BAB 1

PENDAHULUAN

A. Latar Belakang

Kegiatan reformasi pendidikan di Indonesia, hingga saat ini semakin ditingkatkan karena pendidikan bagi rakyat Indonesia merupakan program penting yang sangat mendasar bagi kemajuan bangsa Indonesia di masa yang akan datang. Berbagai perombakan sistem pendidikan terus dikembangkan dan disosialisasikan. Perombakan tersebut diantaranya adalah perubahan kurikulum. Hal ini mengingat bahwa salah satu kunci untuk menentukan kualitas lulusan adalah kurikulum. Setiap kurun waktu tertentu kurikulum selalu dievaluasi untuk kemudian disesuaikan dengan perkembangan ilmu pengetahuan, kemajuan teknologi, dan kebutuhan pasar. Selain itu, dalam proses pengendalian mutu, kurikulum merupakan perangkat yang sangat penting karena menjadi dasar untuk menjamin tercapainya kompetensi yang diharapkan.

Sejak tahun 1945 hingga saat ini, kurikulum pendidikan nasional telah mengalami perubahan, yaitu pada tahun 1947, 1952, 1964, 1968, 1975, 1984, 1994, 2004, dan 2006. Perubahan kurikulum dalam perkembangan terakhir, telah diberlakukannya kurikulum 2013. Perubahan tersebut merupakan konsekuensi logis dari terjadinya perubahan sistem politik, sosial budaya, ekonomi, ilmu pengetahuan dan teknologi. Kurikulum sebagai seperangkat rencana pendidikan terus dikembangkan secara dinamis sesuai dengan tuntunan dan perubahan yang terjadi di masyarakat. Semua kurikulum nasional dirancang berdasarkan landasan yang sama yaitu Pancasila dan UUD 1945, perbedaannya terletak pada penekanan pokok dari tujuan pendidikan serta pendekatan dalam merealisasikannya.

Seiring dengan berubahnya kurikulum yang berlaku, sistem penilaiannya pun tentu saja juga mengalami perubahan. Hasil dari penilaian tersebut dapat dijadikan gambaran mengenai kompetensi atau penguasaan siswa terhadap materi pelajaran yang telah ditempuhnya. Penilaian memang merupakan salah satu

komponen penting dalam sistem pendidikan karena penilaian dapat berfungsi selain untuk memantau kualitas belajar siswa juga dapat digunakan untuk tujuan akuntabilitas. Penilaian hasil belajar siswa yang akuntabel dan akurat dapat dicapai hanya jika ada kesesuaian antara kurikulum dan apa yang muncul dari siswa pada penilaian. Oleh karena itu perlu untuk memastikan bahwa ada kesesuaian atau kesejajaran antara penilaian dan kurikulum dalam rangka memperoleh kesimpulan yang valid dari hasil penilaian.

Salah satu strategi yang dapat digunakan untuk mengevaluasi kesesuaian antara penilaian dan kurikulum adalah dengan melakukan uji kesejajaran. Bhola, Impara, dan Buckendahl (2003) mendefinisikan kesejajaran sebagai tingkat kesesuaian antara standar isi yang ditetapkan pemerintah dengan penilaian yang digunakan untuk mengukur hasil belajar siswa. Studi kesejajaran akan menunjukkan sejauh mana penilaian yang dilakukan mencerminkan standar isi yang harus dicapai.

Hasil dari studi kesejajaran dapat juga digunakan sebagai bukti validitas untuk mendukung interpretasi skor tes. Ananda (2003a) menyatakan, kesejajaran dapat menjadi sumber untuk bukti validitas isi dan konstruk. Kesejajaran bisa menjadi sumber bukti validitas isi karena berusaha untuk menetapkan sejauh mana tes mencerminkan kurikulum. Selama lebih dari satu dekade, metode untuk mengevaluasi kesejajaran antara penilaian dan kurikulum terus berkembang. Menurut Bhola et al. (2003), metode kesejajaran dapat dikategorikan dalam tingkatan rendah, sedang, dan tinggi terkait kompleksitasnya. Kompleksitas rendah apabila hanya fokus pada perbandingan antara isi butir dan standar. Sedangkan kompleksitas tinggi apabila selain membandingkan butir dan standar juga mempertimbangkan dimensi lain seperti kedalaman isi dan tingkat penekanan dalam penilaian dan kurikulum. Oleh sebab itu, hampir semua metode kesejajaran melibatkan ahli (pakar).

Pemetaan butir berdasarkan teori respons butir dalam perkembangannya menunjukkan bahwa dapat digunakan untuk mengevaluasi kesejajaran antara

penilaian dan kurikulum. Hal ini mengingat bahwa hasil dari evaluasi kesejajaran tidak hanya menunjukkan tingkat kesepakatan antara penilaian dan standar yang tertuang dalam kurikulum, tapi juga perbandingan antara standar dan kinerja siswa yang sebenarnya. Menggunakan kinerja siswa dalam kesejajaran memerlukan definisi pemetaan yang jelas terkait apa yang siswa tahu dan bisa lakukan.

Berdasarkan uraian di atas menunjukkan bahwa metode pemetaan butir berdasarkan teori respons butir dapat digunakan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum. Oleh sebab itu, penelitian ini akan mengembangkan model *alignment* antara penilaian dan kurikulum melalui pemetaan butir. Hasil pengembangan tersebut diharapkan dapat digunakan guru-guru di sekolah untuk mengevaluasi kesejajaran antara penilaian yang dilakukan dengan kurikulum yang digunakan. Melalui penelitian ini, model yang dikembangkan berupa prosedur beserta buku panduannya yang dalam hal ini mengacu pada model kesejajaran Webb.

B. Rumusan Masalah

Rumusan masalah penelitian untuk tahun pertama (2013) adalah:

(1) Bagaimanakah langkah-langkah mengevaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir?
(2) Bagaimanakah buku panduan guna mempermudah kegiatan mengevaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir?

BAB 2

TINJAUAN PUSTAKA

A. Kesejajaran antara Penilaian dan Kurikulum

Kesejajaran antara penilaian dan kurikulum dimaksudkan sebagai terkoordinasinya dengan baik antara penilaian yang dilakukan dengan kurikulum yang digunakan. Hasil dari beberapa studi kesejajaran menginformasikan tentang seberapa baik penilaian telah dilaksanakan sesuai dengan kurikulum dan juga memberikan wawasan tentang apa yang diajarkan di sekolah-sekolah. Kesenjangan konten dalam penilaian dapat ditentukan (Ananda, 2003a) dan informasi tersebut penting bagi para pembuat kebijakan untuk membuat keputusan tentang penilaian dan kurikulum.

Tindal (2005) menambahkan bahwa hasil dari studi kesejajaran dapat digunakan untuk mengidentifikasi daerah di mana standar isi mungkin perlu diperjelas sehingga perkembangan pengetahuan di kelas juga lebih jelas. Hasil dari studi kesejajaran juga dapat digunakan dalam menentukan apakah restrukturisasi penilaian diperlukan atau tidak. Jika restrukturisasi diperlukan, hasil kesejajaran akan membantu untuk mengidentifikasi perubahan yang diperlukan dalam penilaian. Ananda (2003b) juga menyebutkan bahwa hasil kesejajaran dapat digunakan untuk memberikan bukti validitas isi dari sumber eksternal.

Ada lima model yang bisa digunakan untuk studi kesejajaran. Menurut Bhola et al. (2003), model kesejajaran bisa dikategorikan dalam kompleksitas rendah, sedang, dan tinggi. Kategorisasi ini didasarkan pada banyaknya dimensi yang dipertimbangkan dalam model tersebut.

Model CBE (*Council for Basic Education*)

Model CBE menggunakan empat dimensi: konten, keseimbangan konten, ketelitian, dan jenis respon butir (Bhola et al, 2003.). Dimensi konten terlihat pada

perbandingan antara isi butir dan standar. Keseimbangan konten berkaitan dengan distribusi butir-butir menilai standar sementara ketelitian berkaitan dengan perbandingan kompleksitas kognitif antara butir dan standar. Jenis respons butir dilihat pada jenis respon yang dinilai dari peserta didik dengan keterampilan yang ditetapkan dalam standar. Model ini memiliki kelemahan yakni tidak menjelaskan kriteria yang jelas untuk menilai kesejajaran.

Model SEC (*Survey of Enacted Curriculum)*

SEC adalah model kesejajaran dengan kompleksitas sedang. Pengembangan model ini didorong oleh kebutuhan yang dirasakan untuk mengembangkan deskriptor yang seragam dari suatu topik dan kategori kognitif yang bersama-sama dapat menggambarkan isi dari proses pembelajaran (Porter, 2002). Salah satu keunikan dari SEC adalah bahwa model ini tidak hanya berusaha membangun kesejajaran antara penilaian dan kurikulum (standar), tetapi juga termasuk isi dari penggambaran proses pembelajaran secara nyata. Model SEC memiliki dua dimensi dasar yaitu perbandingan isi dan kategori kognitif, yang dinilai secara bersamaan oleh panelis (ahli). Model SEC memuat lima kategori kebutuhan kognitif yakni menghafal, melakukan prosedur, berkomunikasi, memecahkan masalah non-rutin, dan generalisasi/membuktikan.

Model La Marca

Salah satu model kesejajaran dengan kompleksitas tinggi diusulkan oleh La Marca dan rekan-rekannya (2000). Model ini memiliki dimensi perbandingan konten secara mendalam, penekanan konten, perbandingan kinerja, dan aksesibilitas (Bhola et al., 2003). Dimensi perbandingan konten mengevaluasi kesesuaian antara isi dan konten standar penilaian. Perbandingan konten mendalam menilai tingkat kesepakatan antara kompleksitas kognitif yang digariskan dalam standar dan yang tercermin dalam penilaian. Dimensi penekanan mengevaluasi kesesuaian antara bobot yang diberikan pada daerah konten tertentu dalam penilaian dan dalam standar. Menurut La Marca et al. (2000), aksesibilitas dapat dicapai jika penilaian meliputi butir yang bervariasi dalam kesulitan untuk

menutupi berbagai tingkat prestasi di tingkat kelas tertentu. Dengan demikian penilaian harus memberi kesempatan kepada semua peserta didik untuk menunjukkan berbagai pengetahuan dan keterampilan. Keterbatasan utama dari model ini adalah tidak memberikan petunjuk tentang bagaimana masing-masing dimensi dapat dievaluasi. Dengan kata lain, model tersebut tidak memberikan penjelasan pedoman seperti apa tingkat kesepakatan antara penilaian dan standar yang diterima.

Model Webb

Webb (1997) mengembangkan model kesejajaran dengan lima kategori yaitu fokus konten, artikulasi lintas kelas dan usia, keadilan dan kejujuran, implikasi pedagogis, dan sistem penerapan. Setiap kategori memiliki beberapa kriteria untuk menilai kesejajaran. Namun, fokus konten adalah kategori yang telah diterapkan secara luas di sebagian besar studi kesejajaran yang menerapkan model Webb. Model kesejajaran Webb merupakan model yang dapat digunakan untuk membandingkan hasil pada seluruh wilayah negara. Perbandingan ini dimungkinkan karena data kuantitatif yang dihasilkan dari model ini. Namun, hasil dari kesejajaran model Webb kadang bisa menyesatkan. Misalnya, Martone dan Sireci (2009) mencatat bahwa butir yang mengukur hanya bagian dari tujuan yang lebih luas dinyatakan masih dianggap sesuai dengan tujuan. Dengan demikian, hasil dari kesejajaran dapat meningkat sejauh persetujuan kategoris dari berbagai pengetahuan dan keseimbangan representasi yang bersangkutan. Adapun terkait dengan level kemampuan, pada model kesejajaran Webb meliputi empat level yakni Level 1: *recall*, Level 2: *skills and concepts*, Level 3: *strategic thinking*, dan Level 4: *extended thinking*.

Model *Achieve*

Model keselarasan *Achieve* memiliki enam kriteria yaitu akurasi tes, sentralitas konten, sentralitas kinerja, tantangan, keseimbangan, dan jangkauan (Bhola et al. 2003). Proses model kesejajaran ini menggunakan tiga tahap. Pertama adalah analisis butir di mana butir yang dibandingkan dengan standar

untuk mengkonfirmasi draft tes, menilai sentralitas konten, dan mengevaluasi sentralitas kinerja. Tahap kedua, menilai tantangan dalam hal sumber dan tingkat. Tahap terakhir menilai keseimbangan dan jangkauan. Konfirmasi dari draft tes yang diuji melibatkan ahli yang mencocokkan setiap butir ke draft untuk memastikan bahwa setiap butir dalam penilaian tersebut terkait dengan setidaknya satu tujuan dalam standar. Para ahli melakukan ini dengan cara diskusi untuk mencapai konsensus tentang tingkat kecocokan antara butir dan tujuan yang berkaitan. Suatu butir dianggap sesuai dengan tujuan jika mengukur konten yang sama dengan yang ditentukan dalam standar (Rothman, Slattery & Vranek, 2002). Ketersediaan data kualitatif pada model *Achieve* menyediakan pemahaman menyeluruh untuk tingkat kesejajaran. Penggunaan model ini membutuhkan banyak waktu dan personal yang terampil, serta biaya yang tinggi.

Semua model kesejajaran di atas mengandalkan ahli *(judgement)* untuk menilai derajat kesesuaian antara penilaian dan kurikulum (standar). Kualitas hasil kesejajaran tergantung pada seberapa baik ahli memahami kriteria penilaian selama pelatihan. Dalam hal menilai kesejajaran, semua model mengevaluasi perbandingan dalam konten antara penilaian dan standar. Hal ini membantu untuk memeriksa bahwa setiap butir pada penilaian mengukur konten dalam beberapa tujuan. Model-model kesejajaran juga mengevaluasi sejauh mana penilaian mencerminkan luasnya pengetahuan dalam kurikulum (standar). Semua model menilai tingkat kesepakatan antara tuntutan kognitif yang ditentukan dalam kurikulum dan yang dibutuhkan untuk ujian dalam hal memberikan respons yang benar untuk tiap butir pada penilaian.

Sejumlah perbedaan dari beberapa model kesejajaran di atas diantaranya adalah kriteria untuk menilai kesejajaran, kurangnya kriteria untuk menilai kesejajaran membatasi utilitas dari model tersebut. Model kesejajaran yang ada juga berbeda dalam hal tingkat kecermatan untuk mencocokkan penilaian dengan kurikulum. Dalam beberapa metode, pencocokan dilakukan pada tingkat standar yang lebih global. Model Webb adalah satu-satunya model yang dapat mengakomodasi pencocokan pada setiap tingkat standar. Terkait dengan hal ini

menunjukkan bahwa beberapa metode memberikan hasil kesejajaran baik secara kualitatif maupun kuantitatif (misalnya, Webb, SEC, dan Achieve) sementara yang lainnya tidak (misalnya, CBE dan La Marca). Perbedaan lainnya adalah bahwa hanya metode SEC yang menggabungkan instruksi atau proses pembelajaran ke dalam kesejajaran.

Hasil dari evaluasi kesejajaran tidak hanya menunjukkan tingkat kesepakatan antara penilaian dan standar, tapi juga perbandingan antara standar dan kinerja peserta didik yang sebenarnya. Berdasarkan hasil penelitian Kaira, L. T. & Sireci, S. G. (2010) menunjukkan bahwa *item mapping* berdasarkan teori respons butir juga dapat digunakan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum. Penelitian ini didasari pada argumen bahwa kinerja peserta didik dalam kesejajaran memerlukan definisi pemetaan yang jelas terkait apa yang peserta didik tahu dan bisa lakukan.

2. Teori Respons Butir

Pendekatan yang dapat digunakan untuk menganalisis tes selain menggunakan teori tes klasik yakni pendekatan teori respons butir. Pendekatan teori tes klasik memiliki beberapa kelemahan yakni adanya sifat *group dependent* dan *item dependent* (Hambleton, Swaminathan, & Rogers, 1991), juga indeks daya pembeda, koefisien validitas, koefisien reliabilitas skor tes tergantung kepada peserta tes yang mengerjakan tes tersebut. *Group dependent* artinya hasil pengukuran tergantung pada kemampuan kelompok peserta yang mengerjakan tes. Jika tes diujikan kepada kelompok peserta dengan kemampuan tinggi, tingkat kesulitan butir soal akan rendah. Sebaliknya jika tes diujikan kepada kelompok peserta dengan kemampuan rendah, tingkat kesulitan butir soal akan tinggi. *Item dependent* artinya hasil pengukuran tergantung pada tes mana yang diujikan. Jika tes yang diujikan mempunyai tingkat kesulitan tinggi, estimasi kemampuan peserta tes akan rendah. Sebaliknya jika tes yang diujikan mempunyai tingkat kesulitan rendah, estimasi kemampuan peserta tes akan tinggi.

Untuk mengatasi kelemahan-kelemahan yang ada pada teori tes klasik, para ahli pengukuran mengembangkan model pengukuran yang disebut dengan teori respons butir (*Item Response Theory/IRT*). Menurut Hambleton, Swaminathan, & Rogers (1991) serta Hulin, Drasgow, & Parsons (1983), model ini memiliki sifat: (a) statistik butir yang tidak tergantung pada kelompok subjek, (b) skor tes dapat menggambarkan kemampuan subjek, (c) model dinyatakan dalam tingkatan (*level*) butir, tidak dalam tingkatan tes, (d) model tidak memerlukan tes paralel untuk menghitung koefisien reliabilitas, dan (e) model menyediakan ukuran yang tepat untuk setiap skor kemampuan.

Menurut Hambleton, Swaminathan, & Rogers (1991), teori respons butir *(Item Response Theory)* dikembangkan berdasarkan dua buah postulat, yaitu: (a) prestasi subjek pada suatu butir soal dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latent traits*), dan (b) hubungan antara prestasi subjek pada suatu butir dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu, yang disebut kurva karakteristik butir (*Item Characteristic Curve/ICC*). Kurva karakteristik butir ini menggambarkan bahwa semakin tinggi level kemampuan peserta tes, semakin meningkat pula peluang menjawab benar suatu butir.

Model matematis dalam teori respons butir mempunyai makna bahwa probabilitas subjek untuk menjawab butir dengan benar tergantung pada kemampuan subjek dan karakteristik butir. Ini berarti bahwa peserta tes dengan kemampuan tinggi akan mempunyai probabilitas menjawab benar lebih besar jika dibandingkan dengan peserta yang mempunyai kemampuan rendah. Hambleton & Swaminathan (1985) dan Hambleton, Swaminathan, & Rogers (1991) menyatakan bahwa ada tiga asumsi yang mendasari teori respons butir, yaitu unidimensi, independensi lokal, dan invariansi parameter.

Unidimensi, artinya setiap butir tes hanya mengukur satu kemampuan. Contohnya, pada tes prestasi belajar bidang studi matematika, butir-butir yang termuat di dalamnya hanya mengukur kemampuan siswa dalam bidang studi

matematika saja, bukan bidang yang lainnya. Pada praktiknya, asumsi unidimensi tidak dapat dipenuhi secara ketat karena adanya faktor-faktor kognitif, kepribadian peserta tes, dan faktor-faktor pelaksanaan tes, seperti kecemasan, motivasi, dan tendensi untuk menebak. Oleh sebab itu, asumsi unidimensi dapat ditunjukkan hanya jika tes mengandung satu saja komponen dominan yang mengukur prestasi subjek.

Jika faktor-faktor yang mempengaruhi prestasi konstan, maka respons subjek terhadap butir yang manapun akan independen secara statistik satu sama lain. Kondisi ini disebut dengan independensi lokal. Asumsi independensi lokal ini akan terpenuhi apabila jawaban peserta terhadap suatu butir soal tidak mempengaruhi jawaban peserta terhadap terhadap butir soal yang lain. Memenuhi asumsi independensi lokal dapat dilakukan dengan membuktikan bahwa peluang dari pola jawaban setiap peserta tes sama dengan hasil kali peluang jawaban peserta tes pada setiap butir soal. Menurut Hambleton, Swaminathan, & Rogers (1991), independensi lokal secara matematis dinyatakan sebagai:

$$P(u_1,u_2,\ldots,u_n/\theta) = P(u_1/\theta).P(u_2/\theta)\ldots.P(u_n/\theta)$$

$$= \prod_{i=1}^{n} P(u_i/\theta) \quad \text{.....(1)}$$

Keterangan :

i : 1, 2, 3, …n

n : banyaknya butir tes

$P(u_i/\theta)$: probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir ke-i dengan benar

$P(u_1,u_2,\ldots,u_n/\theta)$: probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir ke-1 sampai ke-n dengan benar

Invariansi parameter artinya karakteristik butir soal tidak tergantung pada distribusi parameter kemampuan peserta tes dan parameter yang menjadi ciri peserta tes tidak bergantung dari ciri butir soal. Kemampuan seseorang tidak akan berubah hanya karena mengerjakan tes yang berbeda tingkat kesulitannya dan parameter butir tes tidak akan berubah hanya karena diujikan pada kelompok peserta tes yang berbeda tingkat kemampuannya. Menurut Hambleton, Swaminathan, & Rogers (1991), invariansi parameter kemampuan dapat diselidiki dengan mengajukan dua perangkat tes atau lebih yang memiliki tingkat kesukaran yang berbeda pada sekelompok peserta tes. Invariansi parameter kemampuan akan terbukti jika hasil estimasi kemampuan peserta tes tidak berbeda walaupun tes yang dikerjakan berbeda tingkat kesulitannya. Invariansi parameter butir dapat diselidiki dengan mengujikan tes pada kelompok peserta yang berbeda. Invariansi parameter butir terbukti jika hasil estimasi parameter butir tidak berbeda walaupun diujikan pada kelompok peserta yang berbeda tingkat kemampuannya.

Dalam teori respons butir, selain asumsi-asumsi yang harus dipenuhi tersebut, hal penting lain yang perlu diperhatikan adalah pemilihan model yang tepat. Pemilihan model yang tepat akan mengungkap keadaan yang sesungguhnya dari data tes sebagai hasil pengukuran. Pada teori respons butir, digunakan pendekatan probabilistik untuk menyatakan hubungan antara kemampuan peserta dengan harapan menjawab benar. Pada teori ini, model distribusi yang digunakan yakni distribusi logistik, bukan distribusi normal. Hal ini disebabkan karena kurva normal berbentuk lonceng (Walpole, et al., 2002), sehingga kurva tidak monoton naik. Hal ini menyebabkan untuk suatu kemampuan yang lebih tinggi dari rerata, nilai probabilitasnya justru lebih rendah dibandingkan nilai probabilitas rerata kemampuan. Hal ini bertolak belakang dengan prinsip pengukuran, bahwa peserta dengan kemampuan tinggi mempunyai peluang yang tinggi pula untuk menjawab benar suatu butir instrumen. Pada penghitungan luas daerah di bawah kurva, dapat dilakukan dengan integrasi (Hogg & Craig, 1978), karena merupakan fungsi kerapatan peluang yang kontinu. Dengan adanya variabel kemampuan yang dikuadratkan pada fungsi kerapatan peluang normal, menyebabkan integrasi

menjadi lebih rumit. Hal inilah yang menyebabkan digunakannya model logistik pada teori respons butir.

Ada tiga model logistik dalam teori respons butir, yaitu model logistik satu parameter, model logistik dua parameter, dan model logistik tiga parameter. Perbedaan dari ketiga model tersebut terletak pada banyaknya parameter yang digunakan dalam menggambarkan karakteristik butir dalam model yang digunakan. Parameter-parameter yang digunakan tersebut adalah indeks kesukaran, indeks daya pembeda butir, dan indeks tebakan semu (*pseudoguessing*).

Model logistik tiga parameter ditentukan oleh tiga karakteristik butir yaitu indeks kesukaran butir soal, indeks daya pembeda butir, dan parameter tebakan semu. Dengan adanya tebakan semu pada model logistik tiga parameter, memungkinkan subyek yang memiliki kemampuan rendah mempunyai peluang untuk menjawab butir soal dengan benar. Secara matematis, model logistik tiga parameter dapat dinyatakan sebagai berikut (Hambleton, Swaminathan, dan Rogers, 1991, Hambleton, dan Swaminathan, 1985).

$$P_i(\theta) = c_i + (1-c_i)\ \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad \text{.....(2)}$$

Keterangan :

$P_i(\theta)$: probabilitas peserta tes yang memiliki kemampuan θ dipilih secara acak dapat menjawab butir I dengan benar

θ : tingkat kemampuan subjek

a_i : indeks daya pembeda dari butir ke-i

b_i : indeks kesukaran butir ke-i

c_i : indeks tebakan semu butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes

D : faktor penskalaan yang dibuat agar fungsi logistik mendekati fungsi ogive normal yang harganya 1,7

Sesuai dengan namanya, model logistik tiga parameter ditentukan oleh tiga karakteristik butir yaitu indeks kesukaran butir soal, indeks daya pembeda butir, dan indeks tebakan semu (*pseudoguessing*). Dengan adanya indeks tebakan semu pada model logistik tiga parameter, memungkinkan subjek yang memiliki kemampuan rendah mempunyai peluang untuk menjawab butir soal dengan benar.

Nilai kemampuan peserta (θ) terletak di antara –3 dan +3, sesuai dengan daerah asal distribusi normal. Pernyataan ini merupakan asumsi yang mendasari besar nilai b_i. Secara teoretis, nilai b_i terletak di $-\infty$ dan $+\infty$. Suatu butir dikatakan baik jika nilai ini berkisar antara –2 dan +2 (Hambleton & Swaminathan, 1985). Jika nilai b_i mendekati –2, maka indeks kesukaran butir sangat rendah, sedangkan jika nilai b_i mendekati +2 maka indeks kesukaran butir sangat tinggi untuk suatu kelompok peserta tes.

Peluang menjawab benar pada saat kemampuan peserta tes sangat rendah dilambangkan dengan c_i, yang disebut dengan tebakan semu *(pseudoguessing)*. Parameter ini merupakan suatu kemungkinan asimtot bawah yang tidak nol *(nonzero lower asymtote)* pada kurva karakteristik butir (ICC). Parameter ini menggambarkan probabilitas peserta dengan kemampuan rendah menjawab dengan benar pada suatu butir yang mempunyai indeks kesukaran yang tidak sesuai dengan kemampuan peserta tersebut. Besarnya harga c_i diasumsikan lebih kecil daripada nilai yang akan dihasilkan jika peserta tes menebak secara acak jawaban pada suatu butir.

Model 2 parameter merupakan kasus khusus ketika *pseudoguessing* sama dengan 0, dan model 1 parameter merupakan kasus khusus dari model 3

parameter ketika *pseudoguessing* sama dengan 0 dan daya pembedanya sama untuk semua butir. Adapun estimasi parameter butir dapat dilakukan diantaranya dengan bantuan *software* BILOG, BILOGMG, ASCAL, WINSTEP (1,2,3 parameter), RASCAL, dan QUEST (1 parameter).

Pada model logistik satu parameter, probabilitas peserta tes untuk menjawab benar suatu butir soal ditentukan oleh satu karakteristik butir, yaitu indeks kesukaran butir. Menurut Hambleton, Swaminathan, & Rogers (1991), secara matematis model logistik satu parameter dapat dinyatakan pada persamaan berikut.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$, dengan i : 1,2,3, …,n .....……..... (3)

Keterangan:

θ : tingkat kemampuan peserta tes

$P_i(\theta)$: probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir i dengan benar

b_i : indeks kesukaran butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes

Parameter b_i merupakan suatu titik pada skala kemampuan dalam kurva karakteristik butir ketika peluang menjawab benar peserta tes sebesar 50%. Misalkan suatu butir tes mempunyai parameter b_i = 0,3, artinya diperlukan kemampuan minimal 0,3 pada skala untuk dapat menjawab benar dengan peluang 50%. Semakin besar nilai parameter b_i, maka semakin besar kemampuan yang diperlukan untuk menjawab benar dengan peluang 50%. Dengan kata lain, semakin besar nilai parameter b_i, maka makin sulit butir soal tersebut.

Pada model logistik dua parameter, probabilitas peserta tes untuk dapat menjawab benar suatu butir soal ditentukan oleh dua karakteristik butir, yaitu indeks kesukaran butir (b_i) dan indeks daya pembeda butir (a_i). Parameter a_i merupakan indeks daya pembeda yang dimiliki butir ke-i. Pada kurva karakteristik, a_i proporsional terhadap koefisien arah garis singgung *(slope)* pada titik $\theta = b$. Butir soal yang memiliki daya pembeda yang besar mempunyai kurva yang sangat menanjak, sedangkan butir soal yang mempunyai daya pembeda kecil mempunyai kurva yang sangat landai. Secara teoretis, nilai a_i ini terletak antara $-\infty$ dan $+\infty$. Pada pada butir yang baik nilai ini mempunyai hubungan positif dengan performan pada butir dengan kemampuan yang diukur, dan a_i terletak antara 0 dan 2 (Hambleton & Swaminathan, 1985). Menurut Hambleton, Swaminathan, & Rogers (1991), secara matematis model logistik dua parameter dapat dituliskan sebagai berikut.

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \text{dengan i : 1,2,3, …,n .....……..... (4)}$$

Keterangan :

θ : tingkat kemampuan peserta tes

$P_i(\theta)$: probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir i dengan benar

a_i : indeks daya pembeda

b_i : indeks kesukaran butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes

D : faktor penskalaan yang harganya 1,7

Fungsi informasi butir *(item information functions)* merupakan suatu metode untuk menjelaskan kekuatan suatu butir pada perangkat soal dan menyatakan kekuatan atau sumbangan butir soal dalam mengungkap kemampuan laten (*latent trait*) yang diukur dengan tes tersebut. Dengan fungsi informasi butir diketahui butir mana yang cocok dengan model sehingga membantu dalam seleksi butir soal. Secara matematis, fungsi informasi butir didefinisikan sebagai berikut.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \quad \text{.....} \quad (5)$$

keterangan :

i : 1,2,3,…,n

$I_i(\theta)$: fungsi informasi butir ke-i

$P_i(\theta)$: peluang peserta dengan kemampuan θ menjawab benar butir i

$P'_i(\theta)$: turunan fungsi $P_i(\theta)$ terhadap θ

$Q_i(\theta)$: peluang peserta dengan kemampuan θ menjawab salah butir i

Fungsi informasi butir untuk model logistik tiga parameter dinyatakan oleh Birnbaum (Hambleton & Swaminathan, 1985) dalam persamaan berikut.

$$I_i(\theta) = \frac{2{,}89a_i^2(1-c_i)}{\left[\ (c_i + \exp(Da_i(\theta - b_i))\ \right]\left[1 + \exp(-Da_i(\theta - b_i\ )\ \right]^2} \quad \text{.....} \quad (6)$$

keterangan :

$I_i(\theta)$: fungsi informasi butir i

θ : tingkat kemampuan subjek

a_i : parameter daya beda dari butir ke-i

b_i : parameter indeks kesukaran butir ke-i

c_i : indeks tebakan semu *(pseudoguessing)* butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

Berdasarkan persamaan fungsi informasi di atas, maka fungsi informasi memenuhi sifat: (1) pada respons butir model logistik, fungsi informasi butir mendekati maksimal ketika nilai b_i mendekati θ. Pada model logistik tiga parameter nilai maksimal dicapai ketika θ terletak sedikit di atas b_i dan indeks tebakan semu butir menurun; (2) fungsi informasi secara keseluruhan meningkat jika parameter daya pembeda meningkat. Fungsi informasi tes merupakan jumlah dari fungsi informasi butir-butir tes tersebut (Hambleton & Swaminathan, 1985). Berkaitan dengan hal ini, nilai fungsi informasi perangkat tes akan tinggi jika butir-butir penyusun tes mempunyai fungsi informasi yang tinggi pula. Fungsi informasi perangkat tes *(I(θ))* secara matematis dapat didefinisikan sebagai:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad \text{.....} \quad (7)$$

Nilai-nilai indeks parameter butir dan kemampuan peserta merupakan hasil estimasi. Karena merupakan hasil estimasi, maka kebenarannya bersifat probabilistik dan tidak terlepaskan dengan kesalahan pengukuran. Dalam teori respons butir, kesalahan pengukuran standar *(Standard Error of Measurement, SEM)* berkaitan erat dengan fungsi informasi. Fungsi informasi dengan *SEM* mempunyai hubungan yang berbanding terbalik kuadratik, semakin besar fungsi informasi maka *SEM* semakin kecil atau sebaliknya (Hambleton, Swaminathan, & Rogers, 1991). Jika nilai fungsi informasi dinyatakan dengan $I_i(\theta)$ dan nilai estimasi SEM dinyatakan dengan $SEM\ (\hat{\theta})$, maka hubungan keduanya, menurut Hambleton, Swaminathan, & Rogers (1991) dinyatakan dengan:

$$SEM\ (\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad \text{.....} \quad (8)$$

3. Metode *Item Mapping* Berdasarkan Teori Respons Butir

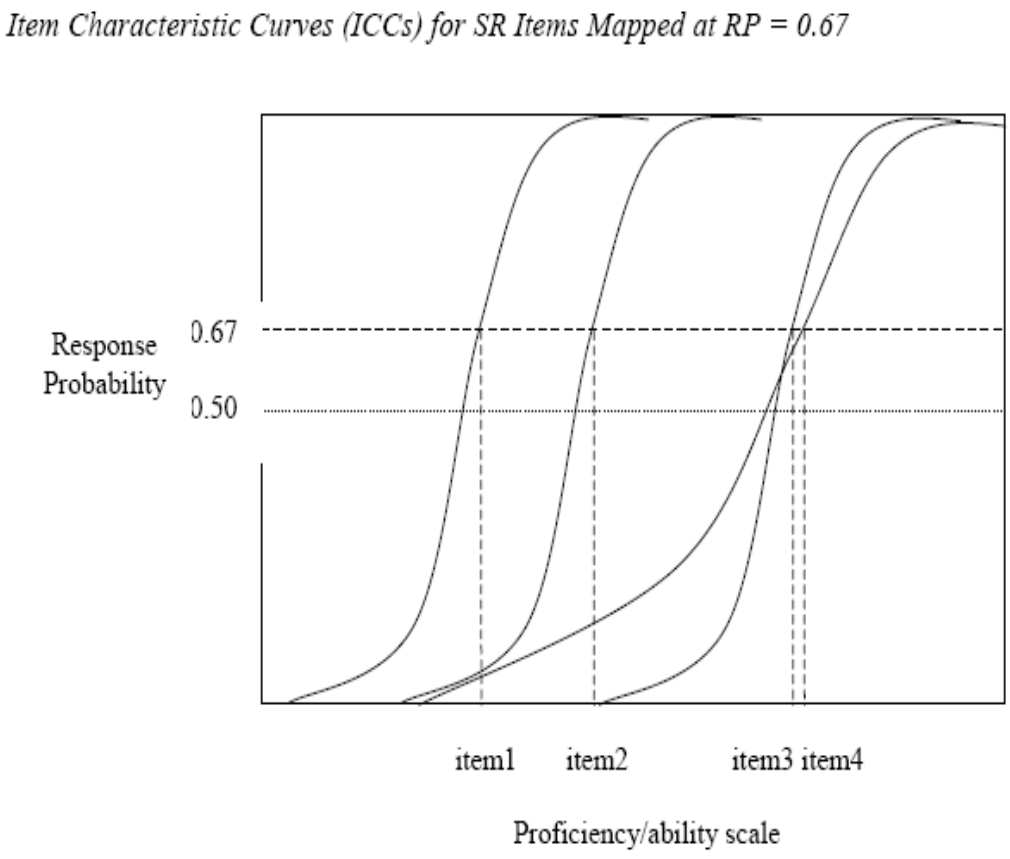
Metode *Item Mapping* (pemetaan butir) dimaksudkan untuk mengidentifikasi dan menjelaskan apa yang siswa tahu dan mampu lakukan pada tingkat penguasaan tertentu. Salah satu pendekatan yang umum digunakan untuk pemetaan butir adalah penggunaan IRT. Beaton dan Allen (1992) menyebutkan ada metode pemetaan butir yaitu metode langsung dan metode *smoothing*.

Salah satu keputusan yang perlu dibuat dalam pemetaan butir adalah mendefinisikan tingkat kesuksesan atau disebut *Respon Probability* (RP). *Respon Probability* digunakan untuk mencari atau memetakan butir sepanjang skala skor dengan tujuan menggambarkan keterampilan siswa pada titik-titik skor tertentu. NRC (2005) menyatakan, keputusan tentang nilai-nilai RP merupakan salah satu hal penting karena mempengaruhi interpretasi tingkat skor.

Berbagai nilai RP seperti 0,50, 0,65, 0,67, dan 0,80 telah diusulkan dan digunakan dalam studi pemetaan butir (Kolstad, et al, 1998;. Zwick, et al,. 2001). Zwick dan rekan-rekannya lebih menyarankan penggunaan RP 0,50 dengan alasan bahwa titik 0,50 atau 50% menandai garis pemisah antara siswa yang tidak bisa dan bisa mengerjakan. Hal ini secara teoritis didukung oleh IRT karena berdasarkan IRT, informasi suatu butir akan maksimum jika probabilitas respon menjawab benar adalah 0,5. (Kolstad, et al.,1998).

Meskipun mendukung penggunaan RP 0,50, Zwick et al. (2001) melaporkan bahwa 50% tidak cukup untuk menunjukkan penguasaan siswa. Argumen untuk RP 0,65 (atau RP 0,67) memajukan gagasan bahwa penguasaan beberapa keterampilan akan menjadi terbukti jika siswa melebihi pada tingkat prestasi tertentu sehingga benar-benar dapat melakukan tugas dibandingkan dengan mereka yang tidak bisa. Para pendukung untuk RP 0,67 berpendapat bahwa jika jumlah peserta ujian yang memberikan respons benar untuk butir sama dengan mereka yang tidak (RP 0,50), maka tidak bisa mengatakan bahwa sebagian besar siswa telah menguasai keterampilan, dengan kata lain, lebih baik digunakan RP besar seperti 0,67.

Huynh (2006) memberikan justifikasi bahwa untuk penggunaan RP 0,67 menunjukkan bahwa untuk setiap butir dikotomus, total informasi yang diberikan oleh respons yang benar dimaksimalkan jika nilai RP lebih besar dari 0,50 untuk satu, dua, dan tiga-parameter model logistik. Huynh (2006) berpendapat bahwa untuk Rasch dan model logistik dua parameter, informasi butir dari respon yang benar diberikan oleh p (1-p) p yang dimaksimalkan ketika p = 0,67. Untuk 3PLM informasi ini diberikan oleh p = (2 + c) / 3, dimana c adalah parameter *pseudo-guessing*. Berikut gambar kurva karakteristik butir yang dipetakan pada RP 0,67.

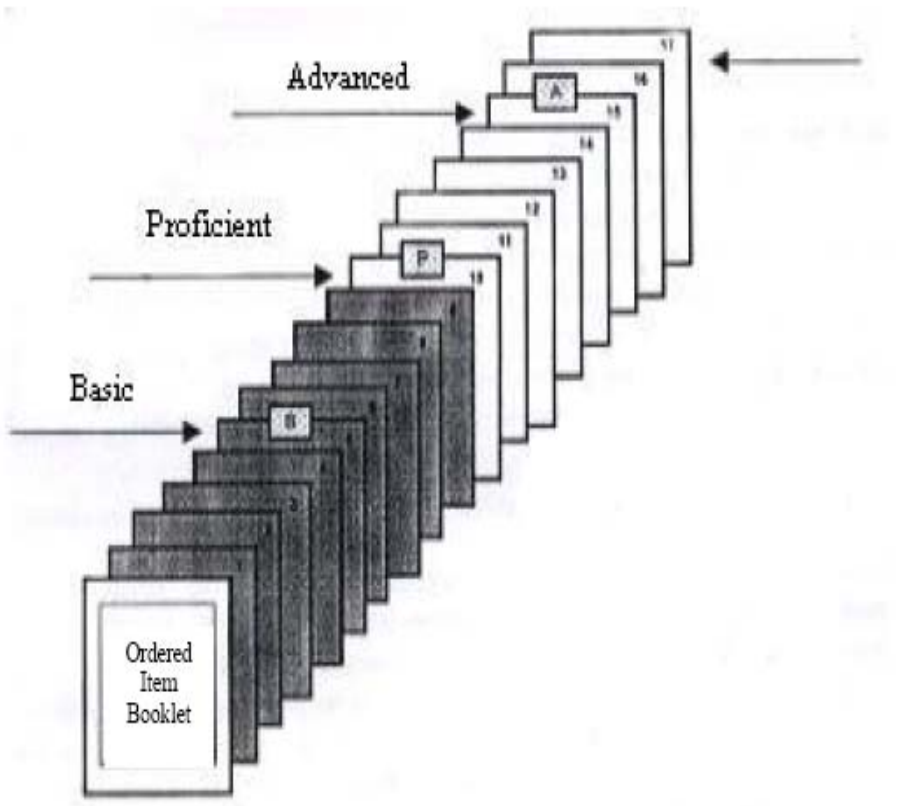


Gambar 1. Kurva Karakteristik Butir yang Dipetakan pada RP 0,67

(Diadaptasi dari Mitzel, Lewis, Patz, & Green (2001), p. 261)

Bahan berupa buku tes atau *Ordered Item Book (OIB)* juga digunakan dalam metode *item mapping*. Menggunakan parameter tingkat kesukaran (b), butir diurutkan dari yang mudah ke yang sulit dalam buku tes. Seperti diilustrasikan

pada Gambar 2 berikut, buku tes dengan butir terurut memiliki satu butir tiap halaman dengan halaman pertama berisi butir termudah dan yang terakhir butir tersulit. Tujuan dari buku tes dengan butir terurut dinyatakan oleh Lewis dkk (1998), yakni untuk membantu panelis menyusun suatu konsep terintegrasi dalam membuat nilai batas (*cut off score).*



Gambar 2. Ilustrasi Buku Tes dengan Butir Terurut (*Ordered Item Book/OIB)*

(Diadaptasi dari Mitzel, Lewis, Patz, & Green (2001), p. 263)

BAB 3

TUJUAN DAN MANFAAT PENELITIAN

A. Tujuan Penelitian

Tujuan penelitian tahun pertama (2013) ini adalah sebagai berikut:

1. Mengembangkan model *alignment* antara penilaian dan kurikulum melalui pemetaan butir yang dalam penelitian ini berupa prosedur atau langkah-langkah dengan mengacu pada model kesejajaran Webb,
2. Menyusun draft buku panduan untuk menunjang pelaksanaan kegiatan mengevaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir.

B. Manfaat Penelitian

Secara teoretis, penelitian ini memberikan manfaat dalam memberikan sumbangan teoretis berkaitan dengan penggunaan metode *item mapping* berdasarkan teori respons butir untuk mengevaluasi kesejajaran antara penilaian dan kurikulum. Selain itu, mengembangkan penelitian yang berkaitan dengan model kesejajaran antara penilaian dan kurikulum. Adapun secara praktis, manfaat penelitian ini diantaranya adalah: (1) Bagi MGMP, guru, dan calon guru pada umumnya, hasil penelitian memberikan pengetahuan dan pemahaman lebih mendalam mengenai metode untuk mengevaluasi kesejajaran antara penilaian dan kurikulum, (2) Buku panduan yang dihasilkan dapat digunakan MGMP untuk mempermudah kegiatan mengevaluasi kesejajaran antara penilaian dan kurikulum di daerah, (3) Bagi pengambil kebijakan di bidang pendidikan, hasil penelitian dapat digunakan sebagai pertimbangan untuk keperluan asesmen baik dalam skala kecil maupun dalam skala besar guna peningkatan dan penjaminan kualitas mutu pendidikan, dan (4) Sebagai bahan pembanding bagi peneliti lain tentang model *alignment* antara penilaian dan kurikulum di masa yang akan datang.

BAB 4

METODE PENELITIAN

A. Langkah-Langkah Penelitian

Penelitian ini menggunakan pendekatan penelitian dan pengembangan (*research and development*), yang terdiri dari dua tahap. Tahap I (tahun I) merupakan penelitian yang dilakukan secara analitis merancang dan menyusun model *alignment* antara penilaian dan kurikulum melalui pemetaan butir dan menyusun draft buku panduan untuk pelaksanaan model *alignment* antara penilaian dan kurikulum melalui pemetaan butir. Model *alignment* antara penilaian dan kurikulum melalui pemetaan butir berupa prosedur atau langkah-langkah yang dalam hal ini model yang dikembangkan mengacu pada model kesejajaran Webb. Pada tahap II (tahun II), model yang dikembangkan diterapkan secara riil menggunakan data dan panelis dari sekolah yang dipilih, selanjutnya dilakukan analisis data untuk mengetahui kelebihan dan kekurangan model yang dikembangkan dan melengkapi serta merevisi buku panduan pelaksanaan model *aligment* antara penilaian dan kurikulum melalui pemetaan butir.

Adapun secara rinci tahapan penelitian tahun pertama (2013) dan tahun kedua (2014) untuk kegiatan penelitian dapat dilihat pada Tabel 1 dan Tabel 2 sebagai berikut.

Tabel 1. Tahapan Penelitian Tahun Pertama (2013)

Kegiatan Penelitian	Hasil yang Ingin Dicapai	Pendekatan yang Digunakan	Metode Pengumpulan Data	Analisis Data
Mengembangkan model *alignment* antara penilaian dan kurikulum melalui pemetaan butir	Prosedur atau langkah-langkah melakukan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir	Kajian analitis	Dokumentasi	Analisis konten berbagai referensi

Menyusun draft buku panduan untuk pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir	Draft buku panduan untuk pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir	Kajian analitis	Dokumentasi	Analisis konten berbagai referensi

Tabel 2. Tahapan Penelitian Tahun Kedua (2014)

Kegiatan Penelitian	Hasil yang Ingin Dicapai	Pendekatan yang Digunakan	Metode Pengumpulan Data	Analisis Data
Penerapan model *aligment* antara penilaian dan kurikulum melalui pemetaan butir	Mengetahui kelebihan dan kekurangan model *aligment* antara penilaian dan kurikulum melalui pemetaan butir	Kuantitatif dan kualitatif	Dokumentasi Wawancara Observasi Pemberian Angket	Analisis kuantitatif dan kualitatif
Merevisi model dan buku panduan pelaksanaan kegiatan mengevaluasi *aligment* antara penilaian dan kurikulum melalui pemetaan butir	Model dan buku panduan pelaksanaan kegiatan mengevaluasi *aligment* antara penilaian dan kurikulum melalui pemetaan butir yang lebih baik dan lengkap	Kuantitatif dan kualitatif	Dokumentasi Wawancara	Analisis Kuantitatif dan kualitatif

B. Lokasi Penelitian

Kegiatan penelitian ini dilakukan di Kotamadya Yogyakarta. Penerapan model dilakukan untuk mata pelajaran matematika pada Sekolah Menengah Atas kelas X di Kotamadya Yogyakarta yang telah menerapkan kurikulum 2013.

C. Judge (Panelis)

Panelis kegiatan ini adalah guru SMA yang dalam penelitian ini dipilih pada mata pelajaran Matematika, berpengalaman mengajar dalam bidang studi yang diteliti minimal 5 tahun dan mengajar kelas X pada mata pelajaran tersebut minimal selama 5 tahun.

BAB 5

HASIL DAN PEMBAHASAN

A. Hasil Penelitian

Penelitian tahun pertama ini lebih menekankan pada ketajaman dan keluasan referensi yang mendukung sehingga tersusun prosedur atau langkah-langkah guna pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui *item mapping* berdasarkan teori respons butir yang benar-benar berkualitas dan dapat dipertanggungjawabkan. Hal ini dilakukan dengan terus meningkatkan dan memperbanyak studi pustaka serta intensif dalam menelusuri berbagai referensi terbaru.

Melalui kajian analitis berbagai literatur diperoleh bahwa untuk mengevaluasi kesejajaran antara penilaian dan kurikulum melalui *item mapping* berdasarkan teori respons butir dilakukan setidaknya dengan langkah-langkah sebagai berikut:

1. Mengurutkan butir dari yang mudah ke yang sulit berdasarkan parameter tingkat kesukaran hasil kalibrasi IRT berdasarkan nilai RP yang dipilih, membuat grafik serta tabel peta butir (*item map*) yang mencakup nomor urut berdasarkan kesulitan butir, nomor asal butir, kunci jawaban, tingkat kesulitan butir, dan *content strand*;
2. Membagi panelis dalam beberapa kelompok kecil dan memilih pimpinan panelis untuk masing-masing kelompok yang duduk terpisah;
3. Memberikan *item map* kepada panelis dan meminta panelis memcermati *item map* dan menempatkan suatu tanda pada titik antara pertanyaan terakhir yang peserta tes kemungkinan menjawab dengan benar dan pertanyaan pertama yang mereka kemungkinannya tidak mampu menjawab dengan benar;

4. Dalam penelitian ini ada dua nilai batas yang ditentukan, karena menempatkan kemampuan siswa pada tiga kategori yaitu *basic* (dasar), *proficient* (mahir), dan *advanced* (lanjut), maka para panelis diminta melanjutkan hingga akhir *item map* dan menempatkan tanda yang lain untuk nilai batas selanjutnya (langkah 3 dan 4 di atas dilakukan dalam tiga putaran);

5. Panelis diminta mendiskusikan antar sesama panelis (diskusi kelompok) dan menyampaikan kepada seluruh panelis mengenai rentang dari penempatan tanda mereka di tiap kelompok dan ada kesempatan pimpinan kelompok menjelaskan khususnya pada butir yang tidak mereka sepakati, lalu berikan kesempatan panelis dari kelompok lainnya untuk menanggapi sebelum kembali ke diskusi masing-masing kelompok;

6. Panelis diminta menempatkan tanda untuk putaran yang ketiga dan mengkalkulasi nilai batas yang diperoleh berdasarkan median penempatan tanda;

7. Menghitung estimasi kemampuan atau theta berdasarkan nilai RP yang dipilih kemudian berdasarkan estimasi theta dan menggunakan nilai batas yang telah diperoleh, memetakan butir ke tiap level yang telah ditentukan dan seluruh panelis juga diminta mengklasifikasi peta butir untuk tiap level dalam bentuk persentase. Level kemampuan yang digunakan dalam penelitian ini menggunakan model kesejajaran Webb;

8. Kesejajaran antara panilaian dan kurikulum terpenuhi jika ada kecocokan antara klasifikasi yang dibuat panelis dengan klasifikasi yang diperoleh berdasarkan peta butir yang dibuat berdasarkan estimasi kemampuan atau theta.

Sebelum langkah-langkah di atas, terlebih dahulu dilakukan kegiatan pelatihan kepada seluruh panelis guna mengkomunikasikan tentang tujuan dan teknis pelaksanaan kegiatan sehingga panelis benar-benar memahami yang harus dilakukan.

Menggunakan data simulasi dengan hasil analisis estimasi parameter tingkat kesukaran dan parameter kemampuan (terlampir), Tabel 3 dan Tabel 4 berikut merupakan contoh *Ordered Item Book (OIB)* dan *item map* yang dapat dibuat.

Tabel 3. Contoh *Ordered Item Book(OIB)*

Butir Soal Matematika	Kemampuan yang Diperlukan untuk Mempunyai Peluang Menjawab Benar	Tingkat Kesukaran	Halaman pada *OIB*
5	-2.277	-3,000	1
7	-1.327	-1.542	2
12	-0.877	-1.531	3
13	-0.567	-1.334	4
27	-0.298	-1.160	5
2	-0.197	-1.020	6
10	-0.150	-0.989	7
9	-0.047	-0.888	8
1	0.072	-0.826	9
.			
.			
.			
.			
14	1.642	1.432	38
6	1.987	1.982	39
19	2.211	3.000	40

Tabel 4. Contoh *Item Map*

Matematika	Kemampuan yang diperlukan untuk mempunyai peluang menjawab benar	Tingkat Kesulitan	Halaman
09	0.06	-0.821	30

Perhatikan premis-premis berikut ini!

1. Jika semua manusia tidak menjalankan agamanya maka kehidupan di dunia rusak.
2. Jika kehidupan di dunia rusak maka kiamat datang
3. Kiamat belum datang

Kesimpulan yang sah dari premis-premis di atas adalah.....

A. Semua manusia menjalankan agamanya.
B. Tidak semua manusia menjalankan agamanya.
C. Sebagian manusia menjalankan agamanya.
D. Sebagian manusia tidak menjalankan agamanya
E. Tidak ada manusia yang menjalankan agamnya.

Kunci: C

Adapun contoh hasil penentuan nilai batas oleh panelis disajikan pada Tabel 5 dan Tabel 6 berikut.

Tabel 5. Nilai Batas dari Panelis pada Tiga Putaran

Antara Level *Basic* (Dasar) dan *Proficient* (Mahir)

Panelis	Putaran 1	Putaran 2	Putaran 3
1	0,265	0,279	0,279
2	0,279	0,279	0,333
3	0,265	0,333	0,333
4	0,265	0,279	0,333
5	0,279	0,279	0,279

6	0,333	0,333	0,333
7	0,333	0,533	0,533
8	0,533	0,533	0,533
9	0,533	0,533	0,533
Median	0,279	0,333	0,333
Rata-Rata	0,315		

Tabel 6. Nilai Batas dari Panelis pada Tiga Putaran

antara Level *Proficient* (Mahir) dan *Advanced* (Lanjut)

Panelis	Putaran 1	Putaran 2	Putaran 3
1	0,813	0,813	0,792
2	0,792	0,813	0,813
3	1,033	0,792	0,813
4	0,813	0,792	0,813
5	0,792	0,792	0,813
6	0,792	1,033	0,813
7	1,033	1,033	1,033
8	0,813	1,033	1,033
9	0,813	0,792	1,033
Median	0,792	0,813	0,813
Rata-Rata	0,806		

Berdasarkan Tabel 5 dan Tabel 6, menunjukkan bahwa nilai batas (*cut off score*) pertama *(Basic-Proficient)*= 0,315 dan nilai batas (*cut off score*) kedua *(Proficient-Advanced)*= 0,806.

Selanjutnya Tabel 7 berikut merupakan contoh level pemetaan butir menurut panelis dan *Item Map*. Level dalam penelitian ini mengacu pada model kesejajaran Webb yakni yakni level 1: *recall*, Level 2: *skills and concepts*, Level 3: *strategic thinking*, dan Level 4: *extended thinking*.

Tabel 7. Level Pemetaan Butir Menurut Panelis dan *Item Map*

Butir	*Panelis*	*Item map*
1	1	1
2	1	1
3	2	1
4	1	1
5	2	2
6	4	3
7	2	2
8	1	2
9	3	2
10	2	2
11	2	2
12	2	2
13	2	2
14	1	1
. .		
38	3	3
39	2	2
40	2	2

Berdasarkan Tabel 7, kesejajaran ditunjukkan oleh kesamaan level baik oleh panelis maupun hasil dari *item map*. Adapun ketidaksesuaian level antara panelis dan *item map* mengindikasikan adanya ketidaksejajaran pada butir tersebut. Ketidaksejajaran ini lebih lanjut dapat ditelusuri melalui diskusi antar panelis.

B. Pembahasan

Pelaksanaan penelitian tahapan tahun pertama (2013) ini dilakukan berdasarkan langkah-langkah kegiatan yang telah direncanakan. Penelitian diawali dengan melakukan kajian analitis berupa analisis konten mencari hubungan berbagai referensi guna tersusunnya prosedur atau langkah-langkah guna pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir. Hasil yang telah diperoleh yakni model *alignment* antara penilaian dan kurikulum yang mengacu pada model webb yang dalam penelitian ini berupa prosedur yang dapat dilakukan setidaknya dalam delapan langkah kegiatan. Contoh simulasi disajikan dalam penelitian ini untuk memperjelas gambaran mengenai pelaksanaan kegiatan mengevaluasi kesejajaran antara penilaian dan kurikulum dengan model yang dikembangkan.

Berdasarkan hasil penelitian tahapan tahun pertama (2013) ini telah dihasilkan dua artikel yang dipublikasikan dalam seminar nasional. Selain itu telah disusun draft buku panduan yang selesai 3 bab dari 5 bab yang direncanakan. Bab 4 dan 5 buku panduan merupakan hasil penerapan dari model yang dikembangkan. Mengingat penerapan model direncanakan baru akan dilakukan pada tahun kedua (2014) maka penyempurnaan buku panduan secara lebih lengkap baru bisa dihasilkan setelah tahun kedua. Demikian halnya dengan publikasi yang dilakukan, dalam penelitian tahun pertama ini masih dalam forum seminar nasional karena hasil penelitian tahun pertama baru berdasarkan hasil kajian analitis berbagai referensi sehingga belum didukung data empiris. Publikasi secara lebih luas melalui jurnal baik nasional atau internasional akan diupayakan mampu dihasilkan setelah penelitian tahun kedua (2014).

Faktor-faktor yang mendukung pelaksanaan penelitian ini antara lain adalah adanya akses untuk memperoleh berbagai referensi serta sarana dan prasarana yang memadai sehingga penelitian dapat dilaksanakan dengan cukup lancar. Faktor yang menghambat pelaksanaan penelitian ini, diantaranya adalah masih jarangnya referensi terkait *alignment* antara penilaian dan kurikulum khususnya di Indonesia. Selain itu, sistem pendidikan di Indonesia yang sekarang untuk beberapa sekolah telah menerapkan kurikulum 2013 menjadi salah satu pertimbangan aplikasi baru akan dilakukan pada penelitian tahapan kedua karena bagi sekolah yang menerapkan kurikulum 2013, saat ini baru akan menyelesaikan proses pembelajarannya untuk satu semester sehingga penilaian yang akan dikaji kesejajarannya juga belum sepenuhnya tersedia.

BAB 6

RENCANA TAHAPAN BERIKUTNYA

Ketercapaian hasil penelitian pada tahapan penelitian tahun pertama (2013) meliputi: (1) Telah dikembangkan secara teoritis, prosedur atau langkah-langkah pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir, (2) Telah dirancang draft buku panduan, dan (3) Telah dihasilkan 2 artikel yang dipublikasikan pada seminar nasional.

Adapun secara riil kegiatan yang masih akan diselesaikan pada tahapan tahun kedua (2014) adalah menerapkan model yang dikembangkan secara riil guna mengetahui kelebihan dan kekurangannya serta merevisi dan melengkapi draft buku panduan pelaksanaan kegiatan *alignment* antara penilaian dan kurikulum melalui pemetaan butir.

Prosedur pelaksanaan kegiatan evaluasi *alignment* antara penilaian dan kurikulum melalui pemetaan butir dalam penelitian ini dibuat secara rinci dalam bentuk buku panduan yang merupakan salah satu dari produk penelitian ini. Produk dari penelitian ini yang berupa buku panduan sementara ini baru tersusun 3 bab dari 5 bab yang direncanakan. Bab 4 dan 5 akan disusun dan keseluruhan isi buku panduan disempurnakan di tahun kedua. Isi buku panduan secara keseluruhan mencakup 5 bab sebagai berikut:

1. Pendahuluan (Pengertian Kesejajaran antara Penilaian dan Kurikulum)
2. Model-Model Kesejajaran antara Penilaian dan Kurikulum
3. *Item Mapping* Berdasarkan Teori Respons Butir
4. *Item Mapping* Berdasarkan Teori Respons Butir untuk Kesejajaran antara Penilaian dan Kurikulum
5. Penutup (Kelebihan dan Kekurangan)

BAB 7

KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian tahapan tahun pertama (2013), telah dikembangkan model *alignment* antara penilaian dan kurikulum yang mengacu pada model kesejajaran Webb berupa prosedur yang memuat delapan langkah yang dalam pelaksanaannya terlebih dahulu memberikan pelatihan kepada seluruh panelis guna mengkomunikasikan tentang tujuan dan teknis pelaksanaan kegiatan. Adapun secara riil, pendalaman tentang hasil pelaksanaan mengevaluasi kesejajaran antara penilaian dan kurikulum yang dilakukan panelis dapat diungkap melalui pengamatan selama proses berlangsung, wawancara, dan pemberian angket. Selain itu, dari hasil penelitian tahun pertama ini telah disusun draft buku panduan pelaksanaan kegiatan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum melalui *item mapping* berdasarkan teori respons butir.

B. Saran

Model yang dikembangkan pada penelitian tahun pertama ini baru disusun berdasarkan kajian analitis berbagai referensi, belum dilengkapi dengan hasil riil dari lapangan. Untuk itu, perlu dilakukan aplikasi metode yang telah dihasilkan guna mengetahui kelebihan dan kekurangannya sehingga model yang dikembangkan lebih akuntabel dan buku panduan yang dihasilkan lebih lengkap dan utuh karena dilengkapi dengan contoh dan data empiris pelaksanaan di lapangan.

DAFTAR PUSTAKA

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Ananda, S. (2003a). *Rethinking Issues of Alignment Under No Child Left Behind*. San Francisco: WestEd.

Ananda, S. (2003b). Achieving Alignment. *Leadership, 33*(1), 18-21.

Beaton A. E., & Allen, N. L. (1992). Interpreting Scales Through Scale Anchoring. *Journal of Educational Statistics, 17*(2), 191-204.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning Tests with Content Standards: Methods and Issues. *Educational Measurement, Issues and Practice*, 2003 (22), 21-29.

Crocker, L. &Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont: Wadsworth Group.

Gall, M. D., Borg, W. R. & Gall, J. P. (1996). *Educational Research: An Introduction.* (6th ed). New York: Longman Publishers.

Herman, J., Webb, N., & Zuniga, S. (2005). *Measurement Issues in Alignment of Standards and Assessment: A Case Study*. (CSE Report 653). Los Angeles; University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Kaira, L. T. & Sireci, S. G. (2010). *Using Item Mapping to Evaluate Alignment between Curriculum and Assessment*. Center for Educational Assessment Research Report Amherst, Massachusetts: School of Education, University of Massachusetts Amherst.

Kolen, M. (2001). Linking Assessments Effectively: Purpose and Design. *Educational Measurement: Issues and Practice*, *20*(1), 5 – 9.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The Response Probability Convention Used in Reporting Data from IRT Assessment Scales: Should NCES Adopt a Standard?* Washington, DC: American Institutes for Research.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A. & Despriet, L. (2000). *State Standards and State Assessment Systems: A Guide to Alignment.* Washington, DC; Council of Chief State Officers.

Martone, A. & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessments, and Instruction. *Review of Educational Research*, *79*(4), 1332 - 1361.

Martone, D., Sireci, S. G., & Delton, J. (2006). *Methods for the Alignment Between State Curriculum Frameworks and State Assessments: A Literature Review*. Center for Educational Assessment Research Report No 603. Amherst, MA: University of Massachusetts, School of Education.

Sireci, S. G. (1998). The Construct of Content Validity. *Social Indicators Research*, *45*, 83 – 117.

Tindal, G. (2005). *Alignment of Alternate Assessments using the Webb System*. Washington, DC; Council of Chief State Officers.

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Research Monograph No. 6). Washington DC: Council of Chief State Officers.

Webb, N. L., M., Ely, R., Cormier, M. &Vesperman, B. (2005). *The WEB Alignment Tool: Development, Refinement, and Dissemination*. Washington, DC; Council of Chief State Officers.

Webb, N. L (2006). *Alignment Analysis of Mathematics Standards and Assessments. Wisconsin, Grades 3-8 and 10*. Diambil pada 2 Oktober 2012 dari http://www.dpi.state.wi.us/oea/pdf/mathsummary06.pdf

LAMPIRAN

Lampiran 1.

Personalia Tenaga Peneliti
Beserta Kualifikasinya

No.	Nama/NIDN	Instansi Asal	Bidang Ilmu	Jabatan dalam Penelitian	Kualifikasi Pendidikan
1	Elly Arliani, M. Si./ NIDN. 0016086705	UNY	Pendidikan Matematika	Ketua Peneliti	S1 Pend Mat Unimed S2 Statistika UGM
2	Kana Hidayati, M. Pd./ NIDN. 0010057702	UNY	Evaluasi Pendidikan Matematika	Anggota Peneliti	S1 Pend Mat UNY S2 PEP UNY

Lampiran 2.

Artikel Hasil Penelitian

ARTIKEL 1

Penerapan *Item Mapping* Berdasarkan Teori Respons Butir dalam Pengukuran Pendidikan Matematika

Abstrak

Oleh: Elly Arliani [*)] dan Kana Hidayati [*)]

[*)] Dosen Jurusan Pendidikan Matematika FMIPA UNY

Item Mapping (pemetaan butir) berdasarkan teori respons butir dimaksudkan untuk mengidentifikasi dan menjelaskan apa yang siswa tahu dan mampu lakukan pada tingkat penguasaan tertentu. Salah satu keputusan yang perlu dibuat dalam pemetaan butir adalah mendefinisikan tingkat kesuksesan atau disebut *Response Probability (RP)* yang digunakan untuk mencari atau memetakan butir sepanjang skala skor dengan tujuan menggambarkan keterampilan siswa pada titik-titik skor tertentu. Keputusan tentang nilai-nilai *Response Probability (RP)* ini merupakan salah satu hal penting dalam *Item Mapping* karena mempengaruhi interpretasi tingkat skor.

Seiring berkembangnya teori pengukuran, *item mapping* berdasarkan teori respons butir dapat digunakan dalam berbagai kegiatan pengukuran pendidikan. Melalui studi literatur, artikel ini membahas penerapan *item mapping* berdasarkan teori respons butir dalam kegiatan pengukuran khususnya dalam pendidikan matematika di Indonesia.

Kata kunci: *Item Mapping,* Teori Respons Butir, Matematika

A. Pendahuluan

Kegiatan reformasi pendidikan di Indonesia, hingga saat ini terus ditingkatkan karena pendidikan merupakan program penting yang sangat mendasar bagi kemajuan bangsa Indonesia di masa yang akan datang. Berbagai perombakan sistem pendidikan terus dikembangkan dan disosialisasikan. Perombakan tersebut diantaranya adalah peningkatan standar kelulusan pada Ujian Nasional (UN) dan perubahan kurikulum yang digunakan.

Sejak tahun 1945 hingga saat ini, kurikulum pendidikan nasional telah mengalami perubahan, yaitu pada tahun 1947, 1952, 1964, 1968, 1975, 1984, 1994, 2004, dan 2006. Perubahan kurikulum dalam perkembangan terakhir, telah mulai diberlakukannya kurikulum 2013. Perubahan tersebut merupakan konsekuensi logis dari terjadinya perubahan sistem politik, sosial budaya, ekonomi, ilmu pengetahuan dan teknologi. Kurikulum sebagai seperangkat rencana pendidikan terus dikembangkan secara dinamis sesuai dengan tuntunan dan perubahan yang terjadi di masyarakat.

Seiring dengan terus berubahnya sistem pendidikan di Indonesia, berbagai konsep dalam teori pengukuran juga terus berkembang, diantaranya konsep *item mapping* berdasarkan teori respons butir. Selama ini kajian mengenai *item mapping* berdasarkan teori respons butir telah cukup banyak dikaji manfaatnya untuk berbagai kegiatan pengukuran pendidikan. Namun pemanfaatannya dalam kegiatan pendidikan di Indonesia masih sangat jarang digunakan. Melalui kajian literatur, artikel ini membahas tentang penerapan *item mapping* berdasarkan teori respons butir dalam kegiatan pengukuran khususnya dalam pendidikan matematika di Indonesia.

B. *Item Mapping* Berdasarkan Teori Respons Butir

Item mapping (pemetaan butir) dimaksudkan untuk mengidentifikasi dan menjelaskan apa yang peserta didik tahu dan mampu lakukan pada tingkat penguasaan tertentu. Secara lebih spesifik, tujuan utama pemetaan butir adalah untuk mengidentifikasi dan menjelaskan prestasi siswa pada tingkat tertentu, apa yang siswa tahu, dan mampu lakukan. Salah satu pendekatan umum untuk pemetaan butir adalah penggunaan teori respon butir atau *Item Response Theory* (IRT).

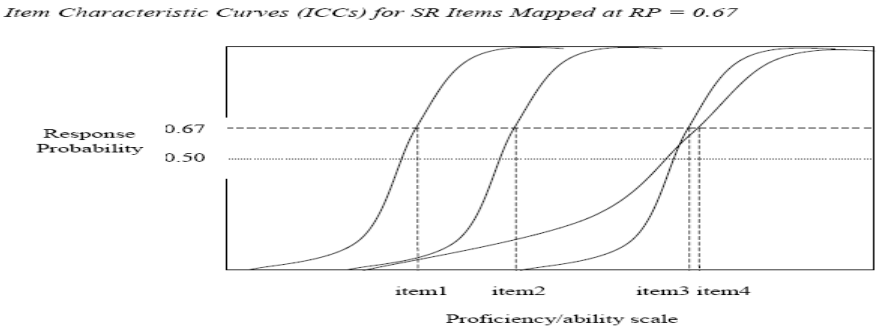
Menurut teori ini, kemampuan peserta tes untuk memberikan jawaban dengan benar dapat diidentifikasi. Terdapat ungkapan "paling mungkin menjawab dengan benar" yang dalam IRT biasanya didefinisikan dengan probabilitas peserta tes memberikan jawaban yang benar untuk suatu butir yang disebut sebagai probabilitas respons atau *Response Probability* (RP). Pemilihan nilai *Response Probability* (RP) berdampak pada hasil pemetaan butir. Melalui pemetaan butir, diperoleh informasi yang menggambarkan apa yang dapat dilakukan siswa.

Informasi ini akan memberikan indikasi tentang bagaimana siswa telah belajar dan berapa banyak yang masih harus dipelajari. Beaton & Allen (1992) menyebutkan ada dua metode pemetaan butir berdasarkan teori respons butir yaitu metode langsung dan metode *smoothing*.

Salah satu hal yang perlu diperhatikan dalam pemetaan butir adalah mendefinisikan tingkat kesuksesan atau disebut *Response Probability (RP)*. *Response Probability (RP)* digunakan untuk mencari atau memetakan butir sepanjang skala skor dengan tujuan menggambarkan keterampilan peserta didik pada titik-titik skor tertentu. Keputusan tentang nilai-nilai *RP* yang digunakan dalam pemetaan butir merupakan salah satu hal penting karena mempengaruhi interpretasi tingkat skor. Berbagai nilai *RP* seperti 0,50, 0,65, 0,67, dan 0,80 telah diusulkan dan digunakan dalam studi pemetaan butir (Kolstad, et al., 1998; Zwick, et al., 2001).

Zwick et al. (2001) lebih menyarankan penggunaan *RP* 0,50 dengan alasan bahwa titik 0,50 atau 50% menandai garis pemisah antara peserta didik yang tidak bisa dan bisa mengerjakan. Hal ini secara teoritis didukung oleh *IRT* karena berdasarkan *IRT*, informasi suatu butir akan maksimum jika probabilitas respons menjawab benar adalah 0,50 (Kolstad, et al.,1998). Meskipun mendukung penggunaan RP 0,50, Zwick et al. (2001) menyatakan bahwa 50% tidak cukup untuk menunjukkan penguasaan peserta didik. Argumen untuk penggunaan *RP* 0,65 atau *RP* 0,67 diajukan, mengingat bahwa penguasaan beberapa keterampilan tertentu akan terbukti jika peserta didik memiliki kemampuan lebih pada tingkat prestasi tertentu sehingga benar-benar dapat melakukan tugas dibandingkan dengan mereka yang tidak bisa. Para pendukung untuk *RP* 0,67 berpendapat bahwa jika jumlah peserta ujian yang memberikan respons benar untuk butir sama dengan mereka yang tidak (*RP* 0,50), maka tidak bisa mengatakan bahwa sebagian besar siswa telah menguasai keterampilan, dengan kata lain, lebih baik digunakan *RP* besar seperti 0,67.

Huynh (2006) memberikan justifikasi bahwa pada penggunaan *RP* 0,67 menunjukkan bahwa untuk setiap butir dikotomus, total informasi yang diberikan oleh respons yang benar dimaksimalkan jika nilai *RP* lebih besar dari 0,50 untuk model logistik satu, dua, dan tiga-parameter. Lebih lanjut Huynh (2006) menyatakan bahwa untuk model Rasch dan model logistik dua parameter, informasi butir dari respons yang benar diberikan oleh $p(1-p)$, yang dimaksimalkan ketika p=0,67. Adapun untuk model logistik tiga parameter, informasi ini diberikan oleh $p=(2+c)/3$, dimana c adalah parameter *pseudo-guessing*. Berikut gambar kurva karakteristik butir yang dipetakan pada *RP* 0,67 dari Mitzel, Lewis, Patz, & Green (2001).



Gambar 1. Kurva karakteristik butir yang dipetakan pada RP 0,67

(Diadaptasi dari Mitzel, Lewis, Patz, & Green (2001), p. 261)

C. Penerapan *Item Mapping* Berdasarkan Teori Respons Butir

Penerapan *Item mapping* berdasarkan teori respons butir dalam pengukuran pendidikan matematika di Indonesia diantaranya adalah dalam kegiatan *standard setting*. Hal ini mengingat bahwa di Indonesia, peningkatan standar berupa batas kelulusan UN khususnya dalam pelajaran matematika merupakan bagian dari kegiatan pengaturan standar (*standard setting)* untuk meningkatkan kualitas pendidikan matematika secara nasional dalam bentuk penentuan batas kelulusan *(cut score)*. Batas kelulusan UN yang sebelumnya dikenal sebagai Evaluasi Belajar Tahap Akhir Nasional (EBTANAS) mengalami peningkatan sejak dikenal sebagai Ujian Akhir Nasional (UAN) pada tahun 2000. *Cut Score* yang digunakan pada tahun 2000 adalah 3,01 atau peserta didik tidak boleh memiliki nilai 3,0 ke bawah. Batas nilai tersebut ditingkatkan pada tahun 2004 yakni menjadi 4,00. Sejak tahun 2005, Ujian Akhir Nasional (UAN) dikenal sebagai Ujian Nasional (UN). Tahun pelajaran 2008/2009 digunakan kriteria rata-rata minimum 5,50, boleh memilki nilai 4,0 pada paling banyak 2 mata pelajaran, lainnya minimum 4,25. Perkembangan selanjutnya, pada tahun 2011 batas kelulusan ditetapkan 5,50 untuk semua mata pelajaran yang diujikan. Adanya peningkatan batas kelulusan, diharapkan kualitas pendidikan juga akan mengalami peningkatan. Selain itu, seiring dengan berubahnya kurikulum yang berlaku, sistem penilaiannya pun tentu saja juga mengalami perubahan. Pada jenjang pendidikan Sekolah Menengah Atas (SMA), salah satu penilaian yang digunakan untuk mengetahui tercapai tidaknya kompetensi berdasarkan kurikulum yang digunakan adalah Ujian Akhir Semester (UAS) dan Ulangan Umum Kenaikan Kelas (UUKK) yang disusun oleh Musyawarah Guru Mata Pelajaran (MGMP) di setiap kabupaten. Kegiatan *Standard setting* kiranya juga dapat diterapkan pada hasil UAS dan UUKK untuk lebih mengetahui kondisi kemampuan siswa yang sebenarnya. Selain itu juga dapat diterapkan dalam kegiatan penentuan Kriteria Ketuntasan Minimal (KKM) yang lebih sesuai dengan kondisi siswa.

Standard setting merupakan kegiatan penentuan batas kelulusan, yakni proses menentukan *cut score* terhadap instrumen pendidikan atau psikologi untuk menjawab pertanyaan "seberapa bagus yang disebut cukup bagus" (George Engelhard, Jr. & Stephen E. Cramer, 1995 yang dikutip Wilson, dkk; 1997). Penentuan standar berupa *cut score* dimaksudkan untuk memutuskan bahwa seseorang dikatakan sudah lulus/kompeten bila telah melewati nilai batas tersebut yakni berupa nilai batas antara peserta didik yang sudah menguasai kompetensi tertentu dengan peserta didik yang belum menguasai kompetensi tertentu. Pengertian tentang *standard* telah banyak dikemukakan para pakar dan juga definisi menurut kamus. *Standard* dapat diartikan sebagai ukuran atau patokan yang disepakati. *Standard setting* adalah proses yang digunakan untuk menentukan atau memilih suatu *passing score* pada suatu ujian. Dari semua langkah-langkah di dalam proses pengembangan tes, *standard setting* merupakan tahapan yang lebih dekat pada seni daripada sains (ilmu pengetahuan) sedangkan metode statistik yang sering digunakan di dalam pelaksanaan suatu *standard setting*, juga lebih banyak melalui pertimbangan dan atau kebijakan.

Komponen esensial dari *standard setting* melalui *judgment* seperti yang dikemukakan oleh Angoff (1971), Ebel (1972), Jaeger (1982), and Nedelsky (1954) adalah panelis atau penilai ahli (Plake, Melican, & Mills, 1991). Jaeger (1991) mengidentifikasi delapan kualifikasi ahli bidang studi (*Subject Matter Expert*, SME) yakni: (1) terbaik dalam bidang spesialisasinya, (2) memiliki wawasan yang luas dalam bidang keahliannya, (3) memiliki kemampuan menyelesaikan masalah dengan cepat sesuai bidangnya, (4) mampu mengkaji secara mendalam level konseptual dalam bidangnya dibandingkan orang baru, (5) menganalisis problem-problem dalam bidangnya secara kualitatif, (6) menilai problem secara lebih akurat dibandingkan orang baru, dan (8) mempunyai daya ingat semantik yang lebik komplek.

Metode *item mapping* dikembangkan berdasarkan IRT (Lord, 1980) yang menggabungkan secara simultan antara karakteristik kemampuan peserta dan tingkat kesulitan butir. Setiap butir yang terskalakan dalam IRT dapat dinyatakan dengan kurva karakteristik yang menyatakan hubungan antara kemampuan peserta terhadap suatu butir. Teori respons butir menyebabkan hal ini memungkinkan untuk mengurutkan berdasarkan kemampuan atau skor skala yang diperlukan suatu probabilitas khusus dari kesuksesan. Butir yang dipetakan tersebut pada suatu lokasi dalam skala IRT sedemikian hingga siswa dengan skor skala dekat pada butir spesifik dapat disimpulkan memiliki pengetahuan keterampilan dan kemampuan yang diperlukan untuk merespon secara sukses pada butir dengan probabilitas khusus.

Pemetaan butir dapat digunakan untuk kegiatan pengaturan standar. Wang (2003) menggambarkan sebuah studi di mana pemetaan butir dapat diterapkan pada penentuan *cut score.* Hasil dari metode pemetaan butir dibandingkan dengan hasil penetapan standar menggunakan metode Angoff. Wang (2003) menemukan bahwa konsistensi antar panelis adalah lebih tinggi untuk metode pemetaan butir daripada untuk metode Angoff. Selain itu, kesepakatan yang teramati antar panelis lebih tinggi dalam metode pemetaan butir daripada Angoff .

Bahan utama yang sering digunakan pada penentuan *standard setting* dengan *item mapping* berdasarkan teori respons butir adalah gambar grafik yang memetakan butir. Menggunakan parameter tingkat kesukaran, butir diurutkan dari yang mudah ke yang sulit dalam bentuk grafik. Prosedur *item mapping* dapat dilaksanakan sebagai berikut: Mengurutkan item dari yang mudah ke yang sulit berdasarkan parameter b hasil kalibrasi IRT; membuat grafik dan tabel peta item (*item map*) yang mencakup nomor urut berdasarkan kesulitan butir, nomor asal item, kunci jawaban, tingkat kesulitan butir, "*content strand*", dan komentar; memilih pimpinan panelis dari panelis yang ada untuk masing-masing kelompok; Menempatkan panelis dalam kelompok kecil yang duduk terpisah; Memberikan Peta Item kepada panelis; Meminta panelis memcermati *item map* dan menempatkan suatu tanda pada titik antara pertanyaan terakhir yang peserta tes kemungkinan menjawab dengan benar dan pertanyaan pertama yang mereka kemungkinanya tidak mampu menjawab dengan benar; Tetap minta panelis mencermati lebih lanjut *item map* untuk mencegah mereka memberikan tanda tanpa melihat dengan cermat; Jika lebih dari satu *cut score* ditentukan, katakan pada panelis untuk melanjutkan hingga akhir *item map* dan menempatkan tanda yang lain untuk *cut score* selanjutnya; Disarankan melakukan langkah-langkah di atas dalam tiga putaran, dimana setelah putaran pertama, berikan masukan kepada panelis tentang tanda yang mereka buat dibandingkan dengan yang lain; Selanjutnya dorong panelis untuk mendiskusikannya antar sesama panelis (diskusi kelompok); Sampaikan kepada seluruh panelis mengenai rentang dari penempatan tanda mereka di tiap kelompok; Beri kesempatan pimpinan kelompok menjelaskan khususnya pada butir yang tidak mereka sepakati, lalu berikan kesempatan panelis dari kelompok lainnya untuk menanggapi sebelum kembali ke diskusi masing-masing kelompok; Minta panelis menempatkan tanda untuk putaran yang ketiga; Mengkalkulasi *cutscore* berdasarkan median penempatan tanda; Meminta panelis me-*review* deskriptor tingkat performansi dan memastikannya kongruen dengan titik potong (*cutpoints*) yang ditentukan saat pertemuan.

Selain dalam *standard setting*, penerapan *Item mapping* berdasarkan teori respons butir juga dapat digunakan untuk mengevaluasi kesejajaran antara

penilaian dan kurikulum pendidikan di Indonesia khususnya dalam mata pelajaran matematika. Penilaian merupakan salah satu komponen penting dari sistem pendidikan karena penilaian dapat berfungsi untuk memantau kualitas belajar peserta didik dan untuk tujuan akuntabilitas. Pemantauan kualitas hasil belajar peserta didik semestinya sesuai dengan kondisi sebenarnya yang terjadi pada peserta didik. Penentuan standar kelulusan mestinya juga tidak hanya berdasarkan keputusan *judgement* semata tetapi juga dapat dipertanggungjawabkan secara nyata kepada masyarakat. Selain itu, penilaian hasil belajar peserta didik yang akurat dapat dicapai hanya jika ada kesesuaian antara kurikulum, apa yang dipelajari peserta didik, dan apa yang muncul dari peserta didik pada penilaian. Oleh sebab itu, perlu untuk memastikan bahwa ada kesesuaian atau kesejajaran antara penilaian dan kurikulum dalam rangka memperoleh kesimpulan yang valid dari hasil penilaian.

Kegiatan mengevaluasi kesejajaran antara penilaian dan kurikulum adalah untuk memastikan bahwa antara penilaian dan kurikulum terkoordinasi dengan baik. Hasil dari beberapa studi kesejajaran menginformasikan tentang seberapa baik penilaian telah dilaksanakan sesuai dengan kurikulum dan juga memberikan wawasan tentang apa yang diajarkan di sekolah-sekolah. Kesenjangan konten dalam penilaian dapat ditentukan (Ananda, 2003a) dan informasi tersebut penting bagi para pembuat kebijakan untuk membuat keputusan tentang panilaian dan kurikulum.

Tindal (2005) menambahkan bahwa hasil dari studi kesejajaran dapat digunakan untuk mengidentifikasi daerah di mana standar isi mungkin perlu diperjelas sehingga perkembangan pengetahuan di kelas juga lebih jelas. Hasil dari studi kesejajaran juga dapat digunakan dalam menentukan apakah restrukturisasi penilaian diperlukan atau tidak. Jika restrukturisasi diperlukan, hasil kesejajaran akan membantu untuk mengidentifikasi perubahan yang diperlukan dalam penilaian. Ananda (2003b) juga menyebutkan bahwa hasil kesejajaran dapat digunakan untuk memberikan bukti validitas isi dari sumber eksternal.

Hasil dari evaluasi kesejajaran tidak hanya menunjukkan tingkat kesepakatan antara standar dan penilaian, tapi juga perbandingan antara standar dan kinerja peserta didik yang sebenarnya. Berdasarkan hasil penelitian Kaira, L. T. & Sireci, S. G. (2010) menunjukkan bahwa *item mapping* berdasarkan teori respons butir dapat digunakan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum. Penelitian ini didasari pada argumen bahwa kinerja peserta didik dalam kesejajaran memerlukan definisi pemetaan yang jelas terkait apa yang peserta didik tahu dan bisa lakukan.

Penerapan *item mapping* berdasarkan teori respons butir dalam mengevaluasi kesejajaran antara penilaian dan kurikulum diantaranya sudah dilakukan oleh Kaira, L. T. & Sireci, S. G. (2010). Namun dalam penelitiannya, memuat kegiatan *standard setting* yang dilakukan secara terpisah oleh pihak lain. Padahal mengingat bahwa baik pada kegiatan *standard setting* maupun evaluasi kesejajaran antara penilaian dan kurikulum, peran utama terletak pada harus adanya *judgement*, sangat dimungkinkan menentukan *cut score* dalam kegiatan *standard setting* dan mengevaluasi kesejajaran antara penilaian dan kurikulum yang dilakukan dalam satu rangkaian kegiatan sekaligus.

Langkah-langkah secara rinci untuk mengevaluasi kesejajaran antara kurikulum dan penilaian melalui item mapping berdasarkan teori respons butir dapat diuraikan sebagai berikut: (1) Mengurutkan butir dari yang mudah ke yang sulit berdasarkan parameter tingkat kesukaran hasil kalibrasi IRT berdasarkan nilai RP yang dipilih, membuat grafik serta tabel peta butir (*item map*) yang mencakup nomor urut berdasarkan kesulitan butir, nomor asal butir, kunci jawaban, tingkat kesulitan butir, dan *content strand*; (2) Membagi panelis dalam beberapa kelompok kecil dan memilih pimpinan panelis untuk masing-masing kelompok yang duduk terpisah; (3) Memberikan *item map* kepada panelis dan meminta panelis memcermati *item map* dan menempatkan suatu tanda pada titik antara pertanyaan terakhir yang peserta tes kemungkinan menjawab dengan benar dan pertanyaan pertama yang mereka kemungkinannya tidak mampu menjawab dengan benar; (4) Jika lebih dari satu *cut score* yang ditentukan, para panelis diminta melanjutkan hingga akhir *item map* dan menempatkan tanda yang lain untuk *cut score* selanjutnya (Disarankan melakukan langkah 3 dan 4 di atas dalam tiga putaran); (5) Panelis diminta mendiskusikan antar sesama panelis (diskusi kelompok) dan menyampaikan kepada seluruh panelis mengenai rentang dari penempatan tanda mereka di tiap kelompok dan ada kesempatan pimpinan kelompok menjelaskan khususnya pada butir yang tidak mereka sepakati, lalu berikan kesempatan panelis dari kelompok lainnya untuk menanggapi sebelum kembali ke diskusi masing-masing kelompok; (6) Panelis diminta menempatkan tanda untuk putaran yang ketiga dan mengkalkulasi *cut score* berdasarkan median penempatan tanda; (7) Menghitung estimasi kemampuan atau theta berdasarkan nilai RP yang dipilih kemudian berdasarkan estimasi theta dan menggunakan *cut score* yang telah diperoleh, memetakan butir ke tiap level yang telah ditentukan dan seluruh panelis juga diminta mengklasifikasi peta butir untuk tiap level; dan (8) Kesejajaran antara panilaian dan kurikulum terpenuhi jika ada kecocokkan antara klasifikasi yang dibuat panelis dengan klasifikasi yang diperoleh berdasarkan peta butir yang dibuat berdasarkan estimasi kemampuan atau theta.

Sebelum langkah-langkah di atas, terlebih dahulu perlu dilakukan kegiatan pelatihan kepada seluruh panelis guna mengkomunikasikan tentang tujuan dan teknis pelaksanaan kegiatan sehingga panelis benar-benar memahami yang harus dilakukan. Informasi mendalam tentang pelaksanaan penentuan *cut score* dan kesejajaran antara penilaian dan kurikulum yang dilakukan panelis dapat diungkap melalui pengamatan selama proses berlangsung, wawancara, dan pemberian angket kepada para panelis terutama terkait keluasan dimensi yang diungkap sesuai model kesejajaran yang digunakan, alasan yang mungkin jika ditemukan adanya ketidaksejajaran, serta tanggapan panelis terhadap serangkaian kegiatan yang telah dilakukan.

F. Simpulan

Berdasarkan kajian di atas, dapat diambil simpulan bahwa dalam pengukuran pendidikan matematika di Indonesia, *item mapping* berdasarkan teori respons butir dapat digunakan untuk kegiatan *standard setting* baik dalam UN maupun UAS dan UUKK serta dalam penentuan KKM. Selain itu, *item mapping* berdasarkan teori respons butir juga dapat digunakan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum yang setidaknya dapat dilakukan dalam delapan langkah dengan terlebih dahulu memberikan pelatihan kepada seluruh panelis guna mengkomunikasikan tentang tujuan dan teknis pelaksanaan kegiatan. Adapun pendalaman tentang pelaksanaan penentuan *cut score* dan kesejajaran antara penilaian dan kurikulum yang dilakukan panelis dapat diungkap melalui pengamatan selama proses berlangsung, wawancara, dan pemberian angket sesuai dengan model kesejajaran yang digunakan.

DAFTAR PUSTAKA

Ananda, S. (2003a). *Rethinking Issues of Alignment Under No Child Left Behind*. San Francisco: WestEd.

Ananda, S. (2003b). Achieving Alignment. *Leadership, 33*(1), 18-21.

Beaton A. E., & Allen, N. L. (1992). Interpreting Scales Through Scale Anchoring. *Journal of Educational Statistics, 17*(2), 191-204.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning Tests With Content Standards: Methods and Issues. *Educational Measurement, Issues and Practice*, 2003 (22), 21-29.

Hambleton, R. K. (1997). Enhancing the Validity of NAEP Achievement Level Score Reporting. *Proceedings of achievement levels workshop*. National Governing Board, Washington, DC.

Huynh, H. (2006). A Clarification on The Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Impara, J.C., & Barbara S Plake. (2000). *A Comparison of Cut Scores Using Multiple Standard Setting Methods.* Universitas Nebraska- Lincoln, Paper presented at the Large Scale Assessment Conference. Snowbird, UT, June, 2000.

Jaeger, R. M. (1991). Certification of Student Competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1994). Selection of Judges for Standard-Setting. *Educational Measurement: Issues and Practice,* 10(2), 3-6, 10.

Kaira, L. T. & Sireci, S. G. (2010). *Using Item Mapping to Evaluate Alignment between Curriculum and Assessment*. Center for Educational Assessment Research Report Amherst, Massachusetts: School of Education, University of Massachusetts Amherst.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The Response Probability Convention Used in* R*eporting* Data *from IRT Assessment Scales: Should NCES Adopt a Standard?* Washington, DC: American Institutes for Research.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The Bookmark Procedure: Psychological Perspectives. In G.J. Cizek (Ed), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors Influencing Intrajudge Consistency During Standard-Setting. *Educational measurement: Issues and Practice,* 10(2), 15-16, 22, 25.

Plake, B. S. & Impara, J. C. (2001). Ability of Panelists to Estimate Item Performance for A Target Group of Candidates: An Issue in Judgmental Standard Setting. *Educational Assessment*, *7*(2), 87 – 97).

Reckase, M. D. (2006b). Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice, 25*(3), 14- 17.

Tindal, G. (2005). *Alignment of Alternate Assessments Using the Webb System*. Washington, DC; Council of Chief State Officers.

Wang, N. (2003). Use of the Rasch IRT Model in Standard Setting: An item mapping Method. *Journal of Educational Measurement, 40*(3); 231-253.

Webb, N. L (2006). *Alignment Analysis of Mathematics Standards and Assessments. Wisconsin, Grades 3-8 and 10*. Retrieved October 14, 2012, from http://www.dpi.state.wi.us/oea/pdf/mathsummary06.pdf

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15-25.

ARTIKEL 2

Model-Model *Aligment* Antara Penilaian dan Kurikulum dalam Pembelajaran Matematika

Abstrak

Oleh: Kana Hidayati [*)] dan Elly Arliani [*)]

[*)] Dosen Jurusan Pendidikan Matematika FMIPA UNY

Berbagai perubahan kebijakan guna meningkatkan kualitas pendidikan di Indonesia terus dilakukan pemerintah. Perubahan kebijakan tersebut diantaranya adalah perubahan kurikulum yang diberlakukan di sekolah saat ini. Terkait dengan perubahan kurikulum tersebut, salah satu komponen penting yang hasus diperhatikan adalah penilaian hasil belajar siswa. Penilaian merupakan salah satu komponen penting dalam sistem pendidikan karena penilaian dapat berfungsi selain untuk memantau kualitas belajar siswa juga dapat digunakan untuk tujuan akuntabilitas. Penilaian hasil belajar siswa yang akuntabel dan akurat dapat dicapai hanya jika ada kesesuaian atau kesejajaran *(aligment)* antara kurikulum, apa yang dipelajari siswa, dan apa yang muncul dari siswa pada penilaian. Oleh sebab itu perlu untuk memastikan bahwa ada kesesuaian atau kesejajaran *(aligment)* antara penilaian dan kurikulum dalam rangka memperoleh kesimpulan yang valid dari hasil penilaian.

Studi kesejajaran dalam pendidikan dimaksudkan untuk menunjukkan sejauh mana penilaian yang dilakukan mencerminkan standar isi yang harus dicapai. Hasil dari studi kesejajaran dapat digunakan juga sebagai bukti validitas untuk mendukung interpretasi skor siswa. Selama lebih dari satu dekade, berbagai metode atau model *aligment* untuk mengevaluasi kesejajaran antara penilaian dan kurikulum terus berkembang. Namun berbagai model *aligment* tersebut selama ini secara spesifik masih jarang dikaji atau diterapkan dalam kegiatan pengukuran pendidikan khususnya di Indonesia. Melalui studi literatur, artikel ini membahas berbagai model *alignment* yang dapat digunakan dalam kegiatan pengukuran pendidikan yakni untuk mengevaluasi kesejajaran antara penilaian dan kurikulum khususnya dalam pembelajaran matematika.

Kata Kunci: Model *Aligment* antara Penilaian dan Kurikulum, Matematika

A. Pendahuluan

Reformasi bidang pendidikan di Indonesia terus ditingkatkan pemerintah karena pendidikan bagi rakyat Indonesia merupakan program penting yang sangat mendasar bagi kemajuan bangsa Indonesia di masa yang akan datang. Berbagai perombakan sistem pendidikan terus dilakukan dan disosialisasikan. Perombakan tersebut diantaranya adalah pengembangan kurikulum. Hal ini mengingat bahwa salah satu kunci untuk menentukan kualitas lulusan adalah kurikulum.

Setiap kurun waktu tertentu, kurikulum selalu dievaluasi untuk kemudian disesuaikan dengan perkembangan ilmu pengetahuan, kemajuan teknologi, dan kebutuhan pasar. Selain itu, dalam proses pengendalian mutu, kurikulum merupakan perangkat yang sangat penting karena menjadi dasar untuk menjamin tercapainya kompetensi yang diharapkan. Sejak tahun 1945 hingga saat ini, kurikulum pendidikan nasional telah mengalami perubahan, yaitu pada tahun 1947, 1952, 1964, 1968, 1975, 1984, 1994, 2004, 2006, dan 2013. Seiring dengan berubahnya kurikulum yang berlaku, sistem penilaiannya pun tentu saja juga mengalami perubahan.

Penilaian merupakan salah satu komponen penting dari sistem pendidikan karena penilaian dapat berfungsi untuk memantau kualitas belajar siswa dan untuk tujuan akuntabilitas. Penilaian hasil belajar siswa yang akurat dapat dicapai hanya jika ada kesesuaian antara kurikulum, apa yang dipelajari siswa pelajari, dan apa yang muncul dari siswa pada penilaian. Oleh karena itu perlu untuk memastikan bahwa ada kesesuaian atau kesejajaran antara kurikulum dan penilaian dalam rangka memperoleh kesimpulan yang valid dari hasil penilaian. Penilaian seharusnya memberikan informasi tentang seberapa baik siswa telah mencapai pengetahuan dan keterampilan yang diharapkan.

Salah satu strategi yang dapat digunakan untuk mengevaluasi kesesuaian antara kurikulum dan penilaian adalah dengan melakukan uji kesejajaran. Menurut Bhola, Impara, dan Buckendahl (2003), kesejajaran merupakan tingkat kesesuaian antara standar isi yang ditetapkan pemerintah dengan penilaian yang digunakan untuk mengukur prestasi siswa. Studi kesejajaran akan menunjukkan sejauh mana penilaian yang dilakukan mencerminkan standar isi yang harus dicapai. Hasil dari studi kesejajaran dapat juga digunakan sebagai bukti validitas untuk mendukung interpretasi skor tes. Ananda (2003a) menyatakan, kesejajaran dapat menjadi sumber untuk bukti validitas isi dan konstruk. Kesejajaran bisa menjadi sumber bukti validitas isi karena berusaha untuk menetapkan sejauh mana tes mencerminkan kurikulum. Bila komponen penilaian dan kurikulum dalam pendidikan memiliki kesejajaran, maka dari proses pendidikan yang

dilangsungkan diharapkan menjadi efisien dan siswa memperoleh serta mampu mencapai kemampuan sesuai dengan apa yang diharapkan (Biggs, 2003).

Selama lebih dari satu dekade, berbagai metode atau model untuk mengevaluasi kesejajaran antara penilaian dan kurikulum terus berkembang. Namun berbagai model *aligment* tersebut selama ini secara spesifik masih jarang dikaji atau diterapkan dalam kegiatan pengukuran pendidikan khususnya di Indonesia. Melalui studi literatur, artikel ini membahas berbagai model *alignment* yang dapat digunakan dalam kegiatan pengukuran pendidikan yakni untuk mengevaluasi kesejajaran antara penilaian dan kurikulum khususnya dalam pembelajaran matematika di Indonesia.

B. Kesejajaran antara Penilaian dan Kurikulum

Tujuan utama dari evaluasi kesejajaran antara penilaian dan kurikulum adalah untuk memastikan bahwa antara penilaian dan kurikulum terkoordinasi dengan baik. Hasil dari beberapa studi kesejajaran memberikan informasi tentang seberapa baik penilaian telah dilakukan sesuai dengan kurikulum yang digunakan. Selain itu, kesenjangan konten dalam penilaian dapat ditentukan (Ananda, 2003a) dan informasi tersebut penting bagi para pembuat kebijakan untuk membuat keputusan tentang penilaian dan kurikulum.

Tindal (2005) menambahkan bahwa hasil dari studi kesejajaran dapat juga digunakan untuk mengidentifikasi muatan dari standar isi yang mungkin perlu diperjelas sehingga perkembangan pengetahuan di kelas juga lebih jelas. Hasil dari studi kesejajaran juga dapat digunakan dalam menentukan apakah restrukturisasi penilaian diperlukan atau tidak. Jika restrukturisasi diperlukan, hasil kesejajaran akan membantu untuk mengidentifikasi perubahan yang diperlukan dalam penilaian. Ananda (2003b) juga menyebutkan bahwa hasil kesejajaran dapat juga digunakan untuk memberikan bukti validitas isi dari sumber eksternal. Terkait penilaian dan kurikulum, Webb (1997) menyatakan bahwa studi kesejajaran memberikan informasi tentang sejauh mana dan seberapa baik antara penilaian dan kurikulum tersebut memfasilitasi dan mampu meningkatkan hasil belajar siswa.

Penilaian yang baik semestinya sejajar dengan kurikulum yang digunakan. Hal ini mengingat bahwa kesejajaran ini penting bagi efektivitas sistem pendidikan (Webb, 1997), pembelajaran siswa (Anderson, 2002; Biggs, 2003; Farenga, Joyce & Ness, 2002; La Marca, Redfield, Musim Dingin, Bailey & Hansche, 2000), keputusan akuntabilitas (Koretz & Hamilton, 2006; La Marca, 2001), evaluasi reformasi pendidikan (Herman, Webb & Zuniga, 2007), validasi terhadap interpretasi dari skor hasil penilaian (La Marca, 2001; Rothman, 2003),

dan memberikan informasi kepada siswa, orang tua, masyarakat dan pengambil kebijakan (Herman, Webb & Zuniga, 2007). Menurut Fuhrman (2001), kesejajaran ini merupakan landasan penting yang harus dapat dipenuhi terutama dalam pendidikan berbasis standar (Fuhrman, 2001).

Berdasarkan uraian di atas, menunjukkan bahwa studi kesejajaran merupakan salah satu kegiatan penting dalam kegiatan pendidikan di Indonesia. Hal ini mengingat bahwa kurikulum yang berlaku di Indonesia merupakan kurikulum yang berbasis standar.

C. Model-Model Kesejajaran

Berdasarkan arah pendekatannya, studi kesejajaran dapat dibedakan menjadi dua yaitu kesejajaran horizontal dan kesejajaran vertikal (Niebling et al., 2008). Kesejajaran horizontal, misalnya menguji kesejajaran pada dua komponen seperti isi kegiatan pembelajaran dengan standar komptetensi pada tingkat kelas yang sama atau menguji kesejajaran pada satu komponen misalnya isi pembelajaran pada dua guru yang berbeda. Adapun kesejajaran vertikal dimaksudkan bahwa uji kesejajaran dilakukan misalnya pada komponen penilaian pada berbagai tingkat kelas yeng berbeda.

Menurut Bhola et al. (2003), model kesejajaran yang berkembang dapat dikategorikan dalam tingkatan rendah, sedang, dan tinggi terkait kompleksitasnya. Kompleksitas rendah apabila hanya fokus pada perbandingan antara butir penilaian dan standar. Sedangkan kompleksitas tinggi apabila selain membandingkan antara butir penilaian dan standar juga mempertimbangkan dimensi lain seperti kedalaman isi dan tingkat penekanan dalam kurikulum dan penilaian. Oleh sebab itu, hampir semua metode kesejajaran melibatkan ahli (pakar). Para ahli ini awalnya dilatih untuk memastikan bahwa mereka mengerti dengan jelas standar, kriteria kesejajaran, dan skala yang digunakan untuk menilai kesejajaran.

Ada lima model yang bisa digunakan untuk studi kesejajaran. Menurut Bhola et al. (2003), model kesejajaran bisa dikategorikan dalam kompleksitas rendah, sedang, dan tinggi. Kategorisasi ini didasarkan pada jumlah dimensi dipertimbangkan dalam model tersebut.

Model CBE (*Council for Basic Education*)

Model CBE menggunakan empat dimensi: konten, keseimbangan konten, ketelitian, dan jenis respons butir (Bhola et al, 2003.). Dimensi konten terlihat pada perbandingan antara isi butir dan standar. Keseimbangan konten berkaitan dengan distribusi butir-butir menilai standar. Jenis respon butir mengevaluasi

kesesuaian jenis respon yang dicari dari siswa dalam menilai kompetensi atau keterampilan yang ditetapkan dalam standar. Namun, model ini memiliki kelemahan diantaranya bahwa pada model ini tidak menjelaskan kriteria yang jelas untuk menilai kesejajaran.

Model Penyelarasan SEC (*Survey of Enacted Curriculum)*

SEC adalah model kesejajaran dengan kompleksitas sedang. Pengembangan model ini didorong oleh kebutuhan yang dirasakan untuk mengembangkan deskriptor yang seragam dari topik dan kategori kognitif yang bersama-sama dapat menggambarkan isi dari proses pembelajaran (Porter, 2002, p. 4). Salah satu keunikan dari SEC adalah bahwa model ini tidak hanya berusaha untuk membangun kesejajaran antara kurikulum (standar) dan penilaian, tetapi juga termasuk isi dari proses pembelajaran ke dalam gambar. Dengan demikian, model keselarasan SEC memuat konten dari penilaian, standar, dan proses pembelajaran. Model SEC memiliki dua dimensi dasar yaitu perbandingan isi dan kategori kognitif, yang dinilai secara bersamaan oleh ahli. Model SEC memuat lima kategori kebutuhan kognitif yakni menghafal, melakukan prosedur, berkomunikasi, memecahkan masalah non-rutin, dan generalisasi/membuktikan.

Model La Marca

Salah satu model kesejajaran dengan kompleksitas tinggi diusulkan oleh La Marca dan rekan-rekannya (2000). Model ini memiliki perbandingan konten secara mendalam, penekanan konten, perbandingan kinerja, dan aksesibilitas sebagai dimensi (Bhola et al., 2003). La Marca, et al. (2000) menganjurkan untuk evaluasi kesejajaran antara penilaian dan standar di luar konten yang sederhana dengan alasan bahwa kecocokan konten dapat dianggap sebagai kondisi yang diperlukan untuk kesejajaran sistem penilaian, tetapi tidak cukup untuk menghasilkan kesejajaran tingkat tinggi saja.

Untuk model La Marca, dimensi perbandingan konten mengevaluasi kesesuaian antara isi penilaian dan konten standar. Perbandingan konten mendalam menilai tingkat kesepakatan antara kompleksitas kognitif yang digariskan dalam standar dan yang tercermin dalam penilaian. Dimensi penekanan mengevaluasi kesesuaian antara bobot yang diberikan pada daerah konten tertentu dalam penilaian dan dalam standar. Menurut La Marca et al. (2000), aksesibilitas dapat dicapai jika penilaian meliputi butir yang bervariasi dalam kesulitan guna mengungkap berbagai tingkat penguasaan di tingkat kelas tertentu. Dengan demikian penilaian harus memberi kesempatan kepada semua siswa untuk menunjukkan berbagai pengetahuan dan keterampilan. Keterbatasan utama dari model ini adalah bahwa hal itu tidak memberikan petunjuk tentang bagaimana

masing-masing dimensi dapat dievaluasi. Dengan kata lain, model tersebut tidak memberikan penjelasan pedoman seperti apa tingkat kesepakatan antara penilaian dan standar yang diterima.

Model Webb

Webb (1997) mengembangkan model kesejajaran dengan lima kategori yaitu fokus konten, artikulasi lintas kelas dan usia, keadilan dan kejujuran, implikasi pedagogis, dan sistem penerapan. Setiap kategori memiliki beberapa kriteria untuk menilai kesejajaran. Namun, fokus konten adalah kategori yang telah diterapkan secara luas di sebagian besar studi kesejajaran yang menerapkan model Webb. Model kesejajaran Webb merupakan alat yang ampuh yang dapat digunakan untuk membandingkan hasil pada seluruh wilayah negara. Perbandingan ini dimungkinkan karena data kuantitatif yang dihasilkan dari model ini. Namun, hasil dari kesejajaran model Webb kadang bisa menyesatkan. Misalnya, Martone dan Sireci (2009) mencatat bahwa butir yang mengukur hanya bagian dari tujuan yang lebih luas, dinyatakan masih dianggap sesuai dengan tujuan. Dengan demikian, hasil dari kesejajaran dapat meningkat sejauh persetujuan kategoris dari ahli berbagai pengetahuan dan keseimbangan representasi yang bersangkutan.

Model *Achieve*

Model keselarasan *Achieve* memiliki enam criteria yaitu akurasi tes, sentralitas konten, sentralitas kinerja, tantangan, keseimbangan, dan jangkauan (Bhola et al. 2003). Proses kesejajaran model ini menggunakan mengikuti tiga tahap. Pertama adalah butir dianalisis dengan analisis butir di mana butir dibandingkan dengan standar untuk mengkonfirmasi draft tes, menilai sentralitas konten, dan mengevaluasi sentralitas kinerja. Tahap kedua, menilai tantangan dalam hal sumber dan tingkat dan tahap terakhir menilai keseimbangan dan jangkauan. Konfirmasi dari uji draft tes melibatkan ahli yang mencocokkan setiap butir draft untuk memastikan bahwa setiap butir dalam penilaian tersebut terkait dengan setidaknya satu tujuan dalam standar. Para ahli melakukan ini dengan cara diskusi untuk mencapai konsensus tentang tingkat kecocokan antara butir dan tujuan yang berkaitan. Butir ini dianggap sesuai dengan tujuan jika mengukur konten yang sama dengan yang ditentukan dalam standar (Rothman, Slattery & Vranek, 2002). Ketersediaan data kualitatif pada model Achieve menyediakan pemahaman menyeluruh untuk tingkat kesejajaran. Informasi ini dapat digunakan untuk meninjau kesejajaran antara penilaian dan standar. Namun, penggunaan model ini membutuhkan banyak waktu dan personal yang terampil, serta biaya yang tinggi.

Berdasarkan uraian di atas, semua model kesejajaran mengandalkan ahli untuk menilai derajat kesejajaran antara penilaian dan kurikulum (standar). Kualitas hasil kesejajaran tergantung pada seberapa baik ahli memahami kriteria penilaian selama pelatihan. Dalam hal menilai kesejajaran, semua model mengevaluasi perbandingan dalam konten antara butir dalam penilaian dan standar dalam kurikulum. Hal ini membantu untuk memeriksa bahwa setiap butir pada penilaian mengukur konten dalam beberapa tujuan. Model-model kesejajaran juga mengevaluasi sejauh mana luasnya pengetahuan dalam penilaian mencerminkan luasnya pengetahuan dalam standar. Kelima model menilai tingkat kesepakatan pada tuntutan kognitif yang ditentukan dalam standar dan yang dibutuhkan untuk ujian guna memberikan respons yang benar untuk butir pada penilaian. Meskipun tingkat tantangan adalah aspek yang sangat penting dalam kesejajaran, semua model menggunakan hasil pembahasan ahli untuk menilai kesejajaran itu.

Sejumlah perbedaan dari beberapa model kesejajaran memberikan kriteria berbeda untuk menilai kesejajaran (misalnya, Webb dan Achieve), sementara yang lainnya tidak (misalnya, La Marca). Kurangnya kriteria untuk menilai kesejajaran membatasi utilitas dari model tersebut. Model kesejajaran juga berbeda dalam hal tingkat detail untuk pencocokan standar dalam penilaian. Dalam beberapa metode, pencocokan dilakukan pada tingkat standar yang lebih global. Model Webb adalah satu-satunya model yang dapat mengakomodasi pencocokan pada setiap tingkat standar seperti perbedaan hasil serta komparabilitas terutama jika komponen yang dievaluasi dalam studi kesejajaran (misalnya, penilaian dan standar) ditulis pada tingkat yang berbeda detail. Terkait dengan hal ini menunjukkan bahwa beberapa metode memberikan hasil kesejajaran baik kualitatif maupun kuantitatif (misalnya, Webb, SEC, dan Achieve) sementara yang lainnya tidak (misalnya, CBE dan La Marca). Hasil kuantitatif dan kualitatif penting dalam membandingkan hasil seluruh wilayah negara dan kekurangan dalam penilaian atau kurikulum. Perbedaan penting lainnya adalah bahwa hanya metode penyelarasan SEC yang menggabungkan proses pembelajaran ke dalam kesejajaran. Hal ini membantu dalam memberikan informasi di bagian dari kurikulum yang berfokus pada guru.

Berdasarkan uraian di atas, menunjukkan bahwa kriteria konten termasuk dalam kerangka taksonomi Webb (Webb, 1997) tapi ternyata secara spesifik tidak menawarkan alat untuk mengkategorikan konten. Oleh karena itu, taksonomi yang ada dalam model Webb masih memilki kelemahan yang perlu diperhatikan khususnya terkait konten kognitifnya.Terkait dengan kompleksitas kognitif, berbagai taksonomi telah dikemukakan para ahli seperti taksonomi Bloom yang direvisi (Anderson & Krathwohl, 2001), DeBlock (de Landsheere, 1990), De

Corte (de Landsheere, 1990), Guilford (1967), Marzano (2001), matriks kinerja-konten Merrill (1994); PISA (OECD, 1999), Porter (Porter & Smithson, 2001a, 2001b) dan TIMSS (Robitaille et al., 1993). Namun taksonomi Bloom yang direvisi (Anderson & Krathwohl, 2001) dalam perkembangannya ternyata juga dapat digunakan untuk mengevaluasi kesejajaran antara penilaian dan kurikulum dan cocok untuk diterapkan dalam pembelajaran matematika (Gunilla N & Henriksson W, 2008).

F. Simpulan

Berdasarkan hasil kajian di atas, dapat diambil simpulan bahwa model-model kesejajaran yang berkembang dan dapat digunakan dalam pendidikan diantaranya adalah Webb, SEC, Achieve, CBE, dan La Marca. Mengingat bahwa sistem pendidikan di Indonesia berbasis standar, penggunaan kelima model tersebut dapat dilakukan namun sebaiknya memperhatikan kelemahan masing-masing model disesuaikan dengan tujuan studi kesejajaran yang dilakukan. Penggunaan model webb dan taksonomi Bloom yang direvisi lebih disarankan dalam kegiatan evaluasi kesejajaran antara penilaian dan kurikulum dalam pembelajaran matematika di Indonesia mengingat kompleksitas kognitif yang dikajinya.

DAFTAR PUSTAKA

Ananda, S. (2003a). *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd.

Ananda, S. (2003b). Achieving alignment. *Leadership, 33*(1), 18-21.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory in Practice, 41*(4), 255-260.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning,teaching, and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning Tests with content Standards: Methods and Issues. *Educational Measurement: Issues and Practice,* 22(3), 21-29.

Gunilla N & Henriksson W. (2008). Alignment of Standards and Assessment: A Theoretical and Empirical Study of Methods for Alignment. *Electronic Journal of Research in Educational Psychology* , 6(3), 667-690

Herman, J., Webb, N., & Zuniga, S. (2005). *Measurement issues in alignment of standards and assessment: A case study*. (CSE Report 653). Los Angeles; University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A. & Despriet, L. (2000). *State Standards and State Assessment Systems: A guide to alignment.* Washington, DC; Council of Chief State Officers.

Martone, A. & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research, 79*(4), 1332 - 1361.

Martone, D., Sireci, S. G., & Delton, J. (2006). *Methods for the alignment between state curriculum frameworks and state assessments: A literature review*. Center for Educational Assessment Research Report No 603. Amherst, MA: University of Massachusetts, School of Education.

Tindal, G. (2005). *Alignment of Alternate Assessments using the Webb System*. Washington, DC; Council of Chief State Officers.

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Research Monograph No. 6). Washington DC: Council of Chief State Officers.

Webb, N. L., M., Ely, R., Cormier, M. &Vesperman, B. (2005). *The WEB Alignment Tool: Development, Refinement, and Dissemination*. Washington, DC; Council of Chief State Officers.

Webb, N. L (2006). *Alignment Analysis of Mathematics Standards and Assessments. Wisconsin, Grades 3-8 and 10*. Retrieved October 20, 2009, from http://www.dpi.state.wi.us/oea/pdf/mathsummary06.pdf

Lampiran 3.

Contoh Hasil Estimasi Parameter Butir dan Kemampuan Menggunakan Program Ascal

MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation

Item Parameter Estimation Program -- ASCAL (tm) Version 3.20

Final Parameter Estimates for Data From File DATAKANA.DAT

Item	a	b	c	N	Chi square	df
1	0.959	-0.826	0.090	2800	32.649	17
2	1.828	-1.020	0.000	2800	71.315	17
3	1.096	-0.514	0.010	2800	39.990	17
4	1.687	0.152	0.340	2800	21.946	17
5	0.745	-3.000	0.320	2800	31.559	17
6	0.539	1.982	0.430	2800	15.965	17
7	0.512	-1.542	0.150	2800	38.777	17
8	0.666	0.524	0.300	2800	12.377	17
9	0.968	-0.888	0.100	2800	34.104	17
10	1.116	-0.989	0.190	2800	34.349	17
11	0.736	0.739	0.090	2800	23.308	17
12	1.057	-1.531	0.200	2800	25.196	17
13	0.857	-1.334	0.130	2800	31.771	17
14	0.804	1.336	0.110	2800	20.268	17
15	1.105	0.122	0.180	2800	17.044	17
16	1.095	0.750	0.170	2800	20.416	17
17	1.407	-0.255	0.220	2800	31.250	17
18	1.983	3.000	0.100	2800	155.453	17
19	2.216	3.000	0.210	2800	149.237	17
20	1.470	0.232	0.270	2800	27.845	17
21	1.042	0.405	0.360	2800	27.982	17
22	1.220	0.674	0.180	2800	25.716	17
23	1.394	0.737	0.500	2800	31.836	17
24	1.344	0.053	0.220	2800	43.769	17
25	1.181	0.511	0.240	2800	33.217	17
26	0.683	-0.737	0.120	2800	18.459	17
27	1.198	-1.160	0.060	2800	43.096	17

28	0.472	-0.687	0.390	2800	31.839	17
29	1.154	1.432	0.190	2800	24.467	17
30	1.519	0.313	0.230	2800	19.404	17
31	0.832	0.914	0.180	2800	19.872	17
32	1.314	1.120	0.230	2800	19.756	17
33	1.452	0.367	0.130	2800	18.573	17
34	2.056	0.055	0.250	2800	8.092	17
35	1.600	0.374	0.220	2800	47.918	17
36	1.495	1.050	0.210	2800	27.926	17
37	0.922	0.156	0.200	2800	27.981	17
38	0.469	-0.201	0.180	2800	52.766	17
39	1.069	0.956	0.100	2800	28.140	17
40	2.129	0.912	0.350	2800	25.834	17

Bayesian theta estimates for examinees from file DATAKANA.DAT

1	1.038
2	1.667
3	0.096
4	2.138
5	1.740
6	1.459
7	1.334
8	1.837
9	0.304
10	1.243
11	0.187
12	1.421
13	0.798
14	1.273
15	2.138
16	0.101
17	0.165
18	0.854
19	1.045
20	0.132
21	1.622
22	1.053
23	1.748
24	0.387
25	0.619
26	-0.008
27	0.973
28	-0.971
29	-0.038
30	-0.184
31	0.728
32	0.871
33	1.329
34	0.655
35	0.102
36	0.368
37	1.175
38	0.269
39	1.125
40	0.096

41	0.670
42	0.707
43	0.420
44	0.615
45	-0.666
46	-0.054
47	2.023
48	-0.299
49	-0.748
50	0.117
51	0.670
52	1.420
53	-0.515
54	1.423
55	1.381
56	1.129
57	1.668
58	0.524
59	0.595
60	0.678
61	0.654
62	1.058
63	-0.020
64	0.863
65	0.256
66	1.188
67	0.686
68	1.297
69	1.183
70	0.824
71	0.394
72	1.737
73	1.026
74	0.987
75	1.168
76	1.458
77	-0.989
78	1.520
79	1.434
80	0.406
81	0.628
82	1.573
83	1.620
84	1.720
85	0.302
86	0.362
87	0.887
88	1.192
89	0.762
90	-0.112
91	0.424
92	-0.518
93	1.915
94	0.925
95	1.489
96	0.578

97	0.186
98	1.267
99	0.675
100	0.178
101	0.764
102	1.239
103	0.493
104	0.372
105	1.665
106	0.294
107	0.763
108	1.021
109	1.092
110	0.825
111	1.575
112	1.706
113	0.381
114	0.978
115	1.554
116	0.911
117	2.138
118	1.149
119	1.097
120	0.465
121	0.947
122	1.266
123	0.761
124	1.192
125	0.521
126	1.074
127	0.553
128	0.365
129	1.278
130	1.594
131	-0.062
132	1.219
133	0.776
134	1.156
135	0.511
136	2.138
137	0.958
138	0.934
139	1.102
140	1.667
.	
.	
.	
2740	-0.389
2741	-0.830
2742	-0.153
2743	0.139
2744	-0.841
2745	-0.314

2746	-0.862
2747	-0.594
2748	-0.395
2749	-0.250
2750	-0.682
2751	-1.012
2752	-0.461
2753	-0.416
2754	-0.280
2755	-0.828
2756	-0.816
2757	-1.673
2758	-0.309
2759	-0.594
2760	-1.592
2761	-0.381
2762	-0.918
2763	-0.632
2764	-1.655
2765	0.110
2766	-0.059
2767	-0.540
2768	-0.385
2769	-1.429
2770	-1.606
2771	-0.094
2772	0.760
2773	0.173
2774	0.484
2775	0.226
2776	-0.898
2777	0.761
2778	-1.031
2779	-0.352
2780	-0.305
2781	0.252
2782	0.427
2783	0.602
2784	0.352
2785	-0.186
2786	1.939
2787	-0.412
2788	-0.672
2289	0.275
2790	0.573
2791	-0.309
2792	0.013
2793	-0.774
2794	-0.551
2795	0.053
2796	0.621
2797	0.132
2798	0.571
2799	0.017
2800	-0.672

Lampiran 4.

Berita Acara dan Daftar Hadir
Seminar Proposal dan Hasil Penelitian

Lampiran 5.

Surat Perjanjian Internal
Pelaksanaan Penelitian Hibah Bersaing

Lampiran 6.

Draft Buku Panduan