PENGEMBANGAN INSTRUMEN PENGUKUR HASIL BELAJAR NIRBIAS DAN TERSKALA BAKU

Djemari Mardapi, Kumaidi, Badrun Kartowagiran
Program Pascasarjana UNY
Jl. Colombo No.1 Yogyakarta 55281
djemarimardapi@yahoo.com, badrunkw@yahoo.com

Abstrak

Penelitian hibah ini bertujuan untuk mengembangkan instrumen pengukur hasil belajar yang nirbias dan terskala baku yang digunakan dalam beberapa mata pelajaran di SMA dan atau SMP. Jenis penelitian ini adalah *research and development* (R&D) yang dilakukan selama dua tahun. Hasil penelitian tahun pertama dihasilkan draf instrumen pengukur hasil belajar yang nirbias dan terskala baku. Tahun kedua, diseminasi draf instrumen pengukur hasil belajar yang dihasilkan tahun pertama ke beberapa guru Matematika SMA dan SMP di Provinsi DIY dan Jawa Tengah. Setelah direvisi, instrumen disosialisasikan ke beberapa guru Matematika SMA dan SMP di Provinsi DIY, Jawa Tengah dan NTB.

Kata kunci: *instrumen yang nirbias dan terskala baku*

DEVELOPING UNBIASED AND STANDARDIZED INSTRUMENTS FOR STUDENT ACHIEVEMENTS IN HIGH SCHOOLS

Djemari Mardapi, Kumaidi, Badrun Kartowagiran
Program Pascasarjana UNY
Jl. Colombo No.1 Yogyakarta 55281
djemarimardapi@yahoo.com, badrunkw@yahoo.com

Abstract

The purpose of this research is to develop unbiased and standardized instruments to measure student achievements that can be used for several subject matters in senior high schools and junior high schools. This was a research development study carried out in two years. The result of this research in the first year was a draft of an unbiased and standardized instuments for student achievements. In the second year, the draft of the unbiased and standardized instruments was disseminated to several mathematics teachers in senior high schools and junior high schools in the provinces of Yogyakarta Special Territory and Central Java. After the revision, the instruments were disseminated again to mathematics teachers in senior high schools and junior high schools in the provinces of Yogyakarta Special Territory, Central Java, and Nusa Tenggara Barat.

Keywords*: unbiased instruments, standardized instruments*

Pendahuluan

Sampai saat ini banyak instrumen hasil belajar, baik yang digunakan oleh guru untuk ulangan harian maupun yang digunakan oleh sekolah untuk ulangan umum belum memenuhi persyaratan sebagai tes yang baik, yakni nirbias dan terskala baku. Instrumen hasil belajar yang mengandung bias butir akan merugikan siswa-siswa yang memiliki kemampuan sama tetapi peluang menjawab benar suatu instrument tes tidak sama, karena berbeda kelamin atau kelompok. Dalam hal ini, perbedaan kelompok itu dapat diartikan perbedaan kultur, gender, agama, dan lainnya. Sementara itu, instrumen yang tidak terskala baku tidak mampu menghasilkan skor yang dapat dibandingkan antarwilayah, antarkelompok, dan antartahun.

Dalam pengembangan tes, ada beberapa tahapan yang harus dilalui, yakni: (1) perancangan tes, (2) uji coba tes, (3) penetapan validitas, (4) penetapan reliabilitas, dan (5) interpretasi skor tes. Kegiatan perancangan tes tercakup di dalamnya yakni: (1) penetapan tujuan, (2) penyiapan tabel spesifikasi, (3) menyeleksi format item yang sesuai, (4) menulis item, dan (5) mengedit item. Kegiatan uji coba tes meliputi kegiatan: (1) analisis item pengujian uji coba pertama, (2) analisis item pengujian uji coba kedua, dan (3) penyiapan format tes. Prosedur ini harus diikuti untuk menghasilkan instrumen tes yang baik.

Tantangan yang dihadapi dalam setiap pengukuran berkaitan dengan panjang tes dan banyaknya kriteria yang digunakan untuk menskala respons yang diberikan oleh siswa. Selain itu, dalam pembakuan item ukuran sampel juga ikut menentukan tingkat kestabilan yang dicapai. Oleh karena itu, perlu dibahas beberapa teori yang terkait dengan penelitian ini.

1. Teori respons butir

Menurut Han & Hambleton (2007: 15-20) juga Theissen et al. (2001: 295-325), bahwa pada model-model respons butir dikotomus, jenis data responsnya yang digunakan adalah biner (yaitu, 0 atau 1). Namun demikian, dalam beberapa situasi tes, respons bisa lebih dari dua kategori. Sebagai contoh, suatu kuesioner yang menanyakan sikap *(attitude)*, dengan menggunakan butir skala Likert, akan menghasilkan respons 5 kategori, yaitu sangat tidak setuju, tidak setuju, setuju, dan sangat setuju.

Di bagian lain Han & Hambleton (2007) menjelaskan bahwa ada dua model analisis respons butir, yakni respon butir dikotomus dan politomus. Analisis respon butir dikotomus mencakup model ogif normal, model logistik satu-parameter, model logistik dua-parameter, model logistik tiga-parameter, sedangkan model politomus mencakup: model Kredit Parsial (PCM), model Kredit Parsial Umum (GPCM), model Skala Rating (RSM), dan model Respons Bertahap (GRM).

Penelitian ini menggunakan *Partial Credit Model* (PCM) sekaligus *Rasch Model* (RM) untuk pengujian *fit* item tes keterampilan proses sains pola divergen untuk mata pelajaran Biologi SMA. Dasar pertimbangan yang digunakan adalah yang pertama bahwa PCM sebagai perluasan RM yang merupakan model 1-PL dapat menggunakan sampel yang tidak sebanyak kalau melakukan kalibrasi data politomus menggunakan model 2-PL atau 3-PL (Keeves & Masters, 1999: 12-13). Kedua, bahwa karakteristik respons terhadap item keterampilan proses sains mengikuti pola PCM. Karakteristik PCM adalah bahwa tingkat kesukaran (delta) dari suatu tahapan kategori di bawahnya ke kategori di atasnya tidak sama antaritem satu dengan yang lain, sehingga besarnya *delta* untuk suatu tahapan kategori di bawahnya dan *delta* untuk tahapan kategori di atasnya tidak sama antaritem satu dengan item lainnya.

Paket program yang digunakan untuk melakukan analisis butir dalam penelitian ini adalah QUEST. Elemen sentral program QUEST adalah *Rasch Model* (RM). Program ini dapat menggunakan data respons yang diskor secara politomus. Program QUEST dalam melakukan estimasi parameter, baik untuk item maupun untuk testi (*case/person/*) menggunakan *unconditional* (UCON) atau *joint maximum likelihood* (Adam & Khoo, 1996: 89).

Penetapan *fit* item secara keseluruhan dengan model program QUEST (Adam & Kho, 1996) didasarkan pada nilai rata-rata INFIT *Mean of Square* (INFIT MNSQ) beserta simpangan bakunya atau nilai rata-rata INFIT *Mean of* INFIT *t*. Penetapan *fit* tiap item pada program QUEST didasarkan pada besarnya nilai INFIT MNSQ atau nilai INFIT *t* item yang bersangkutan. Langkah untuk memperolehnya mengikuti prosedur yang ditulis Wright & Masters (1982, 93-104).

Besarnya kuadrat tengah yang tidak tertimbang (*Unwighted Mean Square* atau u_i)-dalam program QUEST disingkat OUTFIT MNSQ) maupun kuadrat tengah tertimbang (*Wighted Mean Square*) yang diharapkan adalah sebesar 1 dan varians sebesar 0. Besarnya nilai harapan Mean INFIT *t* sama dengan 0 dengan varians sama dengan 1.

Penetapan *fit* testi (*case/person*) secara keseluruhan dengan program QUEST (Adam & Kho, 1996) didasarkan pada besarnya nilai rata-rata INFIT *Mean of Square* (INFIT MNSQ) beserta simpangan bakunya. Penetapan o didasarkan pada besarnya nilai rata-rata INFIT *Mean of* INFIT *t*. Penetapan *fit* tiap testi (*case/person*) dengan model dalam program QUEST didasarkan pada besarnya nilai INFIT MNSQ atau nilai INFIT *t* item yang bersangkutan. Langkah perhitungannya mengikuti langkah yang ditulis Wright & Masters (1982: 108-109).

Program QUEST juga menyajikan hasil realiabilitas tes menurut CTT, yakni berupa indeks konsistensi internal, yang untuk peskoran politomus merupakan indeks alpha Cronbach dan untuk penskoran dikotomus merupakan indeks KR-20 (Adam & Khoo, 1996: 93). Ini berarti bahwa dengan program QUEST dapat diperoleh item atau butir dan testi yang fit, disertai dengan reliabilitas instrumen tes tersebut.

2. Penyetaraan

Jika dalam pengembangan tes disusun beberapa subtes yang diujikan pada kelompok peserta uji yang berbeda, maka diperlukan adanya proses penyetaraan terhadap keseluruhan subtes agar hasil-hasil subtes tadi dapat diskalakan pada satu skala. Menurut Hambleton et al. (1991: 123-143) penyetaraan skor tes atau *equating* (lebih tepat lagi *scaling* atau *linking*. Mengacu pada pendapat Lord bahwa bahwa tindakan mengkonversi skor tes yang satu (skor tes X) menjadi skor metrik sesuai dengan ukuran dari tes yang lain (skor tes Y) agar peserta yang memperoleh skor xc pada tes X akan memperoleh skor baru setelah dikonversi ke dalam tes Y (katakanlah y_c*). Selanjutnya skor baru hasil konversi inilah maka skor peserta tersebut dapat diperbandingkan dengan skor y yang diperoleh peserta lain yang menempuh tes Y. Menurut menurut Kolen dan Brannen (1995: 2) *equiting* skor tes adalah suatu proses statistika yang digunakan untuk melakukan

penyesuaian skor dalam suatu tes sehingga tes yang disesuaikan tersebut dapat dipergunakan bersifat "*interchangeable*". melalui *equiting* skor tes dapat diambil keputusan yang adil, baik dalam mengambil keputusan kelulusan, seleksi, atau sertifikasi yang didasarkan pada paket/format tes yang berbeda.

Sebagian para ahli pengukuran memilih IRT daripada teori klasik untuk melakukan *equating* skor tes karena kalau mendasarkan teori klasik banyak memiliki kelemahan. Metode penyetaraan menurut teori klasik dikelompokkan menjadi dua kategori utama, yakni: (1) penyetaraan equipersentil dan (b) penyetaraan linear.

Menurut teori respons butir/item, parameter kemampuan θ dari seorang peserta adalah invariant lintas subset butir/item. Ini berarti bahwa terlepas dari kesalahan pengukuran, kemampuan mengestimasi juga akan invariant lintas subset butir. Karenanya, dua peserta yang merespons berbeda terhadap subset butir (atau dari tes yang berbeda) di mana nilai-nilai parameter butir diketahui akan memiliki kemampuan mengestimasi pada skala yang sama maka tidak diperlukan penskalaan atau penyetaraan. Model IRT mencakup beberapa metode yakni: (a) metode regresi (*Regression Method*), (b) metode Nilai rata-rata dan Sigma (*Mean and Sigma Method*), (c) metode Nilai Rata-Rata dan Sigma yang "Robust" (*Robust Mean and Sigma Method*), dan (d) Metode kurve Karakteristik (*Characteristic Curve Method*), (e) metode simultan dengan menggunakan "anchor items".

3. Pendeteksian Bias Item

Dalam Permendiknas Nomor 20 Tahun 2007 tentang Standar Penilaian Pendidikan dinyatakan bahwa salah satu prinsip penilaian adalah adanya unsur keadilan. Adil dalam arti bahwa penilaian tidak menguntungkan atau merugikan peserta didik karena berkebutuhan khusus serta perbedaan latar belakang agama, suku, budaya, adat istiadat, status sosial ekonomi, dan gender.

Untuk memberikan penilaian yang adil, instrumen penilaian harus bebas dari adanya unsur bias item/butir tes yang disebabkan adanya *differential item functioning* (DIF). Deteksi bias butir dapat diselidiki dengan menggunakan beberapa metode seperti metode Mantel-Haenzel (Rogers &

Hambleton, 1993:105), sibtest (Gierl, Khaliq, & Boughton (1999: 11), regresi logistik (Embretson & Reise, 2000, 251). Pendeteksian bias butir juga dapat dikaitkan dengan teknik penskoran juga dapat dihitung atas dasar besarnya indek kecurangan terutama pada soal bentuk pilihan ganda.

Idealnya tidak ada kesalahan dalam pengukuran, baik kesalahan yang acak maupun kesalahan yang sistematis. Dengan kata lain, seharusnya tidak ada kesalahan yang dilakukan oleh peserta tes, pelaksanaan tes, dan juga tidak ada kesalahan pengukuran yang disebabkan oleh butir tes. Instrumen yang digunakan untuk mengukur seharusnya memiliki validitas dan reliabilitas mantap, dan adil. Artinya, tidak ada orang atau kelompok orang tertentu yang merasa dirugikan dengan adanya butir soal yang tidak adil itu. Kenyatannya tidak selalu demikian, soal ujian nasional masih saja ada yang mengandung bias butir. Hasil penelitian Budiyono (2005) menunjukkan bahwa hasil UN Matematika tahun 2004 Jurusan IPA di Surakarta memiliki empat butir yang mengandung unsur bias butir. Sementara itu, Kartowagiran (2005) menemukan ada sembilan butir yang bias pada soal Matematika SMP yang digunakan dalam Ujian Nasional tahun 2003 bila dideteksi menggunakan *Likelihood Ratio Test*. Selain itu, hasil ujian nasional juga tidak selalu dapat dibandingkan antarwilayah dan antartahun karena tidak selalu menggunakan pendekatan teori respon butir. Pendekatan teori respon butir yang hasilnya dapat dibandingkan antarwilayah, antartahun ini tidak selalu dapat digunakan di Indonesia karena masyarakatnya belum semuanya memahami pendekatan ini. Salah satu cara yang dapat digunakan adalah, soal itu dikembangkan dari suatu *learning continum.*

Terkait dengan hal di atas, perlu dikembangkan instrumen pengukur hasil belajar beserta teknik penskalaan berdasarkan teori respons butir untuk suatu mata pelajaran di SMP/SMA yang menjadi pijakan, baik bagi penyelenggaraan *assessment for learning* ataupun *assessment of learning.*

Metode Penelitian

Penelitian ini termasuk jenis *research and development* (R&D) yang dilakukan selama dua tahun. Tahun pertama, yang dilakukan adalah: (1) mencermati standar kompetensi (SK) dan kompetensi dasar (KD) dalam

Kurikulum Tingkat Satuan Pendidikan (KTSP), kemudian merumuskan *learning continuum* melalui *forum group discussion* (FGD), (2) menulis butir-butir instrumen pengukur hasil belajar, (3) menelaah dsn merevisi butir butir instrumen, dan (3) melakukan uji coba dan menganalisis butir instrumen untu mendeteksi bias butir dan mengkaji keselarasan SK dan KD atau rumusan *learning continum* dengan butir-butir dalam instrumen. Tahun kedua, kegiatan yang dilakukan adalah: (1) mengeset instrumen pengukur hasil belajar untuk mata pelajaran sain SMP dan Matematik SMA yang telah dihasilkan pada tahun pertama, (2) melakukan diseminasi dan revisi, dan (3) melakukan sosialisasi.

Penelitian tahun pertama berupa kegiatan persiapan, tahap pengkajian dan penyusunan *learning continuum* instrumen pengukur hasil belajar di SMP/SMA serta tahap penulisan instrumen pengukur hasil belajar untuk mata pelajaran yang bersangkutan dilaksanakan di Program Pascasarjana UNY. Adapun pelaksanaan uji coba di lakukan di SMP/SMA pada provinsi sesuai dengan judul anak payung. Demikian pula pelaksanaan pencarian data dilakukan di Dinas Pendidikan terkait sesuai dengan judul anak payung. Analisis data dilaksanakan di Program Pascasarjana UNY. Penelitian tahun kedua berupa kegiatan diseminasi dan kegiatan sosialisasi hasil penelitian kepada pihak terkait dilaksanakan di Program Pascasarjana dengan mengundang baik guru, pengelola satuan pendidikan, maupun Dinas Pendidikan dimana penelitian dilaksanakan.

Pengumpulan data dalam penelitian ini dilakukan dengan cara FGD, tes, angket, dan dokumentasi. *Forum group discussion* (FGD) digunakan sewaktu menyusun *learning continuum,* tes digunakan untuk mencari respon siswa terhadap butir-butir instrumen yang dikembangkan, angket digunakan sewaktu diseminasi dan digunakan untuk mengumpulkan pendapat para guru tentang butir-butir instrumen, dan dokumentasi digunakan untuk mengumpulkan data atau respon siswa yang diperlukan oleh sebagian peneliti anak payung.

Analisis data menggunakan pendekatan kualitatif dan kuantitatif. Analisis data secara kualitatif dalam bentuk analisis deskriptif digunakan untuk menganalisis hasil pengkajian SK dan KD dari silabus KTSP yang

sudah ada, hasil perumusan *learning continuum* mata pelajaran yang bersangkutan di SMP/SMA, juga hasil review dan revisi intsrumen pengukur hasil belajar untuk mata pelajaran yang bersangkutan. Analisis data secara kuantitatif menggunakan pendekatan IRT dikhotomus dan politomus dengan paket program sesuai dengan karakteristik masing-masing penelitian anak payung.

Hasil Penelitian dan Pembahasan

Hasil Penelitian

1. Tahun Pertama

Penelitian diawali dengan perumusan *learning continuum*, penyusunan kisi-kisi, dan penulisan soal untuk keterampilan proses sains oleh peneliti. Selanjutnya, dilakukan penelaahan melalui forum *Focus Group Discussion* (FGD). Peserta FGD untuk tinjauan keilmuan biologi terdiri dari pakar bidang ilmu, pakar pendidikan bidang ilmu, yang dalam hal ini sesuai dengan sampel mata pelajaran yang dikembangkan untuk disertasi mahasiswa peneliti, yaitu Matematika SMA dan Aspek Fisika IPA SMP. Selain itu, ditambah dengan pakar pengukuran, yang dalam hal ini adalah ketua dan anggota anggota peneliti sebagai penanggung jawab sekaligus sebagai promotor disertasi mahasiswa yang berangkutan. Para praktisi diambil dari guru SMA pengampu mata pelajaran Matematika dari SMA dan guru pengampu aspek Fisika IPA SMP di Kotamadya Yogyakarta.

Learning continuum, kisi-kisi dan item tes mata pelajaran Matematika dan Fisika yang sudah memperoleh masukan dari FGD selanjutnya dirakit menjadi perangkat tes yang siap diujicobakan. Perumusan *learning coninuum* selain memperhatikan pendapat para ahli juga memperhatikan cakupan kompetensi yang terumuskan di dalam Standar Isi (SI) mata pelajaran yang bersangkutan dalam Lampiran Permendiknas Nomor 22 Tahun 2006.

Learning continuum dirumuskan untuk pembelajaran Matematika Kelas X SMA. Dasar penyusunan selain mengkaji kompetensi yang terdapat dalam SI yang terdapat di dalam Permendiknas Nomor 22 Tahun 2006 juga memperhatikan pendapat Begle (1979), Trivieri (1989) tentang Matematika Dasar. Untuk mempermudah penelusuran pengetahuan prasyarat

dibuat pemetaan konsep, materi matematika dari SD sampai kelas X SMA. Dasar-dasar Matematika.

Pengembangan *learning continuum* untuk aspek Fisika IPA SMP menggunakan pustaka yang memuat hal-hal yang berkait dengan *High Order Thinking (HOT)*. Dalam hal ini diambil dari pendapat Thomas dan Thorne. Selain itu, diambilkan dari buku *Writing Test Items to Evaluate Higher Order Thinking* yang ditulis oleh Haladyna.

2. Tahun Ke II

Kegiatan utama tahun kedua penelitian ini adalah desiminasi dan sosialisasi instrumen pengukur hasil belajar yang nirbias dan terskala baku yang dihasilkan pada tahun kedua. Kegiatan sosialisasi dilaksanakan di tiga provinsi yakni DIY, Jawa Tengah, dan NTB atau Sulawesi Tenggara. Untuk kegiatan diseminasi dipilih Provinsi DIY dan Jawa Tengah dengan pertimbangan dapat dipilih para peserta yang dapat memberikan masukan untuk penyempurnaan buku panduan. Informasi kemampuan guru diperoleh dengan menghubungi Kepala Sekolah, yakni guru yang sudah mampu mengoperasikan komputer dengan baik, agar kalau ada kegagalan diseminasi bukan akibat kemampuan guru menggunakan komputer tetapi karena panduannya kurang baik. Untuk sosialisasi, atas dasar pertimbangan faktor mahasiswa yang terlibat dipilih dilaksanakan di DIY, Jawa Tengah, dan NTB.

Dalam persiapan diseminasi dan sosialisasi, tim peneliti menyusun panduan cara menulis *learning continuum*, menulis butir soal yang nirbias, dan cara melakukan deteksi bias butir. Untuk dapat menyusun panduan yang cermat maka perlu dilakukan praktik berulang-ulang, terutama cara melakukan deteksi bias butir dengan program QUEST. Untuk kegiatan diseminasi disediakan data riil untuk dianalisis oleh para guru peserta sehingga guru dapat melakukan analisis dan menafsirkan hasilnya sesuai dengan panduan dan secara substansi didiskusikan antarguru.

Dalam penelitian ini, instrumen yang diseminasikan dan sosialisasikan kepada guru Matematika SMP dan SMA adalah: (1) *learning continuum,* (2) cara mengembangkan butir-butir soal yang dijabarkan dari *learning continuum* itu, dan (3) cara melakukan analisis untuk mendeteksi bias butir.

Panduan penulisan item, termasuk di dalamnya cara pengembangan *learning continuum* dan panduan analisis data menggunakan program QUEST.

Diseminsi instrumen di Provinsi DIY dan Jawa Tengah dilaksanakan secara serempak pada tanggal 30-31 Oktober 2010 bertempat di Pascasarjana Universitas Negeri Yogyakarta. Peserta yang diundang untuk mengikuti diseminsi ini 50 orang; 25 guru SMP dan 25 guru SMA. Namun kenyataannya, yang hadir hanya ada 24 guru. Mereka berasal dari Kabupaten Bantul, Kota Yogyakarta, Kabupaten Purworejo, Kabupaten Klaten, dan Kabupaten Magelang.

Pada hari pertama, peserta diberi informasi tentang hasil penelitian ini, termasuk teori dan praktik cara membuat *learning continuum,* menulis butir-butir soal nirbias termasuk contoh-contoh butir nirbias, dan cara melakukan deteksi bias. Hari kedua, mereka berlatih mendeteksi bias dan diminta untuk untuk menggunakan dan memberi masukan pada panduan. Panduan yang dimaksudkan adalah Panduan cara menyusun *learning continuum* dan cara melakukan deteksi bias butir. Menurut pengamatan dan hasil wawancara dengan peserta, diseminasi menguntungkan dua pihak, pihak peserta dan Tim Peneliti. Peserta merasa senang karena mendapatkan materi baru, yaitu mengenai cara menulis *learning continuum,* ciri-ciri butir soal yang bias, dan cara mendeteksi bias butir. Tim Peneliti senang karena peserta memberi masukan pada panduan sangat serius, sehingga ada bahan pertimbangan untuk merevisi panduan.

Panduan cara menulis *learning continuum* dan melakukan deteksi bias yang sudah direvisi berdasarkan masukan pada saat diseminasi digunakan untuk sosialisasi. Kegiatan sosialisasi di Provinsi DIY dan Jawa Tengah dilaksanakan secara serempak pada tanggal 6-7 November 2010 bertempat di Pascasarjana Universitas Negeri Yogyakarta. Kegiatan sosialisasi di NTB dilaksanakan tanggal 13–14 November di Aula Dinas Pendidikan Lombok Utara.

Kegiatan sosialisasi untuk Provinsi DIY dan Jawa Tengah diselenggarakan di Pascasarjana UNY dengan mengundang 40 guru, namun yang hadir hanya 21 orang. Mereka berasal dari Kabupaten Bantul, Kota Yogyakarta, Kabupaten Purworejo, Kabupaten Klaten, dan Kabupaten Magelang.

Kegiatan sosialisasi di NTB direncanakan mengundang 20 peserta. Pelaksanaan memilih di Kabupaten Lombok Utara dengan pertimbangan kemudahan transportasi peserta. Peserta terdiri atas 8 guru matematika SMA, 8 guru matematika SMP, seorang guru matematika MA, seorang guru matematika MTs, seorang peserta dari Dinas Pendidikan KLU, seorang pengawas matematika SMP Dinas pendidikan KLU, dan seorang pengawas matematika SMA Dinas Pendidikan KLU.

Kegiatan sosialisasi pada hari pertama, peserta diberi informasi tentang hasil penelitian ini, termasuk teori dan praktik cara membuat *learning continuum,* menulis butir-butir soal nirbias termasuk contoh-contoh butir nir bias, dan cara melakukan deteksi bias. Pada akhir hari pertama kepada mereka diberi tugas untuk membuat 10 butir pilihan ganda empat pilihan yang nirbias. Pada hari kedua, butir-butir soal yang mereka buat ditelaah, dilihat dari nirbiasnya, sesudah itu mereka berlatih mendeteksi bias butir dengan menggunakan program QUEST. Hasil wawancara dengan peserta menunjukkan bahwa mereka merasa senang karena mendapatkan materi baru, misalnya cara menulis *learning continuum*, ciri-ciri butir yang bias, dan cara mendeteksi bias butir.

Dilihat dari kualitas butir-butir soal yang dibuat peserta untuk daerah DIY dan Jawa Tengah sudah cukup baik dan lengkap; waktu ditelaah, secara teoretik tidak ada butir yang bias. Akan tetapi, untuk peserta dari Lombok Utara tampaknya masih perlu adanya pembinaan lebih lanjut kepada para guru di lapangan. Hal ini disebabkan masih banyak peserta yang hanya mengambil kisi-kisi dan soal dari buku-buku yang beredar di lapangan. Guru belum sepenuhnya berani menyusun instrumen sendiri. Namun demikian, diskusi peserta sudah memahami perihal cara penyusunan instrumen yang baik, termasuk yang nirbias.

Pada tahun kedua, jumlah mahasiswa yang terlibat dalam penelitian hibah pasca ini ada enam orang mahasiswa, dua mahasiswa S3 dan empat mahasiswa S2. Dari dua mahasiswa S3, satu orang sudah lulus doktor dan satu orang lainnya hampir ujian tertutup. Sementara itu, dari empat mahasiswa S2 yang terlibat, dua mahasiswa sudah lulus dan dua mahasiswa lainnya belum ada kemajuan yang berarti karena berbagai alasan.

Pembahasan

Hasil penelitian payung menunjukkan bahwa dari pengembangan instrumen untuk mengukur kemampuan Matematika SMA Kelas X yang *learning continuum* dan kisi-kisi serta itemnya dikembangkan oleh mahasiswa peneliti A. Fauzan dan disempurnakan melalui proses pembimbingan oleh promotor, dan FGD, menunjukkan bahwa hasil uji coba terbukti tes secara keseluruhan *fit* dengan *Rasch Model*. Namun, untuk item ada tiga dari 42 item yang tidak *fit* dengan *Rasch Model*. Hasil tersebut menunjukkan bahwa peranan *expert judgement* dalam proses analisis secara kualitatif berfungsi dengan baik. Ini sesuai dengan pendapat Messick (dalam Wainer & Braun, 1988:34-35). bahwa penilaian oleh para pakar dapat dijadikan alternatif bukti validitas. Meskipun demikian, memang ada kemungkinan validitas oleh para pakar sangat subjektif, tergantung kepada latar belakang pengetahuannya (Mardapi, 2008: 16-19).

Setelah tiga item dikeluarkan, ternyata masih ada dua item yang berubah menjadi tidak *fit* dengan model. Akan tetapi, kedua item yang tidak *fit* masih memiliki nilai *mean ability* yang positif meskipun relatif rendah, yakni 0,15 untuk item nomor 25 dan 0,11 untuk item nomor 28. Mengacu kepada pendapat Frisbie (2005: 26), untuk tes beracuan kriteria, sepanjang tidak memberikan nilai negatif daya beda, yang berarti juga untuk *point biserial* juga untuk *mean ability*, berarti masih dapat dijadikan item pengukur penguasaan kemampuan hasil belajar untuk pembelajaran berbasis kompetensi.

Berdasarkan segi deteksi bias item, yang dalam hal ini dilacak berdasarkan jenis kelamin, ternyata untuk tes Matematika SMA kelas X yang disusun mahasiswa peneliti menunjukkan ada 3 item yang lebih mudah untuk kelompok laki-laki dan ada satu item yang lebih mudah untuk kelompok perempuan. Namun, jika dilihat dari karakteristik item yang diujikan tanpa adanya kasus yang terkait karakteristik gender, tampaknya lebih berkait dengan faktor lain yang masih perlu diselidiki lebih lanjut.

Secara umum hibah ini menunjukkan hasil yang baik dari segi kemampuannya mendorong dan memfasilitasi mahasiswa untuk menyelesaikan studinya. Dengan hibah ini dosen menjadi terus mengontrol kinerja

mahasiswa. Oleh karena itu, hanya karena faktor yang bersifat internal dari diri mahasiswa saja yang menjadi penghambat. Selain itu, penelitian ini juga mendorong guru dalam mengoperasikan komputer terutama untuk analisis butir. Penggunaaan paket program QUEST memerlukan keterampilan pemahaman komputer meskipun tidak perlu tinggi. Faktor lain yang perlu diperhatikan adalah di lapangan masih ada guru yang tidak memiliki pengetahuan yang memadai dalam mengembangkan instrumen. Hal ini menjadi faktor yang memiliki peran kunci untuk meningkatkan pengetahuan guru di lapangan. Dengan demikian, perlu adanya *inhouse training* untuk meningkatkan kemampuan guru terlebih dalam hal mendeteksi bias butir agar dapat menghasilkan butir/item tes hasil belajar yang nirbias.

Simpulan dan Saran

Simpulan

Berdasarkan hasil penelitian dan pembahasan dapat ditarik beberapa kesimpulan sebagai berikut.

1. Pengembangan instrumen dengan mengikuti langkah-langkah yang sistematis dapat memberikan hasil yang optimal, yakni ditunjukkan dengan sedikitnya item yang tidak *fit* dengan *Rasch Model* untuk data dikotomus.
2. *Learning continuum* yang dirumuskan dapat dijadikan *abstract continum* dalam pengukuran hasil belajar.
3. Kegiatan diseminasi dan sosialisasi berjalan dengan baik. Kegiatan tersebut menambah pengetahuan guru tentang *learning continuum,* cara menyusun butir nir bias dan cara mendeteksi bias butir dengan *QUEST* menambah pengetahuan dan keterampilan peserta dalam melakukan deteksi bias butir dan mampu mendorong peserta untuk membuat butir soal yang lebih baik atau yang nirbias.

Saran

Berdasarkan hasil penelitian, sebaiknya Dinas Pendidikan Kabupaten/Kota atau sekolah-sekolah mengadakan *in house training* untuk menyu-

sun *learning continuum,* penyusunan instrumen yang mengacu pada *learning continuum*, dan melakukan analisis secara empiris untuk menghindari adanya bias butir pada instrumen yang disusun. Para pengawas di lapangan harus memperoleh pengalaman yang sama agar kualitas guru di lapangan khususnya dalam kompetensi pengembangan instrumen pengukur hasil belajar dapat terus dikembangkan.

Daftar Pustaka

Adams, R.J. & Kho, Seik-Tom. (1996). *Acer quest version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research.

Budiyono. (2005). *Perbandingan metode mantel-haenzel, sibtest, regresi logistic, dan perbedaan peluang dalam mendeteksi keberbedaan fungsi butir*. Disertasi Program S-3 Penelitian dan Evaluasi Pendidikan UNY. Yogyakarta: Program Pascasarjaa UNY.

Embretson, S. & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*. Washington: Winter 2001. Vol. 38, Iss. 4; pg. 343, 26 pgs.

Frisbie, D.A. 2005. Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*. Fall 2005. Vol. 24. No.3. pp.21 28.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. 1991. *Fundamentals of item response theory*. London: Sage Publications, Inc.

Kartowagiran, Badrun. (2005). *Perbadingan berbagai metode untuk mendeteksi bias butir*. Disertasi Program S-3 Psikometri UGM. Yogyakarta: Fakultas Pascasarjana UGM.

Keeves, J.P. & Masters, G.N. (1999). Introduction. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment*. Amasterdam: Pergamon, An imprint of Elsevier Science.

Kolen, MJ & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag New York Inc.

Peraturan Menteri Pendidikan Nasional Nomor 20 Tahun 2007 tentang Standar Penilaian Pendidikan Satuan Pendidikan Dasar dan Menengah.

Peraturan Menteri Pendidikan Nasional Nomor 22 Tahun 2006 tentang Standar Isi untuk Satuan Pendidikan Dasar dan Menengah.

Peraturan Menteri Pendidikan Nasional Nomor 23 Tahun 2006 tentang Standar Kompetensi Lulusan untuk Satuan Pendidikan Dasar dan Menengah.

Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 24 Tahun 2006 tentang Pelaksanaan Peraturan Menteri Pendidikan Nasional Nomor 22 Tahun 2006 tentang Standar Isi untuk Satuan Pendidikan Dasar dan Menengah dan Peraturan Menteri Pendidikan Nasional Nomor 23 Tahun 2006 tentang Standar Kompetensi Lulusan untuk Satuan Pendidikan Dasar dan Menengah.

Stark, S., Chernyshenko, S., Chuah, D.,Wayne Lee, & Wilington, P. (2001). *IRT modeling lab: IRT tutorial* [Versi elektronik]. Urbana: University of Illinois.

Wright, BD & Masters, G.N. (1982). *Rating scale analsis*. Chicago: Mesa Press.