KEBERFUNGSIAN BUTIR DIFERENSIAL ULANGAN KENAIKAN KELAS VIII PELAJARAN MATEMATIKA SMP DI KABUPATEN SLEMAN

Sugeng Winarno, Badrun Kartowagiran,
LPMP D.I. Yogyakarta, Universitas Negeri Yogyakarta
Winarno_90@yahoo.com, badrunkw@yahoo.com

Abstrak

Penelitian ini bertujuan untuk mengetahui: 1) karakteristik perangkat tes ditinjau dari Teori Tes Klasik dan Teori Respons Butir, dan 2) butir-butir perangkat tes yang memuat keberfungsian butir diferensial pada perangkat Ulangan Kenaikan Kelas VIII mata pelajaran matematika SMP di Kabupaten Sleman. Sasaran penelitian ini adalah perangkat Ulangan Kenaikan Kelas VIII mata pelajaran matematika tahun pelajaran 2011/2012 SMP di Kabupaten Sleman. Pada penelitian ini dipilih respons jawaban siswa dari 11 SMP negeri yang tersebar di Kabupaten Sleman. Pemilihan sampel menggunakan teknik *stratified proportional random sampling.* Untuk mendeteksi *DIF* perangkat tes berdasarkan kelompok jenis kelamin siswa digunakan Metode Mantel-Haenszel. Hasil penelitian menunjukkan bahwa berdasarkan Teori Tes Klasik, informasi indeks keandalan tes sebesar 0,858 dengan indeks kesalahan baku pengukuran *(SEM)* 2,577, yang berarti tes andal dalam melakukan fungsi ukurnya. Sebanyak 23 butir soal (57,5%) mempunyai tingkat kesukaran yang baik, 40 butir soal (100%) mempunyai daya beda yang baik, dan 39 butir soal (97,5%) memiliki pengecoh yang baik. Berdasarkan pendekatan Teori Respons Butir Model Logistik 3 Parameter diperoleh sebanyak 40 butir soal (100%) mempunyai daya beda yang baik, sebanyak 37 butir soal (92,5%) mempunyai tingkat kesukaran yang baik, sebanyak 21 butir soal (52,5%) mempunyai peluang tebakan semu yang baik, serta sebanyak 39 butir soal (97,5%) cocok dengan Model Logistik 3 Parameter. Ditinjau dari keberfungsian butir diferensial, dari 40 butir soal terdapat tiga butir yang mengandung *DIF,* yaitu butir nomor 9, 15 dan 35

Kata kunci: *teori tes klasik, teori respons butir, dif, metode mantel-haenszel.*

DIFFERENTIAL ITEM FUNCTIONS IN THE GRADE PROMOTION TEST OF MATHEMATICS FOR GRADE VIII OF JUNIOR HIGH SCHOOLS IN SLEMAN REGENCY

Sugeng Winarno, Badrun Kartowagiran,
LPMP D.I. Yogyakarta, Universitas Negeri Yogyakarta
winarno_90@yahoo.com, badrunkw@yahoo.com

Abstract

This study aims to investigate: 1) characteristics of a test set in terms of Classical Test Theory and Item Response Theory, and 2) the number of test set items containing differential item functions in the promotion grade test set of mathematics for Grade VIII of junior high schools (JHSs) in Sleman Regency. The research target was the promotion grade test set of mathematics for Grade VIII of JHSs in Sleman Regency in the academic year of 2011/2012. The study selected the answer responses of the students from 11 public JHSs in Sleman Regency. The sample was selected using the stratified proportional random sampling technique. To investigate DIF of the test set based on the students' sex groups, the Mantel-Haenszel method was employed. The results of the study show that based on Classical Test Theory the test reliability index is 0.858 with a standard error of measurement (SEM) of 2.577, indicating that the test is reliable in terms of its measurement functions. A total of 23 test items (57.5%) have good difficulty indices, 40 test items (100%) have good discrimination indices, and 39 test items (97.5%) have good distractors. Based on Item Response Theory using the logistic model of 3 parameters, 40 test items (100%) have good discrimination indices, 37 test items (92.5%) have good difficulty indices, 21 test items (52.5%) have a probability of good pseudo-guessing, and 39 test items (97.5%) fit the logistic model of 3 parameters. In terms of differential item functions, of the 40 test items, 3 (three) test items contain DIF, namely items 9, 15, and 35.

Keywords: *classical test theory, item response theory, dif, mantel-haenszel method*

Pendahuluan

Meskipun pelaksanaan Ulangan Kenaikan Kelas (UKK) bersama di Kabupaten Sleman telah dilakukan beberapa kali namun hasil yang diperoleh belum memuaskan. Hal ini ditunjukkan dengan perolehan nilai UKK yang relatif masih rendah, khususnya pada mata pelajaran matematika. Rendahnya Nilai UKK mata pelajaran matematika mungkin karena peserta mengalami kesulitan mengerjakan butir-butir soal yang disebabkan tingkat kesukaran soal-soal tidak sesuai dengan kemampuan peserta tes, daya beda, atau pengecoh dalam soal-soal tersebut belum mampu berfungsi dengan baik, serta tidak adanya sinkronisasi antara pendekatan pengajaran dengan materi butir-butir tes.

Dalam pengukuran, idealnya tidak ada kesalahan, baik kesalahan acak maupun kesalahan sistematik. Seharusnya tidak ada kesalahan yang dilakukan oleh peserta tes, pelaksanaan tes maupun kesalahan pengukuran yang disebabkan oleh butir tes. Instrumen yang digunakan untuk mengukur seharusnya memiliki validitas dan reliabilitas yang tinggi sehingga instrumen tersebut dapat mengukur apa yang seharusnya diukur. Dengan demikian, tidak ada pihak yang merasa dirugikan.

Penyusunan perangkat pengukuran dituntut harus benar-benar dapat mengukur apa yang seharusnya diukur dan dapat memberikan hasil pengukuran yang dapat dipercaya. Penyusunan butir-butir tes harus berdasarkan penelaahan butir tes secara substansial yang meliputi telaah secara materi, konstruksi dan bahasa yamg dilakukan oleh para ahli di bidangnya dan harus melalui uji coba instrumen di lapangan sehingga diharapkan hasil ujian tersebut dapat memberikan gambaran atau informasi yang akurat tentang tingkat penguasaan peserta ujian yang sebenarnya. Artinya perbedaan skor yang diperoleh seorang peserta dengan peserta lain semata-mata disebabkan perbedaan kemampuan di antara mereka, bukan disebabkan faktor lain semisal karena perbedaan kelompok. Jika ternyata perangkat tes tersebut memihak salah satu kelompok, maka butir pada perangkat tes tersebut mengandung bias butir.

Untuk mengetahui ada/tidak bias butir pada suatu butir tes adalah dengan analisis keberfungsian butir diferensial *(*Differential *Item Functioning/DIF)*. Suatu butir menunjukkan *DIF* jika siswa yang mempunyai kemampuan sama tetapi dari kelompok yang berbeda , tidak mempunyai peluang yang sama untuk menjawab benar. Adanya butir diferensial ini mengakibatkan perangkat tes tersebut bersifat diskriminatif, ditinjau dari berbagai segi, misal ras, budaya, wilayah, jenis kelamin dan ekonomi.

Mata pelajaran Matematika perlu diberikan kepada semua peserta didik mulai dari sekolah dasar untuk membekali peserta didik dengan kemampuan berpikir logis, analitis, sistematis, kritis, dan kreatif, serta kemampuan bekerjasama. Kompetensi tersebut diperlukan agar peserta didik dapat memiliki kemampuan memperoleh, mengelola, dan memanfaatkan informasi untuk bertahan hidup pada keadaan yang selalu berubah, tidak pasti, dan kompetitif. Standar kompetensi dan kompetensi dasar matematika dalam dokumen ini disusun sebagai landasan pembelajaran untuk mengembangkan kemampuan tersebut di atas. Selain itu dimaksudkan pula untuk mengembangkan kemampuan menggunakan matematika dalam pemecahan masalah dan mengkomunikasikan ide atau gagasan dengan menggunakan simbol, tabel, diagram, dan media lain.

Penyusunan perangkat tes matematika harus cermat sehingga dapat mengukur apa yang seharusnya diukur. Perlu dipertimbangkan juga keberfungsian butir diferensial sehingga tidak memihak salah satu kelompok, karena menurut penelitian yang dilakukan di Kanada oleh Gierl, Khaliq dan Boughton menyimpulkan bahwa pada perangkat tes matematika dan sains menguntungkan siswa yang berjenis kelamin laki-laki (1999, p.10). Berdasarkan pada pertimbangan tersebut, penulis memilih keberfungsian butir diferensial berdasarkan jenis kelamin pada perangkat UKK kelas VIII mata pelajaran matematika jenjang SMP di Kabupaten Sleman sebagai kajian dalam tesis ini.

Permasalahan dalam penelitian ini adalah sebagai berikut: 1) bagaimana karakteristik butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 berasarkan teori tes klasik dan teori respons butir; 2) apakah butir-butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 mengandung *DIF* sehingga hasil pengukuran tidak dapat menggambarkan kemampuan siswa yang sebenarnya?

Penelitian ini bertujuan untuk mengetahui: 1) karakteristik butir perangkat tes matematika UKK kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ber-

dasarkan teori tes klasik dan teori respons butir; 2) keberfungsian butir diferensial pada perangkat tes matematika UKK kelas VIII jenjang SMP di Kabupaten Sleman.

Hasil penelitian ini diharapkan akan bermanfaat : 1) sebagai sumber informasi untuk pemilihan butir-butir tes yang layak masuk dalam bank soal; 2) sebagai informasi bagi penelitian lain yang berkaitan dengan pendeteksian keberfungsian butir diferensial; 3) sebagai sumber informasi untuk penyusunan tes yang lebih baik.

Masalah evaluasi hasil belajar meliputi alat ukur yang digunakan, cara menggunakan, cara penilaian dan evaluasinya (Mardapi, 2004, p.14). Pada prinsipnya alat ukur yang digunakan harus memiliki bukti kesahihan dan keandalan. Kesahihan alat ukur dapat dilihat dari konstruk alat ukur, yaitu mengukur seperti seperti yang direncanakan. Menurut teori pengukuran, substansi yang diukur harus satu dimensi. Konstruksi alat ukur dapat ditelaah pada aspek materi, teknik penulisan soal dan bahasa yang digunakan.

Hasil pengukuran harus memiliki kesalahan yang sekecil mungkin. Tingkat kesalahan ini berkaitan dengan kehandalan alat ukur. Alat ukur yang baik memberikan hasil yang konstan bila digunakan berulang-ulang, asalkan kemampuan yang diukur tidak berubah. Kesalahan pengukuran ada yang bersifat acak dan ada yang bersifat sistematik. Kesalahan acak disebabkan kondisi fisik dan mental yang diukur dan yang mengukur bervariasi. Kondisi mental termasuk emosi seseorang yang selalu bervariasi, dan variansinya diasumsikan acak. Kesalahan bersifat sistematik disebabkan oleh alat ukurnya, yang diukurnya, dan yang mengukur. Dalam melakukan pengukuran, pendidik bisa membuat kesalahan yang sistematik. Kesalahan ini bisa terjadi pada saat penskoran, atau ada pendidik yang pemurah dan ada yang mahal.

Sumarna Surapranata (2005, p.19) mendefinisikan tes adalah sehimpunan pertanyaan yang harus dijawab, atau pernyataan-pernyataan yang harus dipilih, ditanggapi, atau tugas-tugas yang harus dilakukan oleh orang yang dites *(testee)* dengan tujuan untuk mengukur suatu aspek (perilaku/atribut) tertentu orang yang dites tersebut (Surapranata, 2005, p.19). Tes pada umumnya dimaksudkan untuk mengukur aspek-aspek perilaku manusia, seperti aspek pengetahuan (kognitif), sikap (afektif) maupun aspek keterampilan (psikomotorik).

Mardapi (2008, p. 67) menyimpulkan bahwa tes adalah sejumlah pertanyaan yang diberi tanggapan. Tujuannya mengukur tingkat kemampuan seseorang atau mengungkap aspek tertentu dari peserta tes. Tes juga didefinisikan sebagai prosedur baku untuk mendapatkan sampel perilaku dari domain yang spesifik. Untuk mendapatkan sampel itu, dibutuhkan pengukuran, yaitu menggunakan tes berupa pemberian pertanyaan yang akan dijawab, selanjutnya diperhitungkan benar tidaknya jawaban tes.

Dapat disimpulkan bahwa tes adalah alat bantu pengukur perilaku atau kemampuan seseorang. Tes juga merupakan prosedur yang sistematis, artinya penyusunan butir tes dan pemberian skor harus terperinci dan jelas. Tes juga merupakan sampel perilaku. Artinya, seberapa banyak tes yang disusun tidak bisa mencakup seluruh isi materi yang akan diukur. Menurut Mardapi (2008, p.67) untuk mengukur kemampuan sesungguhnya dari peserta didik diperlukan perancangan yang baik dalam tes. Hasil itu kemudian digunakan untuk memantau perkembangan mutu pendidikan.

Penggunaan bentuk soal dalam tes prestasi belajar, secara umum dapat dikelompokkan menjadi dua kategori yaitu: 1) tes uraian, terdiri dari uraian bebas, uraian terbatas atau isian singkat, uraian berstruktur, dan 2) tes objektif, terdiri dari pilihan benar-salah, pilihan ganda, dan menjodohkan. Didalam kajian teori pada penelitian ini, hanya akan dibahas tiga bentuk soal yaitu pilihan ganda, isian singkat dan uraian.

Soal bentuk pilihan ganda merupakan bentuk soal yang populer digunakan oleh guru. Bentuk soal ini, jawabannya harus dipilih dari beberapa kemungkinan jawaban yang telah disediakan. Penggunaan tes pilihan ganda, pada umumnya dijumpai pada ujian yang berskala besar/massal karena sifatnya yang obyektif dan mudah penskorannya. Bentuk soal ini juga dianggap pilihan yang tepat untuk ulangan kenaikan kelas atau ujian akhir dimana bahan pelajaran yang hendak diujikan biasanya cukup banyak.

Dilihat dari strukturnya, bentuk soal pilihan ganda terdiri dari pokok soal *(stem)* dan pilihan jawaban *(option)*. Pilihan jawaban terdiri atas satu kunci jawaban dan yang lainnya pengecoh (distraktor). Pokok soal *(stem)* dapat berupa pertanyaan atau pernyataan tidak lengkap.

Setiap bentuk tes, memiliki keunggulan dan keterbatasan. Surapranata (2005, p.178) secara garis besar menyebutkan keunggulan tes yang terdiri dari soal-soal bentuk pilihan ganda yakni: 1) jumlah materi yang dapat ditanyakan relatif tak terbatas dibandingkan dengan materi yang dapat dicakup soal bentuk lainnya. (Blerkom 2009, pp.89 - 91); 2) dapat mengukur berbagai jenis kognitif mulai dari ingatan sampai evaluasi (Kubiszyn dan Borich, 2010, p.144; 3) penskorannya mudah, cepat, objektif, dan dapat mencakup ruang lingkup, bahan dan materi yang luas dalam suatu tes untuk suatu kelas atau jenjang; 4) sangat tepat untuk ujian yang pesertanya sangat banyak sedangkan hasilnya harus segera; 5) reliabilitas soal pilihan ganda relatif lebih tinggi dibandingkan dengan soal uraian.

Keterbatasan tes pilihan ganda antara lain: 1) kurang dapat digunakan untuk mengukur kemampuan verbal; 2) peserta didik tidak mempunyai keleluasaan dalam menulis, mengorganisasikan dan mengekspresikan gagasan yang mereka miliki yang dituangkan kedalam kata atau kalimatnya sendiri; 3) tidak dapat digunakan untuk mengukur memampuan *problem solving*; 3) sangat sensitif terhadap menerka; 4) penyusunan soal yang baik memerlukan waktu yang relatif lama dibandingkan dengan bentuk soal lainnya; 5) sangat sukar menentukan alternatif jawaban yang benar-benar homogen, logis, dan berfungsi.

Menurut Allen & Yen (1979:57) ada tujuh asumsi dasar yang harus dipenuhi dalam teori tes klasik. Ketujuh asumsi tersebut adalah: 1) skor yang diperoleh seseorang dari hasil suatu pengukuran (skor amatan) dapat diuraikan menjadi skor yang sebenarnya dan skor kesalahan pengukuran; 2) nilai harapan skor perolehan sama dengan skor murni; 3) tidak ada hubungan (korelasi) antara skor kesalahan pengukuran dengan skor murni (skor sebenarnya); 4) skor-skor kesalahan pada dua tes (yang mengukur hal yang sama) tidak saling berkorelasi; 5) jika ada dua tes (untuk mengukur atribut yang sama) maka skor-skor kesalahan pada tes pertama tidak berkorelasi dengan skor-skor murni pada tes yang kedua; 6) jika dua perangkat tes (untuk mengukur atribut yang sama) merupakan skor murni dan skor kesalahan sama maka kedua perangkat tes itu disebut paralel; dan 7) Asumsi-asumsi tersebut selanjutnya dijadikan dasar dalam pengembangan formula-formula untuk mengetahui indeks kesahihan *(validity)* dan indeks keandalan *(reliability)*.

Teori tes klasik didasarkan pada model aditif (Allen dan Yen, 1979, p.57) yaitu skor amatan *(observed score)* merupakan penjumlahan dari skor sebenarnya *(true score)* dan skor kesalahan pengukuran *(error score)*.

$$X = T + E$$

dimana :
X : Skor amatan *(observed score)*
T : Skor sebenarnya *(true score)*
E : Skor kesalahan pengukuran *(error score)*

Kesalahan pengukuran yang dimaksud dalam teori ini merupakan kesalahan tidak sistematik atau acak.

Untuk melihat kualitas butir tes sedikitnya diperlukan tiga informasi tentang karakter butir tes. Karakter butir atau parameter butir tersebut adalah tingkat kesukaran butir, daya beda butir dan efektivitas distraktor. Selain itu validitas dan reliabilitas tes dapat digunakan untuk memperkuat informasi yang diperoleh.

Validitas suatu perangkat tes adalah kemampuan suatu tes untuk mengukur apa yang seharusnya diukur (Allen dan Yen, 1979, p.95; Syaifudin Azwar, 2006, p.43; Kerlinger, 1986, p.731). Validitas didefinisikan sebagai ukuran seberapa cermat suatu tes melakukan fungsi ukurnya (Mardapi, 2004, p.25). Dengan demikian dapat disimpulkan bahwa validitas ditentukan oleh ketepatan dan kecermatan hasil pengukuran. Menurut Syaifudin Azwar (2006, p.43) suatu alat ukur yang tinggi validitasnya akan memiliki kesalahan pengukuran yang kecil, artinya skor setiap subyek yang diperoleh oleh alat ukur tersebut ditidak jauh berbeda dari skor yang sesungguhnya.

Reliabilitas merupakan hal yang sangat penting dimiliki oleh suatu tes atau instrumen. Badrun Kartowagiran (2007, p.10) menyatakan bahwa baik buruknya tes, tidak hanya dilihat dari butir-butirnya tetapi juga dilihat secara keseluruhan, atau dilihat dari reliabilitas instrumen atau reliabilitas tes. Nunnaly (1981, p.191) menyatakan bahwa reliabilitas adalah kestabilan skor yang diperoleh orang yang sama ketika diuji ulang dengan tes yang sama pada situasi yang berbeda atau dari satu pengukuran ke pengukuran lainnya. Jadi reliabilitas dinyatakan sebagai tingkat keajegan atau kemantapan hasil dari hasil dua pengukuran terhadap hal yang sama.

Formula Cronbach's Alpha, lebih umum digunakan untuk menghitung koefisien

reliabilitas alat ukur. Bentuk persamaannya secara matematis dinyatakan seperti berikut.

$$\hat{\alpha} = \frac{k}{k-1}(1 - \frac{\sum \hat{\sigma}^2{}_i}{\hat{\sigma}^2{}_x})$$

Keterangan:

$\hat{\alpha}$ = koefisien relibilitas

$\hat{\sigma}^2{}_i$ = varians butir ke-i

$\hat{\sigma}^2{}_x$ = varians tes total

k = banyaknya butir tes

Besarnya koefisien reliabilitas yang dianggap memuaskan tidak dapat ditentukan secara pasti. Hal ini dikarenakan koefisien reliabilitas yang diperoleh berdasarkan perhitungan terhadap data empiris dari sekelompok sujek pada dasarnya hanya merupakan estimasi saja dari reliablitas sesungguhnya dan hanya berlaku bagi kelompok subjek yang dijadikan dasar perhitungan saja (Azwar, 2006, pp.116-117). Meskipun tidak ada perjanjian, secara umum dapat diterima bahwa tes yang digunakan untuk membuat keputusan pada siswa secara perorangan harus memiliki koefisien reliabilitas minimal 0,85 dan untuk membuat keputusan pada siswa secara kelompok, tes harus memiliki koefisien reliabilitas sebesar 0,64.

Tingkat kesukaran suatu butir soal, yang disimbolkan p_i, berguna untuk melihat seberapa baik kualitas suatu butir soal, yang didasarkan pada proporsi siswa yang menjawab benar pada butir tertentu. Secara matematis rumus untuk menentukan besarnya indeks tingkat kesukaran butir (p), sebagai berikut:

$$p_i = \frac{n_i}{N}$$

dimana: p_i adalah indeks tingkat kesukaran butir, n_i adalah banyaknya peserta didik yang menjawab butir dengan benar, dan N adalah banyaknya peserta didik yang menjawab butir.

Tingkat kesukaran yang baik adalah 0,30 sampai dengan 0,70 (Allen & Yen, 1979, p.121 & Surapranata, 2006, p.47). Pada interval ini informasi tentang kemampuan siswa yang diperoleh akan maksimal. Butir soal yang memiliki tingkat kesukaran dibawah 0,3 dikategorikan sebagai butir sukar. Butir soal yang memiliki tingkat kesukaran di atas 0,70 dikategorikan sebagai mudah.

Daya pembeda suatu butir tes berfungsi untuk memberikan informasi seberapa besar kemampuan soal tersebut dapat membedakan peserta tes yang memperoleh jumlah skor yang tinggi dan peserta tes yang jumlah skornya rendah (Allen dan Yen, 1979, p.122). Besaran daya beda butir soal dinyatakan dalam suatu indeks daya beda. Indeks ini menunjukkan kesesuaian antara fungsi soal dengan fungsi tes secara keseluruhan. Indeks daya pembeda dikatakan suatu butir soal dikatakan baik jika memiliki nilai lebih besar atau sama dengan 0,2. Metode indeks korelasi dihitung dengan korelasi *point biserial* dan *biserial.* Korelasi *point biserial* maupun *biserial* adalah korelasi *product moment* yang diterapkan pada data, sedangkan variabel-variabel yang dikorelasikan mempunyai sifat-sifat yang masing-masing berbeda satu sama lain. Variabel skor butir soal bersifat dikotomi, sedangkan variabel skor atau sub skor total bersifat kontinum. Variabel skor tes dinamakan dikotomi karena skor-skor yang terdapat pada tes hanya ada 1 dan 0. Variabel skor total bersifat kontinum, yang diperoleh dari jumlah jawaban benar. Korelasi point biserial ditentukan dengan menggunakan rumus:

$$\rho_{pbis} = \frac{\mu_+ - \mu_\tau}{\sigma_\tau}\sqrt{p.q} \quad (5)$$

ρ_{pbis} = koefisien korelasi point biserial

μ_+ = *mean* skor pada tes dari peserta yang memiliki jawaban benar

μ_τ = *mean* skor total

σ_τ = *deviasi standar* skor total

p = proporsi peserta ujian yang menjawab benar pada butir tes

q = 1-p

Menurut Surapranata (2006, p.43), suatu pengecoh dikatakan berfungsi dengan baik jika paling sedikit dipilih oleh 5% peserta tes. Nitko & Brookhart (2007, p.165) mengatakan distraktor dikatakan berfungsi manakala paling tidak dipilih oleh seorang peserta tes dari kelompok terendah. Apabila pengecoh dipilih secara merata, maka termasuk pengecoh yang sangat baik. Apabila pengecoh lebih banyak dipilih oleh peserta tes kelompok atas dibandingkan dengan kelompok bawah, maka termasuk pengecoh yang menyesatkan.

Teori Respons Butir

Teori respons butir memperbaiki keterbatasan yang ada pada teori tes klasik. Menurut Hambleton, Swaminathan dan Rogers (1991, pp.2-5) teori respons butir bertujuan

membentuk : (1) Karakterisitik butir yang tidak tergantung pada kelompok subyek, (2) Skor tes yang dapat menggambarkan profisiensi subyek dan tidak tergantung pada taraf kesukaran tes, (3) Model yang lebih menekankan pada tingkat butir daripada tingkat tes, (4) Model tes yang tidak memerlukan asumsi paralel dalam pengujian reliabilitasnya, dan (5) Model yang menguraikan sebuah ukuran keputusan untuk tiap skor kemampuan yakni ada hubungan fungsional antara jawaban peserta ujian terhadap tingkat kemampuan yang dimiliki.

Dalam teori respons butir, model matematikanya mempunyai makna bahwa probabilitas subyek untuk menjawab butir dengan benar tergantung pada kemampuan subyek dan karakteristik butir. Ini berarti peserta tes yang berkemampuan tinggi akan mempunyai probabilitas menjawab benar lebih besar jika dibandingkan dengan peserta yang memiliki kemampuan rendah.

Hambleton, Swaminathan dan Rogers (1991, p.9) menyatakan bahwa ada tiga asumsi yang mendasari teori respons butir. Asumsi pertama, peluang menjawab benar suatu butir tidak dipengaruhi oleh peluang menjawab benar butir yang lain. Sifat ini disebut independensi lokal *(local independence)*. Asumsi kedua, tes mengukur pada satu dimensi kemampuan. Sifat ini disebut unidimensi *(unidimensionality)*. Asumsi ketiga, pola respon setiap butir tes dapat digambarkan dalam bentuk kurva karakteristik butir.

Model logistik tiga parameter

Sesuai namanya, model logistik tiga parameter ditentukan oleh tiga karakteristik butir yaitu tingkat kesukaran, daya pembeda, dan parameter tebakan semu. Dengan adanya tebakan semu memungkinkan subyek yang memiliki kemampuan rendah mempunyai peluang untuk menjawab butir soal dengan benar. Peluang menjawab benar terhadap butir soal dapat dihitung dengan rumus (Hambleton & Swaminathan, 1985, p.49) :

$$P_i(\theta) = C_i + (1-C_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

$i = 1, 2, \ldots, n$ (13)

Tambahan parameter pada model ini adalah C_i yang disebut parameter faktor tebakan *(pseudo)* butir ke i. Arti parameter ini adalah besarnya peluang menjawab benar bagi peserta yang memiliki kemampuan rendah. Peserta ujian inilah yang sering menjawab secara untung-untungan atau *pseudoguessing*. Pada suatu butir tes, nilai C_i berkisar antara 0 dan 1. Suatu butir tes dikatakan baik jika nilai C_i tidak lebih dari 1/k, dengan k adalah banyaknya pilihan. Jadi misalkan pada suatu perangkat tes pilihan ganda, ada 4 pilihan untuk setiap butir tes nya, butir ini dikatakan baik jika nilai C_i tidak lebih dari 0,25.

Harga e^D berfungsi sebagai faktor penskalaan pada skala kemampuan. Parameter *bi* merupakan suatu titik pada skala kemampuan, yaitu probabilitas untuk menjawab benar adalah sebesar 0,50. Semakin besar nilai parameter *bi*, maka semakin besar pula kemampuan yang dituntut dari peserta tes untuk memperoleh 50% peluang menjawab dengan benar. Artinya butir soal tersebut semakin sukar dan sebaliknya.

Keberfungsian Butir Diferensial

Hambleton (1991, p.110) mengatakan bahwa bias butir terjadi jika seseorang dengan kemampuan sama berasal dari kelompok yang berbeda memiliki peluang berbeda untuk menjawab benar butir yang sama. Camilli (1993, p.397) mendefinisikan bias sebagai ketidakcermatan atau kesalahan sistematik dalam tes dalam mengukur anggota kelompok tertentu. Lamprianou (2009, p.100) menggambarkan bias butir atau menurut istilahnya *Differential Item Functioning (DIF)* terjadi jika siswa dari kelompok yang berbeda mempunyai peluang yang berbeda untuk menjawab dengan benar suatu butir soal yang sama. Dari pendapat-pendapat tersebut dapat disimpulkan, suatu butir tes dikatakan bias jika butir tersebut tidak memberi peluang yang sama untuk menjawab benar pada tes yang memiliki kemampuan sama hanya karena berasal dari kelompok yang berbeda.

Pendekatan bias dalam penelitian ini dibatasi pada bias internal butir. Suatu tes dikatakan bias jika dua orang peserta tes dengan kemampuan yang sama, dari kelompok yang berbeda tidak memperoleh peluang menjawab benar yang sama. Beberapa ahli psikometri telah melakukan langkah-langkah untuk menghilangkan konotasi yang merendahkan berkaitan dengan istilah bias butir. Istilah untuk menggantikan bias butir adalah *differential item performance (DIP)* atau *differential item functioning*

(DIF). Untuk seterusnya dalam penelitian ini digunakan istilah *DIF* atau keberfungsian butir diferensial.

Mantel dan Haenszel pada tahun 1959 menyampaikan prosedur untuk suatu padanan kelompok, yang oleh Holland dan Thayer (1988, p.129) dipakai untuk mendeteksi DIF, yang kemudian dikenal dengan metode Mantel-Haenszel. Metode Mantel-Haenszel efektif dalam pendeteksian *DIF*, khususnya untuk sampel berukuran kecil. Ahli psikometri seperti Camilli dan Shepard juga masih menggunakan metode ini untuk mendeteksi *DIF*.

Penerapan metode *Mantel-Haenszel* dapat diilustrasikan sebagai berikut. Setiap peserta tes hanya dapat digolongkan pada satu kelompok yaitu kelompok acuan *(R)* atau kelompok fokal *(F)*. Misal masing-masing kelompok *R* dan *F* dibagi menjadi tiga kelompok yaitu kelompok atas, kelompok tengah dan kelompok bawah. Perbandingan dilakukan pada kelompok-kelompok hasil pembagian yang setara dan dilakukan per butir dari perangkat tes yang sedang dianalisis. Misalnya, perbandingan dilakukan antara kelompok *R* atas dengan *F* atas, kelompok *R* tengah dengan *F* tengah dan kelompok *R* bawah dengan *F* bawah. Setiap perbandingan dibuat tabel kontingensinya, yang akan digunakan untuk menghitung besarnya harga statistik Mantel-Haenszel. Harga statistik tersebut digunakan untuk menentukan apakah terdapat *DIF* pada butir soal tersebut.Prosedur tersebut juga diilustrasikan oleh Nabeel Abedalaziz (dalam International Journal for Educational Studies, 2011).

Keunggulan metode Mantel-Haenszel adalah (1) sederhana konsepnya, (2) dapat dikerjakan oleh pengguna yang familier dengan program komputer, (3) dapat digunakan pada sampel kecil. Sedangkan kelemahannya adalah : (1) skor total sebagai variabel pemadanan dapat menyebakan adanya kesalahan, karena dapat memuat skor butir yng terkena *DIF*, (2) tidak dapat mendeteksi *DIF* yang uniform.

Metode Penelitian

Jenis Penelitian

Penelitian ini bersifat *ex-post facto*, karena dalam penelitian ini tidak dilakukan perlakuan apapun terhadap variabel penelitian. Data yang digunakan adalah jawaban/respons peserta Ulangan Kenaikan Kelas VIII mata pelajaran matematika jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012.

Tempat dan Waktu Penelitian

Penelitian dilaksanakan pada SMP di lingkungan Dinas Pendidikan dan Olah Raga Kabupaten Sleman dan memerlukan waktu kurang lebih 3 bulan.

Populasi dan Sampel Penelitian

Populasi penelitian adalah seluruh siswa SMP kelas VIII yang mengikuti ulangan kenaikan kelas mata pelajaran matematika tahun pelajaran 2011/2012 di Kabupaten Sleman sebanyak 12.435 siswa. Subyek dari penelitian ini adalah lembar jawaban peserta Ulangan Kenaikan Kelas VIII mata pelajaran matematika jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012. Sampel yang diambil dalam penelitian ini adalah 710 siswa yang diambil dari 10 SMP di Kabupaten Sleman.

Teknik dan Instrumen Pengumpulan Data

Pengumpulan data dilakukan dengan teknik dokumentasi yaitu mengutip respons/ jawaban siswa pada perangkat Ulangan Kenaikan Kelas VIII mata pelajaran matematika jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012.

Agar diperoleh sampel yang benar-benar mewakili populasi yang jumlahnya sangat besar, maka dilakukan pengambilan sampel dengan metode *stratified proportional random sampling*. *Stratified proportional random sampling* adalah teknik pengambilan sampel yang memperhatikan strata-strata dalam populasi dan subyek yang terdapat dalam tiap strata diambil secara proporsional dan random. Subjek dari penelitian ini adalah lembar jawaban peserta Ulangan Kenaikan Kelas VIII mata pelajaran matematika jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 untuk tipe tes objektif pilihan ganda.

Metode ini memiliki langkah-langkah sebagai berikut, strata berupa sekolah diambil berdasarkan kelompok perolehan rerata nilai Ujian Nasional (UN) SMP/MTs tahun 2011/ 2012 di Kabupaten Sleman. SMP Negeri di Kabupaten Sleman terbagi menjadi 3 kelompok besar, yaitu kelompok tinggi, sedang dan bawah. Penelitian ini terdiri dari dua kategori utama, yaitu: 1) penelitian karakteristik butir pada perangkat tes dengan menggunakan teori tes klasik dan teori respons butir; dan 2) penelitian pendeteksian DIF pada butir-butir tes

dengan pendekatan teori tes klasik menggunakan metode Mantel-Haenszel.

Validasi dan Reliabilitas Instrumen

Validasi perangkat soal Ulangan Kenaikan Kelas menggunakan validasi isi dengan melibatkan guru matematika senior di SMP Negeri 4 Pakem. Sedangkan reliabilitas dihitung menggunakan program *ITEMAN*.

Teknik Analisis Data

Analisis terhadap karakteristik butir tes dilakukan dengan menggunakan pendekatan teori tes klasik dan teori respons butir. Dari teori tes klasik akan diperoleh informasi butir tentang tingkat kesukaran dan daya pembeda. Sementara untuk melihat karakteristik perangkat tes digunakan reliabilitas dan validitas. Untuk memperoleh informasi tersebut, digunakan program Iteman.

Selain menggunakan pendekatan teori tes klasik, untuk melihat karakteristik butir juga digunakan pendekatan teori respons butir yaitu dengan menggunakan model logistik tiga parameter. Ketiga parameter tersebut adalah tingkat kesukaran, daya pembeda dan faktor tebakan. Untuk memperoleh informasi tersebut, digunakan program *Bilog MG* fase pertama. Sementara untuk melihat karakteristik perangkat tes digunakan fungsi informasi. Untuk menghitungnya digunakan program *Bilog MG* fase kedua.

Metode yang digunakan dalam pendeteksian *DIF* pada penelitian ini adalah metode Mantel-Haenszel. Holland dan Trayer (1986) dalam French dan Miller (1995, p.317) menyatakan bahwa teknik Mantel-Haenszel telah terbukti efektif dalam pendeteksian DIF, khususnya pada saat ukuran sampel kecil. Hal ini diperkuat hasil penelitian Heri Retnawati (2003, p.135) menyimpulkan bahwa metode Mantel-Haenszel paling banyak menghasilkan butir yang mengandung DIF dibanding dengan metode lainnya (metode Kh-kuadrat dari Lord dan metode Uji Perbandingan Kemungkinan). Untuk mendeteksi adanya *DIF*, sampel dibagi dalam dua kelompok yaitu kelompok acuan dan kelompok yang diteliti. Kelompok acuan dalam penelitian ini adalah peserta Ulangan Kenaikan Kelas berjenis kelamin perempuan. Sedang kelompok fokalnya adalah peserta Ulangan Kenaikan Kelas berjenis kelamin laki-laki. Setelah dilakukan analisis untuk melihat karakteristik butir soal, maka langkah selanjutnya adalah analisis untuk mendeteksi adanya *DIF*. Dasar analisis ini adalah butir soal yang dinyatakan baik pada analisis karakteristik butir soal.

Hasil Penelitian dan Pembahasan

Analisis butir tes pilihan ganda dengan pendekatan teori tes klasik dilakukan dengan bantuan program *Item and Test Analysis (ITEMAN)* versi 3.00. Tujuan analisis butir tes pilihan ganda dengan program *ITEMAN* adalah untuk mengetahui karakteristik dan kualitas empirik butir tes Matematika yang digunakan pada UKK mata pelajaran Matematika Kelas VIII SMP tahun pelajaran 2011/2012 di Kabupaten Sleman.

Analisis ini akan menghasilkan karakteristik butir soal dan perangkat tes. Karakteristik butir soal, meliputi: (1) tingkat kesukaran, (2) daya beda, dan (3) efektifitas distraktor. Karakteristik perangkat tes, antara lain: mean, median, indeks keandalan, kemencengan, dan kesalahan baku pengukuran.

Statistik Perangkat Tes

Statistik perangkat tes meliputi: rerata, standar deviasi, indeks keandalan, dan estimasi kesalahan pengukuran. Hasil analisis perangkat tes Matematika UKK Kelas VIII SMP tahun pelajaran 2011/2012 disajikan pada lampiran dan Tabel 1.

Tabel 1. Statistik Perangkat Tes dengan Program *ITEMAN*

Aspek	Ket	Aspek	Ket
N of Items	40	Maximun	40
N of Examinees	710	Median	27
Mean	26,439	Alpha	0,858
Variance	46,815	*SEM*	2,577
Std. Dev.	6,842	Mean P	0,661
Skew	-0,221	Mean Item-Tot	0,389
Kurtosis	-0,475	Mean Biserial	0.536
Minimum	9,000		

Berdasarkan Tabel 1 di atas diperoleh informasi bahwa dari 40 butir soal pilihan ganda yang dikerjakan oleh 710 peserta ujian, hasilnya menunjukkan rerata skor 26,439 dengan sebaran skor *(variance)* 46,815 dan simpangan baku (standard deviasi) 6,842. Skor minimum 9,00 untuk peserta ujian 140L, 176L, 258L, 508P, 560P, dan maksimum 40,00 untuk peserta ujian 003L, 044L, 247L, 271L, 398P,

401P, 416P, 419P. Indeks keandalan tes tersebut 0,858 dan indeks kesalahan baku pengukuran *(SEM)* 2,577.

Tingkat Kesukaran

Pada analisis butir soal pilihan ganda dengan bantuan program *ITEMAN*, tingkat kesukaran ditentukan berdasarkan proporsi menjawab benar atau *proportion correct*. Dalam penelitian ini, kriteria tingkat kesukaran (p) butir soal yang dianggap baik adalah $0,3 \leq p \leq 0,7$. Butir soal dikatakan sukar jika nilai $p < 0,3$. Butir soal dikatakan mudah jika nilai $p > 0,7$. Dengan menggunakan *output* hasil analisis *ITEMAN* sebagaimana disajikan pada lampiran 2, maka tingkat kesukaran naskah tes Matematika Ulangan Kenaikan Kelas VIII SMP tahun pelajaran 2011/2012 di Kabupaten Sleman dapat dikategorikan menjadi mudah dan sedang. Selengkapnya hasil analisis tingkat kesukaran butir soal menggunakan program *ITEMAN* dapat dilihat pada Tabel 2 berikut.

Tabel 2. Tingkat Kesukaran Butir Soal Berdasarkan Teori Tes Klasik

Kategori	Soal UKK Kelas VIII		
	No Butir	Jumlah	%
Mudah ($p > 0,70$)	1, 2, 3, 4, 7, 9, 10, 11, 12, 13, 16, 20, 21, 25, 29, 30, 33	17	42,5
Sedang ($0,30 \leq p \leq 0,70$)	5, 6, 8, 14, 15, 17, 18, 19, 22, 23, 24, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39,40	23	57,5
Sukar ($p < 0,30$)	-	0	0

Hasil analis *ITEMAN* memberikan informasi sebanyak 17 butir (42,5%) tergolong mudah yakni butir nomor 1, 2, 3, 4, 7, 9, 10, 11, 12, 13, 16, 20, 21, 25, 29, 30, dan 33. Butir soal yang dinyatakan sedang sebanyak 23 butir (57,5%) yakni butir nomor 5, 6, 8, 14, 15, 17, 18, 19, 22, 23, 24, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39, dan 40. Tingkat kesukaran terendah terletak pada butir soal nomor 31 dan 38 sebesar 0,352 dan tertinggi pada butir nomor 2 sebesar 0,975.

Daya Beda

Daya beda dalam penelitian ini ditentukan berdasarkan nilai korelasi biserial (r_{bis}), dengan batasan kriteria daya beda butir yang baik $r_{bis} \geq 0,30$, cukup baik 0,20 s.d 0,29, dan tidak baik $r_{bis} < 0,20$. Berdasarkan hasil analisis butir soal dengan program *ITEMAN* yang disajikan, diketahui bahwa butir soal yang baik sebanyak 40 butir (100%). Daya beda butir yang tertinggi 0,753 untuk butir soal nomor 33 dan terendah 0,331 pada butir nomor 30.

Efektivitas Distraktor

Distraktor yang efektif apabila mempunyai daya tarik bagi peserta tes yang berkemampuan rendah. Apabila distraktor dipilih secara merata, maka termasuk distraktor yang sangat baik. Selain itu, pada analisis butir soal menggunakan program *ITEMAN version 3.00* distraktor yang baik harus memiliki biserial negatif selain kunci.

Merujuk pada kriteria tersebut, dinyatakan bahwa butir soal yang memiliki pengecoh yang baik untuk naskah tes Ulangan Kenaikan Kelas VIII Mata Pelajaran Matematika SMP Tahun Pelajaran 2011/2012 di Kabupaten Sleman sebanyak 39 butir (97,5%). Distraktor yang dinyatakan kurang baik sebanyak 1 butir (2,5%) yakni butir nomor 27. Kategori distraktor berdasarkan *output* program *ITEMAN* ditampilkan pada dan Tabel 3 berikut.

Tabel 3. Efektivitas Distraktor Butir Soal Berdasarkan Teori Tes Klasik

Kategori	Soal UKK Kelas VIII		
	No Butir	Jumlah	%
Baik	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, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	39	97,5
Kurang baik	27	1	2,5

Hasil Analisis Kuantitatif dengan Teori Respons Butir

Hasil *(output)* analisis butir soal dengan program *BILOG* menghasilkan informasi yang berkaitan dengan kemampuan peserta, daya beda butir, tingkat kesulitan butir, peluang tebakan semu, dan kecocokan antara data dengan model. Model yang digunakan pada analisis ini adalah model logistik tiga parameter.

Kemampuan Peserta

Hasil analisis statistik kemampuan peserta berdasarkan teori respons butir program *BILOG* dirangkum pada Tabel 4 berikut ini.

Tabel 4. Statistik kemampuan peserta dengan Program *BILOG*

Interval Kemampuan	Kemampuan	Jumlah Peserta	%
$1,5 < \theta < +\sim$	Sangat Tinggi	41	5,78%
$0,5 < \theta < 1,5$	Tinggi	129	18,17%
$-0,5 < \theta < 0,5$	Sedang	346	48,73%
$-1,5 < \theta < -0,5$	Rendah	162	22,82%
$-\sim < \theta < -1,5$	Sangat Rendah	32	4,51%

Tabel di atas menunjukkan, kemampuan peserta tes mengikuti distribusi normal. Sebagian besar peserta tes (48,73%) berkemampuan sedang.

Daya Beda Butir (a)

Daya beda butir dikatakan baik menurut teori respons butir jika nilai a terletak pada interval [0,2]. Hasil analisis menunjukkan semua butir soal (100%) memiliki daya beda yang baik.

Tabel 5. Daya Beda Butir Soal Berdasarkan Teori Respons Butir

Kategori	Soal UKK Kelas VIII		
	No Butir	Jml	%
Baik ($0 \le a \le 2$)	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, 29,30,31,32,33,34, 35,36,37,38,39,40	40	100%
Kurang baik ($a < 0$ atau $a > 2$)	-	0	0%

Tingkat Kesukaran (b)

Tingkat kesukaran (b) dalam teori respons butir dikatakan baik jika terletak pada interval [-2,2]. Berdasarkan kriteria *logit*, tingkat kesukaran dapat dikelompokkan dalam 3 kategori yaitu:

1) sukar; dengan indeks kesukaran > 2,00
2) sedang; dengan indeks kesukaran -2,00 sampai dengan 2,00
3) mudah; dengan indeks kesukaran < -2,00.

Rangkuman hasil analisis disajikan pada Tabel 6. Dari hasil analisis menunjukkan sebanyak 37 butir soal (92,5%) mempunyai tingkat kesukaran sedang dan 3 butir soal (7,5%) mudah.

Tabel 6. Tingkat Kesukaran Butir Soal Berdasarkan Teori Respons Butir

Kategori	Soal UKK Kelas VIII		
	No Butir	Jml	%
Mudah ($b < -2$)	22,21,30	3	7,5
Sedang ($-2 \le b \le 2$)	2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16, 17,18,19.20,22,23, 24,25,26,27,28,29, 31,32,33,34,35,36, 37,38,39,40	37	92,5
Sukar ($b > 2$)	-	0	0

Peluang Tebakan Semu (c)

Berdasarkan kajian teori respons butir pada model 3 parameter dikatakan peluang tebakan semu berfungsi baik jika nilai c tidak lebih dari 0,250. Rangkuman hasil analisis dengan program ***BILOG*** disajikan pada tabel 7 berikut.

Tabel 7. Peluang Tebakan Semu Butir Soal Berdasarkan Teori Respons Butir

Kategori	Soal UKK Kelas VIII		
	No Butir	Jml	%
Baik ($c \le 0,250$)	1,2,4,9,10,12,19,20, 22,24,25,27,28,31,32, 34,35,36,37,38,40	21	52,5
Kurang Baik ($c > 0,250$)	3,5,6,7,8,11,13,14,15, 16,17,18,21,23,26,29, 30,33,39	19	47,5

Hasil analisis menunjukkan ada 19 butir soal (47,5%) mempunyai alternatif jawaban kurang baik dan 21 butir soal (52,5%) yang baik.

Kecocokan Model

Analisis kecocokan model digunakan untuk melihat apakah model logistik 3 parameter yang digunakan sesuai dengan butir soal. Butir soal yang cocok berarti berperilaku secara konsisten dengan apa yang diharapkan oleh model. Uji kococokan model menggunakan pendekatan uji Chi Square. Butir soal akan ditolak jika memiliki nilai $\chi^2_{(hitung)}$ > nilai kritis atau nilai tabel. Hasil analisis kecocokan model disajikan pada Tabel 10 berikut.

Hasil analisis menunjukkan, ada 1 (satu) butir soal yang tidak cocok dengan model logistik 3 parameter yaitu butir soal nomor 1.

Butir-butir yang lainnya cocok dengan pemilihan model logistik 3 parameter.

Keberfungsian Butir Diferensial

Untuk mendeteksi keberfungsian butir diferensial, pada analisis ini menggunakan metode Mantel-Haenszel dengan bantuan program *Bilog MG* fase tiga. Program *Bilog MG* fase tiga menghasilkan output informasi kemampuan siswa. Kemampuan siswa tersebut selanjutnya dikelompokkan menjadi 5 (lima), seperti disajikan pada tabel 6. Pengelompokan tersebut selanjutnya digunakan untuk menghitung besarnya nilai Mantel-Haenszel dengan bantuan program *Excel*. Hasil analisis menunjukkan, dari 40 butir soal terdapat 3 (tiga) butir yang mengandung DIF, yaitu butir nomor 9, 15 dan 35. Butir 9 merupakan materi unsur dan bagian-bagian lingkaran. Materi ini lebih bersifat hafalan. Butir 15 menguji materi tentang hubungan sudut pusat dan sudut keliling. Butir 35 menghitung luas permukaan dan volume prisma. Butir soal nomor 9 dan 35 menguntungkan jenis kelamin perempuan. Butir soal nomor 15 menguntungkan jenis kelamin laki-laki.

Deskripsi hasil penelitian di atas selanjutnya dapat untuk menjawab pertanyaan penelitian berikut.

1. Tingkat kesukaran butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ditinjau dari teori tes klasik adalah 42,5% dalam kategori mudah dan 57,5% dalam kategori sedang. Menurut teori respons butir dari 40 butir soal, 92,5% mempunyai tingkat kesukaran dengan kategori sedang dan 7,5% kategori mudah.
2. Daya pembeda butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ditinjau dari teori tes klasik dan teori respons butir semuanya dalam kategori baik (100%).
3. Efektivitas distraktor butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ditinjau dari teori tes klasik 97,5% berfungsi dengan baik dan 2,5% kurang berfungsi.
4. Efektivitas parameter faktor tebakan butir perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ditinjau dari teori respons butir 52,5% berfungsi dengan baik dan 47,5% kurang berfungsi dengan baik.
5. Keberfungsian butir diferensial perangkat tes matematika Ulangan Kenaikan Kelas VIII jenjang SMP di Kabupaten Sleman tahun pelajaran 2011/2012 ditinjau dari pendekatan teori tes klasik 7,5% mengandung *DIF*. Artinya masih ada 3 butir soal yang mengandung *DIF*.

Secara umum, hasil analisis butir soal baik dengan teori tes klasik maupun teori respons butir hampir sama. Analisis butir soal tentang daya beda memberikan hasil yang sama, yaitu semua butir soal mempunyai daya beda yang baik. Artinya semua butir soal sudah berfungsi baik, yaitu dapat memberikan informasi seberapa besar kemampuan soal tersebut dapat membedakan peserta tes yang memperoleh jumlah skor yang tinggi dan peserta tes yang jumlah skornya rendah. Daya pembeda dapat digunakan untuk melihat kemampuan butir soal itu dalam membedakan peserta yang mampu dan yang tidak mampu memahami materi yang ditanyakan dalam butir tersebut. Semakin besar nilai daya pembeda butir soal berarti semakin besar kemampuan butir soal itu membedakan peserta yang mampu dan yang tidak mampu.

Hasil analisis tentang tingkat kesukaran butir soal antara teori tes klasik dan teori respons butir disajikan pada tabel 8. Dari tabel 8 menunjukkan, ada 17 butir soal dengan kategori mudah dengan teori tes klasik sedangkan dengan teori respons butir ada 3 soal. Hal ini disebabkan perbedaan penghitungannya. Penghitungan tingkat kesukaran respons butir pada teori tes klasik semata-mata didasarkan pada siswa menjawab benar. Artinya butir soal yang mampu dijawab oleh sebagian besar siswa dianggap sebagai butir yang mudah. Pada teori respons butir penghitungannya berdasarkan kemampuan siswa.

Dari tabel 9 diperoleh ada perbedaan yang signifikan hasil analisis tentang alternatif jawaban. Hal ini disebabkan oleh dasar perhitungannya. Teori tes klasik memandang, alternatif jawaban sudah berfungsi jika masing-masing alternatif jawaban ada yang memilih. Sementara teori respons butir mendasarkan perhitungan pada rumus model logistik 3 parameter. Hasil analisis tentang alternatif jawaban disajikan pada tabel 9 berikut.

Tabel 8. Perbandingan Tingkat Kesukaran antara Teori Tes Klasik dan Teori Respons Butir

Kategori	Klasik Butir	Jml	Respons Butir Butir	Jml
Mudah	1,2,3,4,7, 9,10,11,12, 13,16,20,21, 25,29,30,33	17	2,21,30	3
Sedang	5,6,8,14,15, 17,18,19,22, 23,24,26,27, 28,31,32,34, 35,36,37,38, 39, 40	23	1,3,4,5,6,7, 8,9,10,11,12, 13,14,15,16, 17,18,19,20, 22,23,24,25, 26,27,28,29, 31,32,33,34, 35,36,37,38, 39,40	37
Sukar	-		-	

Tabel 9. Perbandingan Hasil Analisis Alternatif Jawaban antara Teori Tes Klasik dan Teori Respons Butir

Kategori	Klasik Butir	Jml	Respons Butir Butir	Jml
Baik	1,2,3,4,5,6,7, 8,9,10,11,12, 13,14,15,16, 17,18,19,20, 21,22,23,24, 25,26,28,29, 30,31,32,33, 34,35,36,37, 38,39,40	39	1,2,4,9,10,12, 19,20,22,24, 25,27,28,31, 32,34,35,36, 37,38,40	21
Kurang baik	27	1	3,5,6,7,8,11,13, 14,15,16,17,18, 21,23,26,29,30, 33,39	19

Dari hasil analisis menggunakan pendekatan tes klasik, ada 17 butir soal yang mempunyai tingkat kesukaran mudah (nomor 1, 2, 3, 4, 7, 9, 10, 11, 12, 13, 16, 20, 21, 25, 29, 30 dan 33), dan 23 butir soal yang tingkat kesukaran sedang yaitu butir soal nomor 5, 6, 8, 14, 15, 17, 18, 19, 22, 23, 24, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39, 40. Jika dilihat dari daya pembeda, semua (40) butir soal dalam kategori baik. Jika akan dikembangkan bank soal berdasarkan teori tes klasik, dapat dipilih butir soal yang memiliki daya beda baik dan tingkat kesukaran sedang. Pada perangkat Ulangan Kenaikan Kelas ini, ada 23 butir soal yang memenuhi kriteria ini, yaitu butir soal nomor 5, 6, 8, 14, 15, 17, 18, 19, 22, 23, 24, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39, 40.

Simpulan dan Saran

Simpulan

Analisis kuantitatif (empiris) butir soal dengan pendekatan teori tes klasik menggunakan program *ITEMAN* diperoleh informasi indeks keandalan tes sebesar 0,858 dengan *SEM* 2,577 yang berarti tes andal dalam melakukan fungsi ukurnya. Sebanyak 23 butir soal (57,5%) mempunyai tingkat kesukaran yang baik, 40 butir soal (100%) mempunyai daya beda yang baik serta 39 butir soal (97,5%) memiliki pengecoh yang baik. Hasil ini menunjukkan bahwa kualitas tes Ulangan Kenaikan Kelas VIII mata pelajaran Matematika SMP tahun pelajaran 2011/2012 di Kabupaten Sleman menurut analisis teori tes klasik dikategorikan baik.

Analisis kuantitatif (empiris) butir soal dengan pendekatan teori respons butir Model Logistik 3 Parameter menggunakan program *BILOG* sebanyak 40 butir soal (100%) mempunyai daya beda yang baik, sebanyak 37 butir soal (92,5%) mempunyai tingkat kesukaran yang baik, sebanyak 21 butir soal (52,5%) mempunyai peluang tebakan semu yang baik, serta sebanyak 39 butir soal (97,5%) cocok dengan model logistik 3 parameter. Hasil ini menunjukkan bahwa kualitas tes Ulangan Kenaikan Kelas VIII mata pelajaran Matematika SMP tahun pelajaran 2011/2012 di Kabupaten Sleman menurut analisis teori respons butir dikategorikan baik.

Ditinjau dari keberfungsian butir diferensial, dari 40 butir soal terdapat 3 (tiga) butir yang mengandung *DIF*, yaitu butir nomor 9, 15 dan 35. Butir 9 merupakan materi unsur dan bagian-bagian lingkaran. Materi ini lebih bersifat hafalan. Butir 15 menguji materi tentang hubungan sudut pusat dan sudut keliling. Butir 35 menghitung luas permukaan dan volume prisma.

Saran

Hasil analisis menggunakan metode tes klasik (*ITEMAN*) dan metode respons butir (*BILOG*) secara bersama-sama dalam analisis butir soal secara empiris, cukup konsisten. Namun, disarankan untuk ujian yang berskala besar seperti Ulangan Kenaikan Kelas (UKK), sebaiknya menggunakan metode respons butir karena akan memberi informasi yang lebih banyak, dan tidak tergantung pada sampel yang digunakan.

Pengadaan bank soal untuk mendukung tersedianya butir soal yang bermutu di daerah perlu diupayakan. Butir soal yang dinyatakan baik secara teoritis dan empiris dalam penelitian ini, dapat digunakan untuk keperluan pengembangan bank soal.

Jika terpaksa menggunakan butir soal yang memuat DIF, perlu diseimbangkan jumlah butir yang menguntungkan suatu kelompok siswa berdasarkan jenis kelamin, sehingga dapat diperoleh tes yang adil.

Dinas Pendidikan, Pemuda dan Olah Raga (Disdikpora) Kabupaten Sleman, hendaknya menetapkan soal-soal yang sudah dikalibrasi untuk dirakit menjadi naskah Ulangan Kenaikan Kelas (UKK), sehingga dapat diketahui karakteristik butir soal yang digunakan.

Lembaga Penjaminan Mutu Pendidikan (LPMP) yang mempunyai tugas memberikan bantuan teknis kepada satuan pendidikan dapat bekerjasama dengan dinas pendidikan di daerah, maupun perguruan tinggi untuk meningkatkan kemampuan guru dalam menulis soal, melalui kegiatan pendidikan dan latihan (diklat), workshop, dan lokakarya.

Daftar Pustaka

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Montery, CA: Brook/Cole Publishing Company.

Allen N.C., Holland, P.W. (1993). *Differential item functioning: A model for missing information about the group membership of examinees in DIF studies*. New Jersey : Lawrence Erbaum Associates, Publishers.

Aziz, N.A. (2011). *Evaluation of Mantel-Haenszel statistic for detecting differential item functioning*. Educare: International Journal for Educational Studies, 3(2).

Azwar, Saifuddin. (2007). *Dasar-dasar psikometri*. (Edisi ke-1). Yogyainkarta: Purminatstaka Pelajar.

Azwar, Saifuddin.(2007). *Reliabilitas dan validitas*. (Edisi ke-3). Yogyakarta: Pustaka Pelajar.

Blerkom, M. L. V. (2009). *Measurement and statistics for teachers.* New York: Routledge.

Camilli, G. (1993). *The case against item bias detection techniques based on internal criteria: do item bias procedures obscure test fairness issues?* New Jersey: Lawrence Erbaum Associates, Publishers.

Frenc, A.W. & Miller, T.R. (1996). Logistic regretion and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*. 33(3). 315-332.

Gierl, M., Khaliq, S.N. & Boughton, K. (1999). Gender Differential Item Functioning in Mathematics and Science : Prevalence and Policy Implications. *Makalah dalam symposium Improving Large-Scale Assessment in Education*. Diambil pada tanggal 13 Desember 2013, dari http://www.ncrel.org/sdrs/.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and Aplications.* Boston: Kluwer Nijjhoff, Publisher.

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

Kartowagiran, Badrun. (2005). *Perbandingan berbagai metode untuk mendeteksi bias butir*. Disertasi doktor, tidak diterbitkan, Universitas Gadjah Mada, Yogyakarta.

Kerlinger, F.N. (1986). *Asas-asas penelitian behavioral*. (Terjemahan L.R. Simatupang). Yogyakarta: Gajah Mada University. Press.

Kubiszyn, T., Borich, G. (2010). *Educational testing and measurement: classroom application and practice (9th ed)*. USA: John Wiley & Sons. Inc.

Lamprianou I., Athanasou J.A. (2009). *A teacher's guide to educational assessment (revised edition)*. Rotterdam: Sense Publishers.

Mardapi, Djemari. (1999). *Estimasi kesalahan pengukuran dalam bidang pendidikan dan impikasinya pada ujian nasional*. Yogyakarta, Pidato pengukuhan guru besar.

Mardapi, Djemari. (2004). *Penyusunan tes hasil belajar*. Yogyakarta: Universitas Negeri Yogyakarta.

Mardapi, Djemari. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta : Mitra Cendekia.

Michaelides, Michalis P. (2008). *An Illustration of Mantel-Haenszel Procedur to Flag Misbehaving Common Items in Test Equating. Practical Assessment Research & Evaluation*, 13(7). http:/ pareonline.net/getvn.asp?v=13&n=7

Nitko , A & Brookhart (2007). *Educational Assessment of Students*. New Jersey : Pearson Education, inc.

Surapranata, Sumarna (2006). *Analisis, validitas, reliabilitas dan interpretasi hasil tes implementasi kurikulum 2004*. Bandung : Remaja Rosdakarya.

Surapranata,Sumarna (2005). *Panduan penulisan tes tertulis implementasi kurikulum 2004*. Bandung : Remaja Rosdakarya.

Retnawati, Heri. (2003). *Keberfungsian butir diferensial pada perangkat tes seleksi masuk sekolah lanjutan tingkat pertama (SLTP) mata pelajaran matematika*, Tesis magister, tidak diterbitkan, Universitas Negeri Yogyakarta.

Reynolds, C.R., Livingston, R.B. & Willson, V. (2010). *Measurement and assessment in education*. New Jersey : Pearson Education, Inc.

Wardhani, Sri. (2010). *Implikasi karakteristik matematika dalam pencapaian tujuan mata pelajaran matematika di SMP/ MTs*. Yogyakarta: PPPPTK Matematika

Worthen, B.R. & Sanders, J.R. (1973). *Educational evaluation: theory and practice*. California: Wadsworth Publishing Company

Wright, R. J. (2008). *Educational assessment: test and measurements in the age of accountability*. London: Sage Publications, Inc.