PENGEMBANGAN PERANGKAT TES *TRY OUT* UN SMP BIDANG STUDI MATEMATIKA UNTUK PEMBENTUKAN BANK SOAL

Fieka Nurul Arifa, Djemari Mardapi
Prodi Penelitian dan Evaluasi Pendidikan PPs UNY, Universitas Negeri Yogyakarta
fiekanarifa@gmail.com, djemarimardapi@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengembangkan butir soal, mengetahui karakteristik dan melakukan penyetaraan butir soal pada perangkat soal *Try Out* UN SMP Bidang Studi Matematika untuk pembentukan bank soal. Model pengembangan yang digunakan adalah model prosedural. Sampel sebanyak 1.140 respons peserta didik diperoleh melalui *stratified propotional random sampling*. Perangkat tes yang diujikan meliputi 4 paket soal yang setiap paketnya berisi 40 butir soal dengan 8 butir *anchor*. Analisis butir soal dilakukan melalui analisis kualitatif dan analisis kuantitatif. Analisis kualitatif didasarkan pada kesesuaian butir soal dengan lembar telaah berdasarkan kriteria materi, konstruksi, dan bahasa. Analisis kuantitatif menggunakan Teori Respons Butir dengan bantuan program BILOG-MG 3.0. Hasil penelitian menunjukkan bahwa 107 butir soal memiliki karakteristik baik terdiri dari 26 butir soal P1, 25 butir soal P2, 24 butir soal P3, 24 butir soal P4, dan 8 butir soal *anchor*.

***Kata kunci**: bank soal, karakteristik butir soal, penyetaraan butir soal*

DEVELOPING A SET TEST FOR NATIONAL EXAM'S TRY OUT OF MATHEMATICS STUDY IN JUNIOR HIGH SCHOOL TO ESTABLISH A TEST ITEM BANK

Fieka Nurul Arifa, Djemari Mardapi
Prodi Penelitian dan Evaluasi Pendidikan PPs UNY, Universitas Negeri Yogyakarta
fiekanarifa@gmail.com, djemarimardapi@gmail.com

Abstract

The aim of this study is to develop item tests, identify the characteristics, and equate the test items of National Exam Try Out of Mathematics Study in Junior High School to establish a test item bank. The development model used in this study is procedural model. The sample comprised 1,140 students' responses obtained by stratified propotional random sampling. A set of the test that was tested to the students consist of four test packages, each of which contains 40 test items with 8 common items. The analysis of the test items was conducted through qualitative and quantitative technique. The qualitative analysis was based on the conformity of test items with review sheets on the material criteria, construction, and language. The quantitative analysis use Item Response Theory supported by Bilog MG 3.0 program. The results of the study show that 107 item test have good characteristics, it contains of 26 items in package 1, 24 items in package 2, 24 items in package 3, 24 items in package 4, and 8 common items.

***Keywords**: item bank, characteristics of test items, equation of test items*

Pendahuluan

Pendidikan merupakan proses untuk membantu manusia dalam mengembangkan dirinya agar mampu menghadapi setiap perubahan yang terjadi dalam kehidupan. Perkembangan era globalisasi yang semakin pesat menuntut setiap individu untuk dapat beradaptasi dan bersaing dalam masyarakat. Pendidikan menjadi sarana dan wadah dalam pembinaan sumber daya manusia. Pembinaan sumber daya manusia salah satunya dilaksanakan melalui kegiatan pembelajaran yang yang diselenggarakan oleh lembaga pendidikan. Oleh karena itu pendidikan perlu mendapatkan perhatian dan penanganan baik dari pemerintah, masyarakat, keluarga dan terutama dari lembaga pendidikan.

Pendidikan di Indonesia dari tahun ke tahun selalu mengalami perbaikan. Berbagai upaya telah dilakukan oleh pemerintah untuk meningkatkan kualitas pendidikan, diantaranya melalui perubahan kurikulum yang diharapkan dapat menjadi panduan untuk meningkatkan kualitas belajar mengajar, penerapan program Manajemen Berbasis Sekolah (MBS), pelatihan tenaga pendidik, program Bantuan Operasional Sekolah (BOS), dan lain sebagainya. Akan tetapi upaya tersebut belum memberikan dampak yang memuaskan terhadap peningkatan kualitas pendidikan di Indonesia, apalagi dibandingkan dengan kualitas pendidikan dalam skala internasional.

Kualitas pendidikan merupakan hasil dari serangkaian proses yang melibatkan seluruh komponen sistem pendidikan. Salah satu faktor penting untuk meningkatkan kualitas pendidikan adalah program pembelajaran, dan evaluasi merupakan salah satu komponen penting dalam program pembelajaran untuk mengetahui sejauhmana pencapaian keberhasilan pembelajaran dan mengetahui aspek mana saja yang perlu diperbaiki. Untuk memantau kualitas pendidikan, diperlukan kegiatan evaluasi yang dilaksanakan secara berkesinambungan. Salah satunya dilakukan melalui kegiatan penilaian hasil belajar peserta didik pada tiap jenjang pendidikan.

Kegiatan penilaian hasil belajar dapat dilaksanakan melalui pengujian terhadap penguasaan kompetensi peserta didik pada setiap mata pelajaran. Kegiatan tersebut dilakukan secara internal maupun eksternal. Penilaian secara internal dilakukan melalui kegiatan ulangan harian, ulangan tengah semester, serta ulangan kenaikan kelas untuk memantau perkembangan peserta didik. Penilaian secara eksternal dilakukan melalui Ujian Nasional (UN) atau pengujian yang melibatkan pihak eksternal sekolah/ lembaga pendidikan. Dalam kegiatan pengujian diperlukan perangkat evaluasi sebagai alat ukur.

Tes merupakan salah satu perangkat evaluasi untuk mengukur pencapaian kompetensi dasar pada setiap mata pelajaran yang telah ditempuh dalam waktu yang tertentu. Alat ukur berupa perangkat tes prestasi belajar harus disusun berdasarkan tes standar agar hasilnya dapat dipertanggungjawabkan. Tes standar diperoleh melalui upaya sistematis, mulai dari penyusunan kisi-kisi, penulisan butir soal, penelaahan butir soal, ujicoba paket tes, analisis empirik hasil ujicoba dan kalibrasi soal, hingga sistem pengelolaan. Tes standar merupakan alat ukur yang akurat untuk mengetahui kemampuan peserta didik. Dengan menggunakan tes standar maka hasil-hasil pengujian dapat diperbandingkan lintas daerah dan lintas waktu. Pengembangan tes standar harus terus-menerus dilakukan dalam rangka persiapan pengujian yang baik.

Sebagai pelaksana pendidikan, guru terlibat langsung dalam menyiapkan soal-soal untuk menyusun perangkat tes dan menganalisis butir soal sesuai dengan prinsip, mekanisme, dan prosedur penilaian. Hal ini dikarenakan guru bertanggungjawab dalam pemantauan proses, kemajuan, dan perbaikan hasil belajar anak didik di kelas melalui pengujian secara internal. Namun, pada kenyataanya pengembangan soal yang disusun oleh guru masih sangat terbatas. Karena alasan kepraktisan sekolah-sekolah bekerjasama dengan lembaga eksternal seperti Lembaga Bimbingan Belajar (Bimbel) dalam melakukan pengujian atau untuk memperbanyak latihan soal bagi peserta didik. Sementara dalam hal ini soal-soal yang diperoleh dari Bimbel belum tentu soal-soal tes standar yang sudah diketahui memiliki karakteristik baik.

Perangakat tes yang baik harus memenuhi kriteria tes standar yang telah ditentukan. Persyaratan atau kriteria tersebut meliputi: 1) validitas, yakni perangkat tes harus benar-benar mengukur apa yang hendak diukur, 2) reliabilitas, perangkat tes *reliable* bila menunjukkan ketepatan hasilnya, 3) objektivitas, suatu perangkat tes harus benar-benar mengukur apa yang diukur, tanpa adanya interpretasi yang tidak ada hubungannya dengan perangkat

tes tersebut, 4) efisien, suatu perangkat tes sedapat mungkin dipergunakan tanpa membuang waktu dan biaya yang banyak, dan 5) Kegunaan/kepraktisan, perangkat tes harus mudah digunakan dan berguna.

Perangkat tes yang baik menggunakan butir-butir soal tes standar atau tes yang memiliki karakteristik baik sebagai komponen penyusunnya. Terkait dengan otonomi daerah, setiap daerah perlu membuat butir-butir tes standar untuk setiap mata pelajaran. Butir-butir tes ini kemudian disimpan dan dikelola dalam bank soal. Bank soal menjadi wadah untuk menyimpan butir-butir tes yang telah dibakukan, agar ketika diperlukan dapat dengan mudah diperoleh. Pada bank soal akan memungkinkan pengembangan tes standar untuk pengujian internal di sekolah-sekolah pada setiap mata pelajaran. Selain itu sistem pengujian dengan bank soal dapat menjamin terselenggaranya sistem pengujian yang fleksibel, ekonomis, konsisten, adil, aman, berkesinambungan, serta diharapkan hasil-hasil ujian dapat diperbandingkan.

Pelaksanaan UN merupakan bentuk pertanggung jawaban lembaga pendidikan terhadap pemerintah pada pencapaian kompetensi siswa sebagai hasil dari kegiatan belajar dalam periode waktu yang telah ditentukan. Perolehan nilai UN yang masih rendah menunjukkan bahwa usaha dalam mempersiapkan UN belum maksimal. Salah satu upaya untuk mempersiapkan UN adalah melalui pengujian-pengujian menggunakan perangkat tes uji coba (*try out*) UN. Perangkat tes *try out* UN merupakan perangkat tes yang karakteristiknya disesuaikan dengan perangkat tes standar UN.

Untuk dapat melakukan pengukuran secara tepat maka perlu dibahas beberapa pengertian berikut.

Hakikat belajar matematika

Matematika merupakan ilmu pengetahuan yang memiliki peranan penting dalam dunia pendidikan di Indonesia. Hal ini terbukti dengan dijadikannya matematika sebagai mata pelajaran yang diajarkan dalam setiap jenjang pendidikan, mulai dari sekolah dasar sampai tingkat perguruan tinggi. Materi matematika yang dipelajari di sekolah dipilih sedemikian rupa agar mudah dialihfungsikan kegunanannya dalam kehidupan siswa yang mempelajarinya.

Pada intinya tujuan siswa belajar matematika di sekolah adalah agar mampu menggunakan atau menerapkan matematika yang dipelajari untuk memecahkan masalah dalam kehidupan sehari-hari, bekal belajar matematika lebih lanjut dan bekal belajar pengetahuan lain. Salah satu tugas sekolah adalah berupaya sebaik mungkin untuk mengajar siswa untuk berpikir, dan dari semua mata pelajaran tidak ada yang lebih cocok untuk ini daripada matematika (Dudley, 1997, P. 363).

Pada Standar Isi Mata Pelajaran Matematika Tahun 2006 untuk semua jenjang pendidikan dinyatakan bahwa mata pelajaran matematika dipelajari dengan tujuan agar peserta didik memiliki kemampuan: 1) memahami konsep matematika, menjelaskan keterkaitan antarkonsep dan mengaplikasikan konsep atau algoritma, secara luwes, akurat, efisien, dan tepat dalam pemecahan masalah; 2) menggunakan penalaran pada pola dan sifat, melakukan manipulasi matematika dalam membuat generalisasi, menyusun bukti, atau menjelaskan gagasan dan pernyataan matematika; 3) memecahkan masalah yang meliputi kemampuan memahami masalah, merancang model matematika, menyelesaikan model dan menafsirkan solusi yang diperoleh. 4) mengomunikasikan gagasan dengan simbol, tabel, diagram, atau media lain untuk memperjelas keadaan atau masalah; dan 5) Memiliki sikap menghargai kegunaan matematika dalam kehidupan, yaitu memiliki rasa ingin tahu, perhatian, dan minat dalam mempelajari matematika, serta sikap ulet dan percaya diri dalam pemecahan masalah.

Teori Tes

Tes merupakan alat untuk memperoleh data tentang perilaku individu (Allen & Yen, 1979, P.1). Sementara itu AERA, APA & NCME (Reynolds, Livingston, & Willson, 2010, P.3), menjelaskan tes sebagai suatu prosedur dimana sampel perilaku dari individu didapatkan, dievaluasi, dan dinilai menggunakan prosedur standar. Tes terdiri atas sejumlah pertanyaan yang memiliki jawaban benar atau salah, atau semua benar, atau sebagian benar (Mardapi, 2012, p.108). Berdasarkan definisi di atas, maka dapat disimpulkan bahwa tes adalah cara yang dipergunakan atau prosedur yang ditempuh dalam pengukuran dan penilaian sehingga dihasilkan skor yang menggambarkan tingkah laku atau kemampuan individu.

Soal pilihan ganda merupakan bentuk soal yang jawabannya dapat dipilih dari be-

berapa kemungkinan jawaban yang telah disediakan. Butir pilihan ganda umumnya terdiri atas satu kalimat pertanyaan, yang disebut stem, dan beberapa pilihan jawaban yang disebut alternatif atau options. Salah satu diantar alternatif tersebut merupakan jawaban yang benar atau yang terbaik dan disebut *key* atau kunci jawaban, sedangkan alternatif-alternatif lainnya adalah jawaban yang disebut pengecoh atau distraktor.

Analisis soal dilakukan untuk mengetahui berfungsi tidaknya sebuah soal. Analisis butir tes dilakukan secara kualitatif maupun kuantitatif. Pada analisis kualitatif butir tes atau yang biasa disebut telaah ahli memperhatikan materi, konstruksi, dan bahasa dalam penulisan butir soal. Analisis materi (isi) dimaksudkan sebagai penelaahan khusus yang berkaitan dengan kelayakan pengetahuan yang ditanyakan. Analisis konstruksi (teknis) dimaksudkan sebagai penelaahan yang berkaitan dengan prinsip-prinsip pengukuran dan teknik penulisan soal. Analisis bahasa (editorial) dimaksudkan sebagai penelaahan yang berkaitan dengan penggunaan bahasa Indonesia dengan baik dan benar menurut EYD. Selain itu, penelaahan bahasa juga berkaitan dengan keseluruhan format dan keajegan editorial dari soal yang satu ke soal yang lainnya. Analisis kuantitatif dimaksudkan untuk mengetahui kualitas tes secara empiris. Pada soal pilihan ganda analisis butir bertujuan untuk meneliti tanggapan siswa terhadap item tes individu, menilai kualitas dari butir-butir dan tes secara keseluruhan, serta untuk meningkatkan kualitas/ merevisi butir dan tes (Gajjar, et al, 2014, p.17).

Analisis butir adalah prosedur sederhana namun penting dilakukan pengujian untuk menyediakan informasi mengenai reliabilitas dan validitas item/ test dengan menghitung daya beda, tingkat kesukaran, dan keberfungsian distraktor dan keterkaitan didalamya (Gajjar, 2014, p.20). Berdasarkan hasil analisis dapat diketahui karakter soal. Hasil analisis sangat membantu dalam pengelompokan dan perbaikan butir soal.

Teori Respon Butir

Teori respon butir dikembangkan berdasarkan dua postulat (Hambleton, Swaminathan & Rogers,1991, p.7), yaitu: 1) prestasi peserta uji pada suatu tes dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latens traits*), trait adalah dimensi kemampuan seseorang seperti kemampuan verbal, kemampuan psikometer, kemampuan kognitif, dan sebagainya, dan 2) hubungan antara prestasi uji pada suatu butir tes dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu yang disebut kurva karakteristik butir (*item characteristic curve*).

Secara umum ciri-ciri teori-respons butir adalah sebagai berikut: (1) karakteristik butir tidak tergantung pada peserta ujian, (2) skor yang digambarkan peserta ujian tidak tergantung pada tes, (3) merupakan model yang lebih menekankan pada tingkat butir daripada tingkat tes, (4) merupakan model yang tidak mensyaratkan secara ketat tes paralel untuk menaksir reliabilitas, dan (5) merupakan model yang menguraikan sebuah ukuran keputusan untuk tiap skor kemampuan yakni ada hubungan fungsional antara peserta ujian terhadap tingkat kemampuan yang dimiliki. Beberapa asumsi yang melandasi teori respons butir (Hambleton, Swaminathan, & Rogers, 1991, pp.9–12), yaitu 1) satu dimensi *(unidimensional),* artinya dimensi karakter peserta yang diukur oleh suatu tes tunggal; dan 2) kebebasan lokal *(local independence),* artinya respon peserta tes terhadap suatu butir tidak berhubungan dengan butir lainnya dalam tes tersebut.

Teori respons-butir juga mempunyai suatu sifat yang disebut invariansi parameter butir dan parameter kemampuan (*invariance of item and ability parameter).* Sifat ini memberikan pengertian bahwa parameter karakteristik suatu butir tes tidak tergantung pada distribusi kemampuan peserta dan parameter karakteristik kemampuan peserta tidak tergantung pula pada karakter tes yang digunakan.

Pada umumnya dalam teori respons butir digunakan model distribusi logistik. Beberapa model logistik dalam teori respon butir, diantaranya: 1) model logistik satu parameter, ditentukan oleh satu karakteristik butir yaitu tingkat kesukaran, 2) model logistik dua parameter ditentukan oleh dua karakteristik butir yaitu tingkat kesukaran dan daya pembeda, dan 3) model logistik tiga parameter ditentukan oleh tiga karakteristik butir yakni tingkat kesukaran, daya pembeda, dan faktor tebakan (Hambleton & Swaminathan, 1985, p.49), menyebutkan. Model logistik 3 parameter yang digunakan pada analisis butir memenuhi persamaan:

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

dimana: $P_i(\theta)$ probabilitas menjawab benar butir i oleh peserta berkemampuan θ, e bilangan transenden yang besarnya mendekati 2,718, a_i daya pembeda butir i, b_i tingkat kesukaran butir i, c_i tebakan semu (*pseudo guessing*) butir i, D faktor skala (1,7).

Pada teori respon butir, indeks tingkat kesukaran (b) adalah matriks yang sama dengan keahlian atau sifat (DeMars, 2010, p.4). Butir yang memiliki tingkat kesukaran tinggi maka memiliki indeks tingkat kesukaran sulit. Indeks tingkat kesukaran yang baik berada di antara -2 sampai +2 (Hambleton & Swaminathan, 1985, p.36). Semakin negatif nilai indek kesukaran menunjukkan bahwa butir soal semakin mudah, sebaliknya semakin positif nilai indek kesukaran menunjukkan bahwa butir soal semakin sukar.

Daya beda berguna untuk memilih item yang membedakan dengan baik antara ujian dengan tingkat rendah dan tinggi dari kemampuan atau sikap yang diukur dengan item tes (DeMars, 2010, p. 4). Daya beda yang tinggi dapat diartikan bahwa butir tersebut dapat membedakan antara peserta ujian dengan berbagai tingkat kemampuan. Indeks daya beda yang diterima berada di antara 0 – 2 (Hambleton & Swaminathan, 1985, p.36).

Kecocokan suatu item dengan model *(goodness of fit ststistic)* dapat dilihat dari nilai chi kuadrat item dibandingkan dengan harga kritik distribusi chi kuadrat sesuai dengan dk item yang bersangkutan pada taraf signifikansi = 0,01 atau 0,05. Butir dikatakan cocok model jika nilai chi kuadrat butir lebih kecil dari harga distribusi chi kuadrat pada nilai kritisnya.

Pada analisis teori respons butir dapat diketahui fungsi informasi tes dan kemampuan peserta didik. Pada nilai θ tertentu, nilai fungsi informasi mencapai maksimum yang berarti bahwa jika butir itu dikerjakan oleh peserta dengan "θ" (*theta*) tersebut, maka akan diperoleh informasi yang paling tinggi. Dalam IRT fungsi informasi digunakan untuk menghitung reliabilitas dan *Standard error of estimation* (SE) (DeMars, 2010, p.6). Fungsi informasi butir dengan model logistik 3 parameter memenuhi persamaan:

$$I_i(\theta) = \frac{(1-c_i)D^2 a_i{}^2}{\left[c_i + e^{Da_i(\theta - b_i)}\right]\left[1 + e^{Da_i(\theta - b_i)}\right]^2}$$

Setiap butir soal memiliki fungsi informasi dan jumlahnya merupakan fungsi informasi tes tersebut tes (Hambleton, Swaminathan & Rogers, 1991, p.94) sehingga fungsi informasi paket tes akan tinggi jika butir penyusunnya mempunyai fungsi informasi yang tinggi pula. Secara matematis fungsi informasi tes dapat ditulis:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Kesalahan pengukuran menurut teori respons butir dinyatakan dengan *Standar Error of Estimation* (*SE*) yang besarnya tergantung fungsi informasi tes. Fungsi informasi dengan kesalahan baku pengukuran *(SE*) mempunyai hubungan yang berbanding terbalik kuadratik Makin besar nilai fungsi informasi berarti SE semakin kecil dan sebaliknya. Bentuk hubungan keduanya dirumuskan sebagai:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Pada dasarnya terdapat dua metode estimasi kemampuan, yakni *Maximum Likelihood* Estimation (MLE) dan *Bayesian Maximum Estimation* (BYE). Kedua metode ini menggunakan pola respon peserta tes terhadap butir soal. Pola respon tersebut dimyatakan dengan u. u_1 = i untuk jawaban benar dan u_1 = 0 untuk jawaban salah. Estimasi kemampuan dengan metode MLE memenuhi persamaan berikut.

$$L(U|\theta) = \prod_{i=1}^{n} P_i(\theta)^{U_i} Q_i(\theta)^{1-U_i}$$

dimana U adalah vektor respons, misal U = 1 0 1 1. $P_i(\theta)$ probabilitas menjawab benar butir ke *i*, $Q_i(\theta) = 1 - P_i(\theta)$ yakni probabilitas menjawab salah butir ke *i*

Penyetaraan Perangkat Tes

Penyetaraan bentuk tes dalam pengujian skala besar tidak bisa dianggap remeh, karena melalui proses penyetaraan skor dari berbagai bentuk menjadi sebanding (Keller & Hambleton, 2013, p.390). Di samping itu pada program pengujian, terutama dalam skala besar, keamanan perangkat tes merupakan bagian yang sangat penting. Oleh karena itu penyetaraan diperlukan dalam kegiatan penyusunan beberapa perangkat tes yang setara

mengingat bahwa menyusun tes yang benar-benar paralel tidaklah mudah.

Tujuan penyetaraan adalah untuk menghasilkan hubungan antara skor pada dua bentuk tes sehingga skor dari setiap bentuk tes dapat digunakan seolah-olah berasal pengujian yang sama (Dorans et all.,2010, p3; von Davier, 2011, p.23). Persyaratan yang yang harus dipenuhi dalam penyetaraan diantaranya dua tes harus mengukur konstruk yang sama dengan tingkat kesukaran yang hampir sama dan dengan indeks reliabilitas yang sama.

Proses penyetaraan terhadap beberapa perangkat tes dapat dilakukan dengan dua cara, yaitu penyetaraan horizontal dan penyetaraan vertikal. Penyetaraan horizontal dilakukan untuk menyetarakan dua skor atau lebih yang masing-masing diperoleh dari paket soal yang berbeda tetapi mengukur karakteristik yang sama pada level yang sama. Sedangkan penyetaraan vertikal dilakukan untuk menyetarakan dua skor atau lebih yang yang masing-masing diperoleh dari paket soal yang berbeda tetapi mengukur karakteristik yang sama pada level yang berbeda.

Penyetaraan dilakukan untuk meletakkan estimasi parameter dari dua perangkat tes atau lebih pada skala yang sama. Terdapat empat rancangan atau desain penyetaraan skor di antara beberapa perangkat tes yang berbeda, yaitu: a) *single-group design*; b) *equivalent-group design*; c) *anchor-test design*; dan d) *common-person design* Hambleton, Swaminathan & Rogers, 1991, p.128).

Anchor test design sering disebut juga dengan *nonequivalent groups with anchor test* (NEAT) *design* atau *common-item nonequivalent groups design.* Pada rancangan ini masing-masing kelompok mengerjakan perangkat tes yang berbeda. Di dalam masing-masing perangkat tes terdapat beberapa item yang sama (*common item*).

Dalam rancangan ini, apabila digunakan dua perangkat tes yakni A dan B dan dua kelompok peserta yakni X dan Y, maka masing-masing perangkat tes ditambahkan item-item tes jangkar C sehingga kedua perangkat tes menjadi A + C item dan B + C item. Kelompok peserta X mengerjakan perangkat tes A + C dan kelompok Y mengerjakan B + C sehingga item-item tes anchor C (*common item*) dikerjakan oleh dua kelompok peserta tes. Pada *common items* terdapat hubungan yang linear pada tingkat kesukaran dan daya beda (Hambelton & Swaminathan 1985, p.205). Hubungan linear dari keduanya memenuhi persamaan:

$$b_y = \alpha b_x + \beta$$

$$a_y = \frac{a_x}{\alpha}$$

dimana: b_x parameter tingkat kesukaran *common* items pada kelompok X, b_y parameter tingkat kesukaran *common items* pada kelompok Y, a_x parameter daya pembeda *common items* pada kelompok X, a_y parameter daya pembeda *common items* pada kelompok Y, α, β konstanta konversi penyetaraan tes.

Penyetaraan antara dua perangkat tes atau lebih dapat dilakukan jika konstanta konversi telah diketahui. Metode penyetaraan menurut teori respons butir berfungsi untuk menentukan konstanta konversi. Nilai konversi yang dihasilkan kemudian disubstitusi dalam persamaan skala pada rancangan penyetaraan yang digunakan. Metode penyetaraan menurut teori respons butir (Hambleton, Swaminathan & Rogers, 1991, p.129), meliputi metode regresi, metode rerata dan sigma, metode rerata dan sigma tegar, dan metode kurva karakteristik. Metode kurva karakteristik Stocking dan Lord adalah metode trasnformasi yang menggunakan estimasi item perameter dari pengenaan terpisah untuk menentukan konstanta skala yang diperlukan untuk menempatkan estimasi item parameter pada skala yang sama (Keller & Hambleton, 2013, p.397).

Bank Soal

Pada umumya bank soal dipahami sebagai sekumpulan item atau butir-butir soal saja tapi sebenarnya pengertian bank soal tidaklah sebatas itu. Sebagian ahli pengukuran mendefinisikan bank soal (*item bank*) sebagai kumpulan pertanyaan-pertanyaan terkoordinasi yang dikembangkan didefinisikan, dan dikuantitatifkan sehingga memberikan definisi variable yang operasional (Wood & Skurnik, 1969, p.1). Menurut Choppin bank soal adalah kumpulan dari beberapa butir soal yang terorganisir dan terkumpul dan setiap butir soal dapat dipertanggung jawabkan dan telah diketahui karakteristiknya dalam Umar dalam (Master & Keeves, 1999, p.207). Bank soal adalah sekumpulan item yang semuanya dirancang agar relevan dangan sasaran belajar

yang sama. Bank soal memuat item-item yang dirancang agar relevan dangan sasaran belajar.

Salah satu keuntungan dari tersedianya bank soal adalah fleksibilitas (Lawrence, 1989). Tes yang diambil dari bank soal dapat dibuat panjang atau pandek, mudah atau sukar, atau disesuaikan dengan tingkat kemampuan tertentu. Penggunaanya pun tidak hanya sebatas untuk tes diagnostik, tetapi juga dapat dipergunakan untuk penilaian bersifat komprehensif dalam bentuk ujian akhir. Menurut Umar keuntungan-keuntungan yang dapat diperoleh dengan adanya pengembangan bank soal meliputi: a) kebijakan desentralisasi pada program tes nasional dapat dikenalkan tanpa mengorbankan hasil pengukuran dan dapat dibandingkannya hasil tes; b) biaya dan waktu yang diperlukan pada kegiatan konstruksi tes dapat direduksi; c) makin besar jumlah butir soal yang terdapat pada bank soal, permasalahan keamanan menjadi lebih terjamin; d) kualitas program tes dapat ditingkatkan, dengan adanya butir-butir dalam bank soal yang telah diketahui karakteristiknya; e) pendidik dapat mendesain perangkat tes yang akan digunakan, dengan memanfaatkan butir-butir yang baik dalam bank soal; serta f) guru dapat berkonsentrasi pada usaha untuk meningkatkan kualitas pembelajaran, tanpa harus membelanjakan banyak waktu untuk penyusunan perangkat tes (Suyata, Mardapi & Kartowagiran, 2010, p.48).

Metode Penelitian

Penelitian ini termasuk dalam jenis penelitian pengembangan. Fokus utama penelitian ini adalah kegiatan pengembangan butir-butir soal standar sehingga dapat dimasukkan dalam bank soal. Butir-butir soal standar yang dimaksud adalah butir-butir soal yang memiliki karakteristik sesuai dengan kriteria yang ditentukan.

Model pengembangan dalam penelitian ini menggunakan model prosedural, yakni model yang bersifat deskriptif, menunjukkan langkah-langkah untuk menghasilkan produk. Prosedur pengembangan dalam penelitian ini mengadaptasi dari beberapa model pengembangan. Desain diadaptasi secara operasional, yaitu: 1) penyusunan kisi-kisi soal; 2) penulisan soal berbentuk pilihan ganda; 3) telaah soal dari segi materi, konstruksi, dan bahasa; 4) uji prapenelitian; 5) analisis hasil uji prapenelitian 6); uji penelitian; 7) kalibrasi butir soal dengan menggunakan Program BILOG-MG 3.0. Pada penelitian juga terdapat proses penyetaraan, hal tersebut dikarenakan terdapat 4 (empat) paket soal dengan kisi-kisi yang sama yang akan digunakan dalam penelitian.

Uji coba produk dalam penelitian ini melalui dua tahapan, yakni uji coba prapenelitian (uji skala kecil), dan uji coba produk (uji skala besar). Uji coba prapenelitian dikenakan kepada sekelompok kecil sampel pada situasi yang sebenarnya. Uji coba dilakukan dengan memberikan perangkat tes pada peserta didik dan guru Mata Pelajaran Matematika untuk mengetahui keterbacaan perangkat tes. Butir soal baik pada tahap ini akan diujikan kembali pada uji berikutnya dengan jumlah subjek yang lebih banyak.Uji coba produk dikenakan pada subjek yang lebih banyak dari uji coba pra penelitian. Uji coba dilakukan dengan memberikan perangkat tes untuk dikerjakan oleh siswa.

Subjek coba dalam uji coba skala kecil terdiri dari 5 guru Mata Pelajaran Matematika dan 64 Siswa kelas IX SMP di Kabupaten Magelang yang diambil secara acak. Subjek coba dalam uji coba skala besar adalah siswa kelas IX SMP/MTs di Kabupaten Magelang sebanyak 1.140 siswa sebagai peserta tes yang ditentukan dengan teknik *stratified propotional random sampling* berdasarkan data peringkat sekolah berdasarkan nilai UN 2012/ 2013. Gambaran subjek uji coba penelitian dapat dilihat pada tabel berikut.

Tabel 1. Peringkat Sekolah pada UN 2012/ 2013, Nama Sekolah, dan Jumlah Peserta Uji Coba.

Peringkat Sekolah	Nama Sekolah	Paket Soal			
		P1	P2	P3	P4
Tinggi	SMPN 1 Salaman	36	36	34	34
	SMPN 1 Tempuran	45	46	46	45
Sedang	SMPN 3 Muntilan	51	50	48	51
	SMP Muh. Tempuran	28	29	28	30
	SMP Muh. Sambak	7	8	8	7
	SMP Salaman 1953	17	15	17	19
	MTs Walisongo Kajoran	16	16	20	16
Rendah	SMPN 1 Kajoran	54	51	52	52
	SMP Muh. Borobudur	20	20	20	20
	SMP Purnama Tempuran	12	11	12	13
Jumlah		286	282	285	287

Pengumpulan data menggunakan teknik tes. Instrument pengumpulan data berupa kisi-kisi soal, perangkat soal, lembar telaah yang akan digunakan untuk telaah analisis

dengan *expert judgement*, dan lembar respons jawaban siswa terhadap perangkat soal.

Analisis data dalam penelitian ini berupa analisis secara teoritis dan analisis empiris. Analisis secara teoritis berupa analisis kualitatif melalui telaah butir soal oleh tim ahli (*expert judgement*) dan secara empiris berupa deskriptif kuantitatif dan analisis dengan teori respon butir menggunakan bantuan program komputer *BilogMG*.

Telaah pada butir soal merupakan penilaian terhadap setiap butir soal dengan kriteria telaah berdasarkan aspek materi, konstruksi dan bahasa. Telaah dilakukan oleh Telaah dilakukan oleh tida orang ahli, dua orang praktisi dari guru Matematika SMP yang telah memiliki pengalaman dan kompeten di bidangnya dan satu orang *expert* dari dosen matematika. Setiap penelaah melakukan telaah terhadap perangkat soal yang terpisah, dengan cara mengisi kartu telaah yang telah peneliti buat sesuai dengan penduan dari depdiknas dengan cara memberi tanda pada kolom pernyataan tabel telaah. Selanjutnya peneliti membuat rangkuman hasil telaah dan membuat kesimpulan terhadap butir soal yang memenuhi kriteria pada kartu telaah tersebut. Analisis Butir Soal secara Kuantitatif

Hasil uji coba prapenelitian dianalisis menggunakan analisis statistik deskriptif. Teknik analisis deskriptif bertujuan untuk menggambarkan karakteristik data pada variabel-variabel penelitian yang ada sehingga akan memudahkan dalam kegiatan analisis selanjutnya. Selain itu juga akan diketahui karakteristik atau keadaan dengan deskripsi dan persentase. Adapun pengelompokkan data yang yang dilakukan didasarkan pada rata-rata ideal (Mi) dan simpangan baku (Sbi).

Skor di atas (Mi + 1,5 SBi) : sangat baik
Mi sampai (Mi + 1,5 SBi) : baik
(Mi – 1,5 SBi) sampai Mi : sedang
(Mi – 1,5 SBi) ke bawah : rendah
dimana:
Mi = ½ (skor tertinggi – skor terendah)
SBi = 1/6 (skor tertinggi – skor terendah)

Penentuan jarak 1,5 SB untuk kategori ini dimaksudkan agar jarak kategori tidak terlalu kecil yang menjadikan kategori lebih banyak serta tidak terlalu lebar yang menjadikan kategori menjadi terlalu sedikit. Hal ini didasarkan pada distribusi normal yang terbagi enam bagian atau enam deviasi standar (Azwar, 2006, p.106).

Analisis butir soal secara empiris menggunakan pendekatan teori respons butir. Analisis empiris diawali uji asumsi unidimensi dan independensi lokal. Secara teoretis pengujian unidimensi menggunakan *expert judgement* untuk melihat apakah butir soal hanya mengukur satu dimensi yakni kemampuan matematika, yakni dengan melihat kesesuaian butir soal dengan kisi-kisi soal. Secara empiris uji unidimensi dilakukan menggunakan analisis faktor dengan bantuan program SPSS 16. Pengujian independensi lokal dilakukan dengan *expert judgement* dengan melihat apakah ada ketergantungan antara butir-butir soal pada perangkat soal.

Analisis butir soal secara empiris menggunakan bantuan program *BilogMG*. Analisis dengan program Bilog MG 3.0 terdiri dari 3 (tiga) fase. Fase pertama merupakan estimasi butir berdasarkan teori tes klasik. Fase kedua menujukkan estimasi parameter butir soal berdasarkan teori respons butir. Fase 3 (tiga) menujukkan estimasi kemampuan peserta tes dan reliabilitas tes. Karakteristik butir soal berdasarkan teori respons butir dapat dilihat pada Tabel 2.

Tabel 2. Karakteristik Butir Soal dengan Pendekatan Teori Respons Butir

Model	**Kriteria**		
	Baik	Tidak baik	Belum dapat digambarkan
1 PL	$p > 0,05$ $-2 \le b \le 2$	$p > 0,05$ $b < -2$ atau $b > 2$	$p < 0,05$
2 PL	$p > 0,05$ $0 \le a \le 2$ $-2 \le b \le 2$	Jika salah satu atau lebih kriteria soal baik tidak terpenuhi	$p < 0,05$
3 PL	$p > 0,05$ $0 \le a \le 2$ $-2 \le b \le 2$ $c \le 0,2$	Jika salah satu atau lebih kriteria soal baik tidak terpenuhi	$p < 0,05$

Perhitungan fungsi informasi butir dilakukan berdasarkan hasil estimsi parameter tiap butir soal. Perhitungan dilakukan dalam skala $-3 \le \theta \le 3$ dengan interval 0,3. Untuk mempermudah perhitungan digunakan bantuan komputer program Microsoft Excel.

Pengujian kesetaraan paket tes adalah pengujian yang dilakukan untuk mengetahui apakah paket-paket soal yang digunakan dalam pengujian memberikan hasil yang setara. Proses

penyetaraan dilakukan untuk menghasilkan soal dengan skala yang sama. Metode penyetaraan yang digunakan dalam penelitian ini adalah metode kurva karakteristik Stocking & Lord nilai estimasi parameter butir tes yang akan disetarakan.

Hasil Penelitian dan Pembahasan

Hasil Pengembangan

Hasil pengembangan berupa butir-butir soal butir-butir soal yang telah dikalibrasi, diketahui karakteristiknya, dan dapat dimasukkan dalam bank soal Bidang Studi Matematika. Bank soal yang dihasilkan dapat digunakan untuk pengujian dan evaluasi pembelajaran. Perangkat tes yang dikembangakan sebanyak 4 (empat) paket soal. Masing-masing paket tes terdiri dari 40 butir soal berbentuk pilihan ganda dengan 4 alternatif jawaban (A, B, C, dan D). Pada setiap paket soal terdapat butir *anchor* sebanyak 20% dari jumlah seluruh butir soal.

Telaah butir soal dilakukan untuk mengetahui kualitas soal dari segi materi konstruksi dan kebahasaan sesuai dengan kaidah penulisan butir soal. Hasil telaah 136 butir soal dengan *expert judgement* dalam penelitian ini dapat dilihat pada tabel 7 tentang ringkasan telaah perangkat tes *Try Out* UN SMP Bidang Studi Matematika.

Tabel 3. Rangkuman Hasil Telaah Perangkat Tes oleh *Expert Judgement*.

Paket soal	Jumlah Butir Diterima	Jumlah Butir Direvisi
P1	26	6
P2	28	4
P3	27	5
P4	26	6
Butir Anchor	6	2

Berdasarkan tabel di atas hasil *expert judgement* menunjukkan bahwa sebagian besar butir-butir soal pada perangkat tes P1, P2, P3, maupun P4 termasuk dalam kriteria baik dinilai dari segi materi, konstruk, bahasa yang digunakan. Dari 136 butir soal, sebanyak 113 butir soal termasuk dalam kriteria baik, 23 butir soal termasuk dalam kriteria cukup, dan tidak terdapat butir soal dengan kriteria tidak baik. Dengan demikian 113 butir soal dapat digunakan untuk uji coba, 23 butir soal perlu direvisi.

Hasil Uji Coba Produk

Uji coba prapenelitian (uji coba skala kecil) meliputi uji keterbacaan perangkat tes oleh guru dan keterbacaan perangkat tes oleh siswa. Uji keterbacaan perangkat tes oleh guru dimaksudkan untuk memperoleh penilaian guru secara umum terhadap perangkat soal berdasarkan aspek materi, konstruksi, bahasa dan kesesuaian perangkat soal terhadap tingkat perkembangan intelektual siswa. Hasil uji keterbacaan oleh guru menunjukkan bahwa keseluruhan perangkat *Try Out* UN SMP Bidang Studi Matematika yakni perangkat P1, P2, P3, dan P4 termasuk dalam kategori sangat baik dengan rerata skor antara 3,50 dampai dengan 3,80, baik dari aspek kesesuaian materi, konstruksi, bahasa, maupun tingkat perkembangan intelektual siswa.

Keterbacaan perangkat tes oleh siswa mencakup kejelasan pada rumusan masalah dan pertanyaan, pilihan jawaban, istilah-istilah, notasi, dan gambar atau grafik yang digunakan pada masing-masing butir soal. Hasil uji keterbacaan oleh siswa pada aspek rumusan masalah dan pertanyaan pada perangkat tes P1 terdapat 33 butir termasuk dalam kategori sangat jelas dan 7 butir termasuk dalam kategori jelas. Pada perangkat tes P2 terdapat 35 butir termasuk dalam kategori sangat jelas 5 butir termasuk dalam kategori jelas. Pada perangkat tes P3 terdapat 34 butir termasuk dalam kategori sangat jelas dan 6 butir termasuk dalam kategori jelas. Pada perangkat tes P4 terdapat 34 butir termasuk dalam kategori sangat jelas dan 6 butir termasuk dalam kategori jelas. Pada aspek pilihan jawaban, penggunaan istilah, penggunaan notsi/ simbol, dan gambar/ grafik, pada P1, P2, P3 dan P4 seluruh butir soal termasuk dalam kategori sangat jelas.

Uji coba produk (uji coba skala besar) dilakukan dengan memberikan perangkat tes. Hasil respons berupa jawaban siswa selanjutnya dianalisis untuk mengetahui karakteristik tes dan estimasi kemampuan peserta tes. Analisis dengan teori respons butir diawali dengan pengujiaan asumsu unidimensi dan independensi lokal.

Uji unidimensi dilakukan melalui *expert judgement*. Berdasarkan hasil *expert judgement* menyatakan bahwa perangkat tes paket 1, 2, 3, dan 4 yang digunakan benar-benar mengukur kemampuan matematika siswa. Hal ini dibuktikan dengan adanya kesesuaian antara

butir-butir dalam perangkat tes dengan kisi-kisi soal. Pegujian unidimensi juga dilakukan dengan analisis faktor menggunakan progran SPSS 16.

Sebelum melakukan analisis faktor dilakukan pengujian kelayakan analisis dengan menggunakan uji KMO-MSA dan uji *Barlet's* pada tiap tes. Syarat analisis faktor adalah nilai Kaiser – Meyer Olkin (KMO) – MSA > 0,05 dan signifikansi uji Barlet < 0,05 (Anderson, Hair & Tatham, 1998, p.88). Uji KMO – MSA digunakan untuk melihat kecukupan sampel, sedang uji barlet digunakan untuk normalitas data yang digunakan. Hasil uji kelayakan analisis faktor dapat dilihat pada tabel 4.

Tabel 4. Hasil uji kelayakan analisis faktor

Paket	KMO-MSA	Sig. Balett's Test
P1	0,755	0,000
P2	0,796	0,000
P3	0,838	0,000
P4	0,787	0,000

Hasil uji kelayakan analisis pada perangkat P1, P2, P3 dan P4 menunjukkan nilai KMO-MSA > dari 0,05 dan uji barlett signifikan, artinya semua tes memenuhi persyaratan analisis faktor. Untuk mendapatkan item-item yang mengukur dimensi yang sama dilakukan proses ekstraksi sehinnga dihasilkan beberapa faktor. Uji asumsi terpenuhi jika tes mengandung satu komponen dominan yang mengukur kemampuan siswa (Hambleton dan Swaminathan, 1985, p.16). Senada dengan hal tersebut jika pengukuran menemukan satu dimensi yang dominan, maka dimensi dominan itu menjadi dimensi tunggal atau unidimensi pada respon atau karakteristik butir (Naga, 1992, 165) menyatakan bahwa.

Pada perangkat tes P1 hasil analisis faktor menunjukkan bahwa faktor 1 merupakan faktor dominan dengan nilai eigen sebesar 5,451. Hasil analisis faktor P2 menunjukkan bahwa faktor 1 merupakan faktor dominan dengan nilai eigen sebesar 6,101. Hasil analisis perangkat P3 menunjukkan bahwa faktor 1 merupakan faktor dominan dengan nilai eigen sebesar 7,520. Hasil analisis perangkat P4 menunjukkan bahwa faktor 1 merupakan faktor dominan dengan nilai eigen sebesar 6,119. Dengan demikian dapat dikatakan bahwa tes P1, P2, P3, dan P4 memenuhi unidimensi.

Uji independensi lokal dalam penelitian ini menggunakan *expert judgement* yang dilakukan oleh 3 orang penelaah ahli. Ketiga penelaah menyatakan bahwa butir-butir perangkat tes pada paket 1, 2, 3, dan 4 butir-butirnya tidak tergantung dengan butir-butir yang lain.

Pada analisis dengan program bilog MG 3.0 P1 fase 1 (PH1) terdapat 3 butir soal yang tidak diikutkan dalam analisis, yaitu butir nomor 7, 10, dan 39. Butir-butir tersebut memiliki nilai koefisien korelasi biserial < 0,3. Pada P2 terdapat 3 butir soal yang tidak diikutkan yakni butir soal nomor 28, 34, dan 36. Pada P3 terdapat 2 butir soal yakni butir soal nomir 28 dan 30. Pada P4 terdapat 4 butir soal yakni butir soal nomir 1, 10, 12, dan 21. Seleksi butir hanya dilakukan pada saat pertama kali program dijalankan, selanjutnya jika ada butir yang dibawah 0,3 tetap digunakan asalkan tidak negatif.

Pada analisis dengan program Bilog MG 3.0 fase 2 (PH2), dapat ditentukan model parameter yang akan digunakan untuk analisis. Model logistik yang digunakan adalah model logistik yang menghasilkan paling banyak butir fit model.Semakin banyak fit model yang dihasilkan maka semakin tinggi persentase perangkat soal dapat digambarkan oleh model. Estimasi parameter butir disini menggunakan model logistik 3 PL.

Hasil estimasi parameter karakteristik butir pada pada perangkat tes P1 dengan model logistik 3PL menujukkan rerata tingkat kesukaran butir (*b*) sebesar 0,234 yang termasuk dalam kategori baik, rerata daya beda (a) sebesar 1,462 termasuk dalam kategori baik dan rerata *guessing* (c) 0,201 termasuk baik. Nilai tingkat kesukaran (b) berkisar antara -1,749 sampai 1,819, seluruh butir soal termasuk dalam kategori tingkat kesukaran baik. Nilai daya beda berkisar antara 0,782 sampai 2,003, terdapat 1 butir soal termasuk dalam kategori daya beda tidak baik, yakni butir soal nomor 30. Nilai *guessing* berkisar antara 0,126 sampai 0,285 terdapat 2 butir soal termasuk dalam kategori *guessing* tidak baik yakni butir 20 dan 30.

Untuk kesesuaian model, terdapat 1 butir yang termasuk butir tidak fit dengan model 3PL yaitu butir 38. Kriteria soal berdasarkan tabel dibedakan menjadi tiga yaitu soal yang baik, soal yang tidak baik dan soal-soal yang belum dapat di gambarkan. Dari 37 butir soal terdapat 34 soal masuk dalam kritaria baik, terdapat 2 butir masuk dalam kriteria tidak baik, serta 1 butir soal yang belum dapat digambarkan oleh model.

Hasil estimasi parameter karakteristik butir soal P2 pada model logistik 3PL menujukkan rerata tingkat kesukaran butir (*b*) sebesar 0,242 yang termasuk dalam kategori baik, rerata daya beda (a) sebesar 1,476 termasuk dalam kategori baik dan rerata *guessing* (c) 0,210 termasuk baik. Nilai tingkat kesukaran (b) berkisar antara -1,657 sampai 1,649 seluruh soal termasuk dalam kategori tingkat kesukaran baik. Nilai daya beda berkisar antara 0,875 sampai 2.103, terdapat 1 butir soal masuk kategori daya beda tidak baik yakni butir 13. Nilai *guessing* berkisar antara 0,157 sampai 0,295 terdapat 3 soal termasuk dalam kategori tebakan semu tidak baik yakni butir 13, 30, dan 39.

Untuk kesesuaian model, seluruh butir soal fit dengan model 3PL. Kriteria soal berdasarkan tabel dibedakan menjadi dua yaitu soal yang baik dan soal yang tidak baik. Dari 36 butir soal yang dianalisis, terdapat 33 soal masuk dalam kritaria dan 40, dan terdapat 3 butir masuk dalam kriteria tidak baik.

Hasil estimasi parameter karakteristik butir soal P3 pada model logistik 3PL menujukkan rerata tingkat kesukaran butir (*b*) sebesar 0,305 yang termasuk dalam kategori baik, rerata daya beda (a) sebesar 1,543 termasuk dalam kategori baik dan rerata *guessing* (c) 0,207 termasuk baik. Nilai tingkat kesukaran berkisar antara -1,643 sampai 1,639 seluruh butir soal termasuk dalam kategori tingkat kesukaran baik. Nilai daya beda berkisar antara 0,856 sampai 1,995 seluruh butir soal termasuk dalam kategori daya beda baik. Nilai *guessing* berkisar antara 0,149 sampai 0,273 terdapat 4 soal termasuk dalam kategori *guessing* tidak baik yakni butir 17, 29, 34, dan 40.

Untuk kesesuaian model, terdapat 2 butir yang termasuk butir tidak fit dengan model (3PL) yaitu butir 4 dan 12. Kriteria soal berdasarkan tabel dibedakan menjadi tiga yaitu soal yang baik, soal yang tidak baik dan soal-soal yang belum dapat di gambarkan. Dari 38 butir soal terdapat 32 soal masuk dalam kritaria baik dan 38, terdapat 4 butir masuk dalam kriteria tidak baik, serta 2 butir soal yang belum dapat digambarkan oleh model.

Hasil estimasi parameter karakteristik butir soal P4 pada model logistik 3PL menujukkan rerata tingkat kesukaran butir (*b*) sebesar 0,302 termasuk dalam kategori baik, rerata daya beda (a) sebesar 1,551 termasuk dalam kategori baik dan rerata *guessing* (c) 0,206 termasuk baik. Nilai tingkat kesukaran berkisar antara -1,536 sampai 1,591, seluruh butir soal termasuk dalam kategori tingkat kesukaran baik. Nilai daya beda berkisar antara 0,763 sampai 2,979, terdapat 3 butir soal termasuk dalam kategori butir daya beda baik, yakni butir 16, 20, dan 23. Nilai *guessing* berkisar antara 0,144 sampai 0,259 terdapat 2 soal termasuk dalam kategori *guessing* tidak baik yakni butir 3 dan 4.

Untuk kesesuaian model, seluruh butir soal pada P4 fit dengan model 3PL. Kriteria soal berdasarkan tabel dibedakan menjadi 2 yaitu soal yang baik dan soal yang tidak baik. Dari 36 butir soal terdapat 32 soal masuk dalam kritaria baik dan terdapat 4 butir masuk dalam kriteria tidak baik.

Fungsi informasi menyatakan kekuatan atau sumbangan butir soal dalam mengungkapkan *latent trait* yang diukur oleh perangkat soal. Dengan fungsi informasi dapat diketahui butir mana yang cocok dengan model sehingga membantu dalam seleksi butir soal.

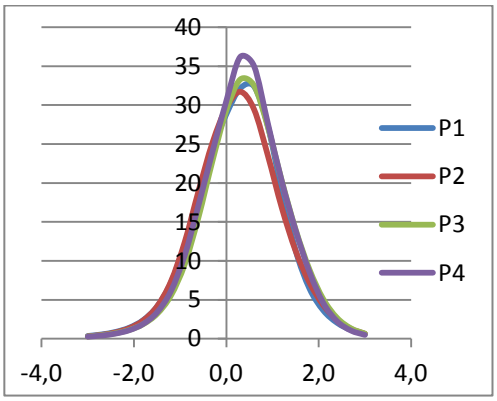
Fungsi informasi maksimum dengan model 3PL berada pada $\theta = 0{,}6$ untuk soal P1 yakni dengan nilai fungsi informasi sebesar 32,404 dan SE 0,176. Dengan demikian dapat dikatakan bahwa perangkat tes P1 dapat memberikan informasi maksilal jika diujikan pada peserta tes berkemampuan 0,6. Nilai informasi yang diperoleh sebesar 32,404 dengan kesalahan pengukuran sebesar 0,176.

Demikian pula untuk perangkat tes P2, P3, dan P4. P2 memberikan informasi maksimal sebesar 31,375 dengan kesalahan pengukuran sebesar 0,178 ketika diujikan pasa peserta tes dengan kemampuan 0,3. P3 memberikan informasi maksimal sebesar 33,292 dengan kesalahan pengukuran sebesar 0,173 ketika diujikan pada peserta tes dengan kemampuan 0,3. P4 memberikan informasi maksimal sebesar 36,186 dengan kesalahan pengukuran sebesar 0,166 jika diujikan pada peserta tes dengan kemampuan 0,3.

Pada semua harga θ $(-3{,}0 \leq \theta \leq 3{,}0)$, estimasi fungsi informasi pada P1 menghasilkan nilai informasi tes sebesar 236,668 dan SE sebesar 0.065. Artinya paket tes P1 mempunyai akurasi untuk mengukur kemampuan peserta didik dalam mata pelajaran matematika sebesar 236,668 dengan tingkat kesalahan pengukuran sebesar 0.065. Begitu pula pada soal P2, P3, dan P4.

P2 mempunyai akurasi untuk mengukur kemampuan peserta didik dalam mata

pelajaran matematika sebesar 231,631 dengan tingkat kesalahan pengukuran sebesar 0,066. P3 mempunyai akurasi untuk mengukur kemampuan peserta didik dalam mata pelajaran matematika sebesar 236,435 dengan tingkat kesalahan pengukuran sebesar 0,065. P4 mempunyai akurasi untuk mengukur kemampuan peserta didik dalam mata pelajaran matematika sebesar 246,190 dengan tingkat kesalahan pengukuran sebesar 0,064. Perbandingan kurva grafik fungsi untuk fungsi informasi soal P1, P2, P3, dan P4 dapat dilihat pada gambar 1.



Gambar 1. Grafik Fungsi Informasi Soal P1, P2, P3, dan P4

Gambar 1 menunjukkan bahwa secara keseluruhan P4 menghasilkan fungsi informasi tertinggi dibandingkan yang lain, artinya P4 memiliki kemampuan untuk mengukur kemampuan peserta didik dalam mata pelajaran matematika secara lebih akurat dibandingkan paket tes lainnya.

Penyetaraan dengan metode kurva karakteristik melibatkan peremeter tingkat kesukaran dan daya beda. Proses penyetaraan dilakukan dengan menggunakan parameter pada *anchor item* yang jumlahnya 8 butir. Penyetaraan paket tes diawali dengan penentuan konstanta penyetaraan, selanjutnya diperoleh formula konversi penyetaraan. Berikut merupakan formula konversi pada perangkat soal *try out* UN bidang studi Matematika tingkat SMP dengan teori respons butir menggunakan metode kurva karakteristik.

a. Penyetaraan P1 terhadap P3 adalah $Y = \alpha X + \beta = 1{,}531X - 0{,}418$
b. Penyetaraan P2 terhadap P3 adalah $Y = \alpha X + \beta = 1{,}285X - 0{,}346$
c. Penyetaraan P4 terhadap P3 adalah $Y = \alpha X + \beta = 1{,}539X - 0{,}340$

Setelah diperoleh formula konversi, selanjutnya dapat dilakukan penyetaraan perangkat soal berdasarkan butir atau konversi parameter butir. Konversi berdasarkan butir dilakukan pada parameter daya beda dan tingkat kesukaran.

Simpulan dan Saran

Simpulan

Jumlah perangkat soal yang dikembangkan adalah 4 paket soal yang masing-masing terdiri atas 40 butir soal dengan 8 butir *anchor* di dalamnya. Pada soal P1 terdapat 26 butir, P2 terdapat 25 butir, P3 terdapat 24 butir, P4 terdapat 24 butir soal, dan 8 butir soal *anchor* memiliki karakteristik baik.

Rerata tingkat kesukaran soal P1 adalah 0,234, P2 adalah 0,242, P3 adalah 0,305, dan P4 adalah 0,302; rerata daya beda soal P1 adalah 1,462, P2 adalah 1,476, P3 adalah 1,543, dan P4 adalah 1,551; dan rerata *pseudo guessing* soal P1 adalah 0,201, P2 adalah 0,210, P3 adalah 0,207, dan P4 adalah 0,206.

Nilai fungsi informasi tertinggi P1 berada pada $\theta = 0{,}6$ yaitu 32,404 dengan SE sebesar 0,176, P2 berada pada $\theta = 0{,}3$ yaitu 31,725 dengan SE sebesar 0,178, P3 berada pada $\theta = 0{,}3$ yaitu 33,291 dengan SE sebesar 0,173, dan P4 berada pada $\theta = 0{,}3$ yaitu 36,186 dengan SE sebesar 0,166. Pada semua harga θ ($-3{,}0 \leq \theta \leq 3{,}0$), P4 mempunyai akurasi untuk mengukur kemampuan peserta didik dalam mata pelajaran matematika paling tinggi dibandingkan paket soal yang lain.

Persamaan penyetaraan pada perangkat soal *try out* UN bidang studi Matematika tingkat SMP dengan teori respons butir menggunakan metode kurva karakteristik adalah: a) penyetaraan P1 terhadap P3 adalah $Y = 1{,}531X - 0{,}418$; b) penyetaraan P2 terhadap P3 adalah $Y = 1{,}285X - 0{,}346$; c) penyetaraan P4 terhadap P3 adalah $Y = 1{,}539X - 0{,}340$.Berdasarkan persamaan di atas dapat disimpulkan bahwa paket tes P3 memiliki tingkat kesulitan yang paling tinggi, disusul P4, kemudian P1 dan terakhir P2.

Saran

Butir-butir soal yang dikembangkan dalam penelitian ini dapat dimanfaatkan sebagai alat ukur hasil pembelajaran matematika tingkat SMP, khususnya dalam mempersiapkan diri

menghadapi UN Matematika SMP. Selain itu butir-butir soal juga dapat digunakan dalam pengujian yang relevan sesuai dengan kebutuhan. Pengembangan produk lebih lanjut diharapkan dapat dilakukan pada mata pelajaran lainnya dan pada jenjang pendidikan dasar menengah maupun perguruan tinggi.

Daftar Pustaka

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

Anderson, R.E., & Hair, J.F.,& Tatham, R.L. et al. (1998). *Multivariate data analysis*. USA. Pretince – Hall. International Inc.

Azwar, Saifuddin. (2006). *Penyusunan skala psikologi*. Yogyakarta: Pustaka pelajar.

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

Depdiknas. (2006). S*tandar Isi Mata Pelajaran Matematika SD/MI dan SMP/MTs (Permendiknas Nomor 22 Tahun 2006).*Jakarta: BSNP, Depdiknas.

Dorans, N.J., Moses, T.P., & Eignor, D.R. (2010). Principles and practices of test score equating. *Research Report*. New Jersey: Educational Testing Service (ETS).

Dudley, U. (1997). Is mathematics nacessary?. *The College Mathematics Journal*, 28,360 – 364.

Gajjar, S. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, 39, 17 – 20.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Application*. Boston, MA: Kluwer Inc.

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

Keller, L.A. & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling method in mixed-format test. *Journal of Educational Measurement*, 50, 390 – 407.

Kolen M. J. & Bremann, R. l. (1995). *Test equiting: methods and practices.* New York: Springer.

Lawrence, R. (1998). *Item banking. practical assessment, research and evaluation, 6(4).* Diakses pada 3 Desember 2013 dari http://pareonline.net.getvn.asp?v=6&n=4.

Mardapi, Djemari. (2012). *Pengukuran penilaian & evaluasi pendidikan*. Yogyakarta: Nuha Medika.

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

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan.* Jakarta: Besbats.

Reynolds, Livingston & willson. (2010). *Measurement and assessment in education*. New Jersey: Pearson Education Inc.

Suyata, P., Mardapi, D., & Kartowagiran, B. (2010). Identifikasi need assessment: studi awal model pengembangan bank soal berbasis guru di Provinsi DIY. *Jurnal Kependidikan*. 1.45-58.

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking*. New York: Springer.

Wood. R. & Skurnik. L. S. (1969). *Item banking*. England: national Fundation for Educational Research.