P – 31

PENERAPAN *ITEM MAPPING* BERDASARKAN TEORI RESPONS BUTIR DALAM PENGUKURAN PENDIDIKAN MATEMATIKA

Elly Arliani dan Kana Hidayati
Jurusan Pendidikan Matematika FMIPA UNY

Abstrak

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

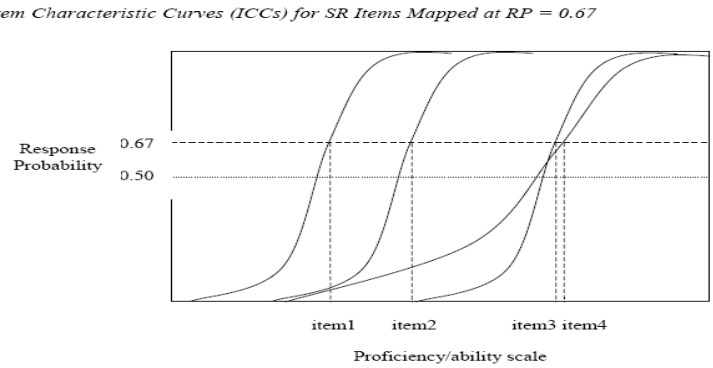
Item mapping (pemetaan butir) dimaksudkan untuk mengidentifikasi dan menjelaskan apa yang peserta didik tahu dan mampu lakukan pada tingkat penguasaan tertentu. Secara lebih

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* Reporting 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.