat diketahui bahwa asumsi ini tidak dapat terpenuhi secara ketat dikarenakan adanya faktor-faktor lain, seperti: motivasi, kepribadian, admistrasi tes, dan sebagainya. Namun yang terpenting adalah adanya satu komponen yang dianggap paling dominan dalam menentukan performansi peserta tes, yaitu persepsi siswa terhadap kegiatan penilaian di kelas. Sehingga dari hasil pengujian tersebut dapat disimpulkan bahwa item-item pada skala SPAQ dapat mewakili persepsi siswa terhadap penilaian di kelas. Atau dengan kata lain, skala SPAQ tersebut secara konstruk merupakan alat ukur yang valid.

Hasil pengujian model PCM menunjukkan statistik *likelihood-ratiochi-square* menunjukkan terdapat 7 item yang tidak fit dengan model (*probability* kurang dari 0.05), yaitu item 10, item 11, item 12, item 19, item 20, item 23 dan item 27. Probability total dari statistik *likelihood-ratiochi-square* juga mengindikasikan bahwa model tersebut tidak fit dengan data.

Hasil pengujian model GPCM dengan program PARSCALE menunjukkan dari 30 item skala SPAQ, seluruhnya fit dengan model. *Probability* total dari statistik *likelihood-ratiochi-square* juga mengindikasikan bahwa model tersebut fit dengan data. Tidak demikian pengujian dengan model GRM dengan program yang sama. Dari 30 item skala SPAQ, terdapat 5 item yang tidak fit dengan model, yaitu Item 9, item 14, item 16, item 20 dan item 26. Berdasarkan probability total dari statistik *likelihood-ratiochi-square* dapat diketahui bahwa model tersebut tidak fit dengan data.

Dari hasil pengujian kesesuaian model terhadap data, dapat disimpulkan bahwa model GPCM lebih sesuai (*fit*) untuk data yang diperoleh. Model tersebut dapat menjelaskan dengan baik item-item dalam sub skala SPAQ. Sehingga untuk analisa lebih lanjut, serta untuk evaluasi skala ini dapat menggunakan model GPCM.

Hasil estimasi parameter untuk ketiga model yang diuji menghasilkan tingkat kesulitan (*threshold*) untuk setiap kategori masing-masing item, seluruhnya menunjukkan nilai yang berurutan (dari kecil sampai besar) dengan posisi menyebar sepanjang rentang *ability* (*trait range*). Hal ini menunjukkan bahwa, paling tidak terdapat satu posisi tingkat kesulitan (*trait level*) tertentu di mana setiap pilihan jawaban lebih disukai.

Pengujian terhadap *Differential Item Functioning* (DIF) dilakukan dengan Metode DIF untuk Model Rasch dengan bantuan program QUEST versi 2.1 (Adams & Khoo, 1996). Dengan membedakan kelompok siswa berdasarkan Jenis Kelamin,

DIF dapat terdeteksi pada Item 21, 24 dan 26. Pembedaan kelompok siswa berdasarkan Program Kelas mendeteksi DIF pada Item 1, 14, 17, 18, 22 dan 25. Sedangkan pembedaan berdasarkan Status Sekolah mendeteksi DIF pada Item 4, 6, 7, 8, 14, 16, 17, 18, 21, 24, 25 dan 29.

Item-item yang terdeteksi DIF tersebut akan mempengaruhi hasil pengukuran persepsi siswa terhadap penilaian di kelas. Karena siswa dengan pandangan yang sama terhadap proses penilaian di kelas, namun berada pada kelompok yang berbeda, akan mempersepsikan secara berbeda proses penilaian tersebut. Atau dengan kata lain, siswa dengan *trait level* (*ability*) yang sama dalam kelompok yang berbeda, memiliki peluang (*probability*) yang berbeda untuk menjawab kategori tertentu ("hampir tidak pernah", "kadang-kadang", "sering", atau "sering sekali") pada item-item skala SPAQ yang teredeteksi DIF.

DISKUSI

Dari beberapa penelitian dapat diketahui bahwa skala *Student Perceptions of Assessment Questionnaire* (SPAQ) merupakan alat ukur yang cukup valid dan reliabel. Skala ini sangat bermanfaat untuk menilai kegiatan pembelajaran di kelas berdasarkan persepsi siswa, khususnya yang berkaitan dengan proses penilaian (*assessment*). Di Indonesia, penilaian kegiatan pembelajaran lebih banyak dilakukan oleh guru terhadap siswa, sangat jarang siswa dilibatkan dalam proses penilaian tersebut. Oleh karena itu, skala SPAQ akan sangat bermanfaat jika diterapkan, dikembangkan dan disesuaikan dengan budaya di Indonesia.

Untuk pengembangan dan penggunaan skala SPAQ, diharapkan temuan-temuan dari penelitian ini dapat dijadikan bahan evaluasi untuk merevisi item-item pada skala SPAQ, khususnya item-item yang terdeteksi DIF. Jika item-item yang terdeteksi tersebut tidak dikeluarkan atau direvisi sebelum dilakukan pengujian, maka hasil pengukuran tersebut dapat menjadi bias. Selain itu, item-item pada skala SPAQ diharapkan dapat disesuaikan dengan pelajaran-pelajaran non IPA, sehingga penggunaan skala SPAQ ini dapat diperluas tidak hanya spesifik untuk kelas IPA. Bahkan, jika diinginkan untuk kegiatan evaluasi secara detail, akan lebih baik jika item-item tersebut direvisi agar dapat spesifik untuk pelajaran (*subject*) tertentu.

Item Response Theory (IRT) memiliki banyak keunggulan dibandingkan *Classical Test Theory* (CTT), di antaranya yaitu *invariance property* yang dapat digunakan untuk pengembangan skala. Akan tetapi keunggulan tersebut hanya akan diperoleh jika asumsi-asumsi dari model IRT dapat terpenuhi. Penerapan model IRT untuk pengembangan skala di Indonesia masih sangat terbatas. Keterbatasan tersebut dikarenakan sulitnya memenuhi asumsi-asumsi model IRT.

Dalam penelitian ini, pengujian hanya dilakukan terhadap asumsi *unidimensionality* dan kesesuaian model dengan data (model fit). Untuk pengujian *unidimensionality*, seharusnya bisa dilakukan dengan *non-linier factor analysis*, namun karena keterbatasan program (*software*), peneliti menggunakan metode *Confirmatory Factor Analysis* (CFA) dengan program LISREL versi 8.3 (Jöreskog& Sörbom,1996), yang merupakan *linier factor analysis*. Penelitian selanjutnya diharapkan dapat menguji asumsi ini dengan metode yang lebih tepat, antara lain dengan: eigen value plot, Bejar's method, *non-linier factor analysis*, dan sebagainya.

Pengujian asumsi model fit dapat dilakukan dengan berbagai cara. Namun karena keterbatasan waktu dan program yang dimiliki peneliti, pengujian hanya dilakukan dengan melihat statistik *chi-square*. Untuk penelitian selanjutnya, pengujian terhadap asumsi model fit tidak hanya dilakukan dengan satu metode saja, akan tetapi diperkuat dengan melakukan investigasi terhadap residual, kekekaran model (*model robustness*), serta membuat prediksi terhadap distribusi dari skor tes.

DAFTAR PUSTAKA

Adams, R.J., & Khoo, S. (1996). *ACER Quest: The Interactive test analysis system*. Australia: The Australian Council for Educational Research.

Black, P., & William, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Phi Delta Kappa

80.2 : 139-148.

Cavanagh, R., Waldrip, B., Romanoski, J., Dorman, J. & Fisher, D. (2005). Measuring student perceptions of classroom assessment. *Annual Conference of the Australian Association for Research in Education:* Sydney.1 Maret 2008.www.aare.edu.au/05pap/cav05748.pdf

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Rinehart and Winston Inc.

Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). *What does research say about assessment?* 4 April 2008.

http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm

Dorman, J. P., Fisher, D. L. & Waldrip, B. G. (2005). *Classroom environment, students' perception of assessment, academic efficacy and attitude to science : A Lisrel analysis*. 1 Maret 2008. www.worldscibooks.com/socialsci/etextbook/5946/5946_chap1.pdf

Dorman, J. P., & Knightley, W. M. (2005). Development and validation of an instrument to assess students' perceptions of their assessment tasks. *European Conference on Educational Research,* Dublin.

Embretson, S.E., & Reise S.P. (2000). *Item Response Theory for Psychologist*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Elliot, A.J. & Harackiewicz, J. M. (1994). Goal setting, achievement orientation, and Intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology,* 72, 218-232.

Fisher, D. L., Waldrip, B. G., & Dorman, J. P. (2005). *Student perceptions of assessment: Development and validation of a questionnaire*. American Educational Research Association, Montreal.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publication, Inc.

Jöreskog, K.G., & D. Sörbom. 1996. *LISREL 8: User's reference guide*. Chicago: Scientific Software International, Inc.

Koul, R. & Fisher, D. (2005). *Using student perceptions in development, validation and application of an assessment questionnaire*. 1 Maret 2008.www.aare.edu.au/06pap/kou06298.pdf

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149 – 174.

Masters, G.N., & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159 – 176.

Moos, R.H. (1979). *Evaluating Educational Environments. San Francisco,* CA: Jossey-Bass, Inc.

Pandia, W.S.S. (2006). *Peran orientasi tujuan, self-efficacy, persepsi mengenai iklim kelas, dan pendekatan belajar terhadap prestasi belajar mahasiswa*. Disertasi Program Pascasarjana Fakultas Psikologi, Universitas Indonesia.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

Suryabrata, S. (2000). *Pengembangan Alat Ukur Psikologis*. Yogyakarta: Andi Offset.

Solomon, M.A. (1996). Impact of motivational climate on student's behaviors and perceptions in a physical education setting. *Journal of Educational Psychology, 88,* 731-738.

Schaffner, M., Burry-Stock, J.A., Cho, G., Boney, T, & Hamilton, G. (2000). *What do kids think when their teachers grade?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Schumacker, R.E. (2005). *Classical test analysis*. Applied Measurement Associates.

Steinberg, J. (2000). Student failure causes states to retool testing programs. *The New York Times,* p. A1. 15 Maret 2008.

http://www.nytimes.com/2000/12/22/national/22EXAM.html?pagewanted=all&ei=5070&en=6cff34d03cab4fa6&ex=1210737600

Van de Watering, G., Gijbels, D., Dochy, F., & van der Rijt, J. (2008). *Students' assessment preferences, perceptions of assessment*

and their Relationships to Study Results. 3 Maret 2008. www.springerlink.com

Walberg, H. J. (1976). Psychology of learning environments: Behavioral, structural, or perceptual? *Review of Research in Education, 4,* 142-178.