

ISBN
978-602-96343-0-3

PROSIDING SEMINAR NASIONAL
HIMPUNAN EVALUASI DAN PENDIDIKAN INDONESIA

Aksesmen untuk memantau Kualitas Pendidikan



JAKARTA, 29-30 JANUARI 2010

Wafiq
UIN SYARIF HIDAYATULLAH





PROSIDING SEMINAR NASIONAL

Dalam rangka Kongres HEPI 2010

29-30 Januari 2010 di UIN Syarif Hidayatullan Jakarta

*Artikel-artikel dalam prosiding ini telah dipresentasikan pada
Seminar Nasional dalam rangka Kongres HEPI ke-4
pada tanggal 29-30 Januari 2010
di Universitas Islam Negeri Syarif Hidayatullah Jakarta*

Tim Penyunting Artikel Seminar :

1. Prof. Djemari Mardapi, Ph.D.
2. Dr. Badrun Kartowagiran
3. Dr. Samsul Hadi
4. Dr. Heri Retnawati
5. Dr. Endang Mulyatiningsih

**Himpunan Evaluasi Pendidikan Indonesia
Jakarta 2010**

KATA PENGANTAR

Puji Syukur ke Hadirat Tuhan Yang Maha Esa atas segala Karunia dan Rahmat-Nya sehingga prosiding ini dapat diselesaikan. Prosiding Seminar Nasional Himpunan Evaluasi Pendidikan Indonesia (HEPI) ini merupakan kumpulan hasil penelitian pendidikan dari berbagai daerah di Indonesia. Makalah yang dipresentasikan meliputi makalah utama dan makalah pendamping, terdiri dari makalah hasil penelitian bidang pengukuran pendidikan, bidang evaluasi, dan hasil penelitian pendidikan pada umumnya.

Pada kesempatan ini panitia mengucapkan terimakasih kepada semua pihak yang telah membantu dan mendukung penyelenggaraan seminar ini. Khususnya, kepada seluruh peserta seminar diucapkan terima kasih atas partisipasinya dan selamat berseminar, semoga bermanfaat.

DAFTAR ISI

	Halaman
Halaman Judul	i
Pengantar	ii
Daftar Isi.....	lii
 A. Keynote Speaker	
Asesmen untuk memantau Kualitas Pendidikan (Bahrul Hayat).....	A1-
 B. Pengukuran	
1. Perbandingan Metode Kai-Kuadrat Lord, Luas Daerah Bertanda dan Tidak Bertanda Raju, dan Uji Perbandingan Kemungkinan untuk Mendeteksi <i>Differential Item Function</i> (Etty Rahmawati & Sugiyanto).....	B1- B16
2. Investigasi Keterbandingan Tes Bahasa Studi Internasional PISA (Programme for International Student Assessment) antara Versi Bahasa Inggris dan Versi Bahasa Indonesia (Rahmawati).....	B17-B30
3. Evaluasi Guru SMAN 9 Menggunakan Media ICT dalam Pembelajaran di Kota Bandar Lampung (Herpratiwi).....	B31-B40
4. Identifikasi Bias Butir Perangkat UN Matematika SMP 2003 Menggunakan Teori Respons Multidimensi (Badrin Kartowagiran).....	B41-B54
5. Pengembangan Bank Soal Berbasis Komputer dalam Upaya Pemetaan Kualitas Pendidikan Matematika Tingkat SMP/MTs di Daerah (Rasmuin).....	B55-B80
6. Validasi Standar Kompetensi dengan Pendekatan Metode Angoff dan Bookmark (Nanik Estidarsani).....	B81-B92
7. Penerapan Model Logistik Satu Parameter Pada Computerized Adaptive Testing Ujian Masuk Perguruan Tinggi (Rukli)	B93-B108
8. Identifikasi Kesulitan Belajar Matematika melalui Penerapan Graded Response Model (Mansyur)	B109-B126

VALIDASI STANDAR KOMPETENSI DENGAN PENDEKATAN METODE ANGOFF DAN *BOOKMARK*

Nanik Estidarsani¹

ABSTRAK

Kemampuan *judgment* ahli dalam mempertimbangkan masalah konsep minimal calon berkompeten adalah masalah yang sangat sulit diatasi. Penekanannya adalah pada pendefinisian batas berkompeten, dan pelatihan akan meningkatkan hasil dari setiap prosedur Angoff. Walaupun standar selalu berubah setiap saat sesuai kebutuhan, namun seharusnya memenuhi kaidah pengaturan standar yang dapat dipertanggungjawabkan.

PENDAHULUAN

Penilaian yang menggunakan kriteria seperti halnya tes yang dirancang untuk mengukur seperangkat tes seharusnya mempunyai standar pencapaian tertentu. Namun standar tidak pernah ada batasnya, karenanya harus dikaji secara terus menerus kelayakannya sesuai tuntutan kebutuhan. Penentuan standar untuk sertifikasi, seleksi atau promosi harus dilakukan secara efektif, efisien, akurat dan berkeadilan, demikian juga kebijakan yang akan menyertainya. Kane (2002a:15) dan Goodwin (1996:249) menyatakan bahwa penetapan standar terutama dihubungkan dengan perkembangan dari tes dengan acuan kriteria (sebagai pertentangan dari acuan norma) dalam pengambilan sebuah keputusan. Lebih jauh Shepard (1979) dalam Goodwin (1996:250) menyatakan bahwa penggunaan penilaian beracuan kriteria dilakukan untuk menetapkan apakah individu bisa berhasil dengan ditunjukkannya nilai-nilai setara atau di atas nilai *cut-off*.

Cut score ditulis *cutoff score* atau *cut-off score* adalah sebuah titik penentuan pada suatu skala skor dengan interpretasi yang berbeda antara daerah di atas dan di bawah titik pada skala yang sama (AERA=American Educational Research Association, APA=American Psychological Association & NCME=National Council on Measurement in Education, 1999:175). *Cut-off score* menjadi salah satu kebutuhan dasar seleksi atau

¹ Anggota HEPPI dan Dosen Jurusan Teknik Sipil FT Unesa

promosi dalam bidang ketenaga-kerjaan (AERA, 1999:53), selain itu dapat juga untuk menentukan kompeten atau tidak kompeten atas hasil proses pembelajaran di dalam satuan pendidikan tertentu. *Cut-off score* dalam praktik pendidikan merupakan informasi persyaratan batas minimal dari hasil belajar yang menentukan kebijakan lolos tidaknya seseorang pada suatu hasil tes yang dinyatakan oleh suatu lembaga, DIKLAT (Pendidikan dan Latihan) atau sekolah. Goodwin (1996:249) menyatakan bahwa studi tentang *cut-off scores* oleh para ahli pengukuran sering diarahkan kepada sebuah penentuan standar (*standard setting*).

Metode *Standard Setting*

Standard setting (penentuan standar) merupakan prosedur untuk menjamin terhadap tingkat keterampilan yang layak. Metode *standard-setting* tradisional mengembangkan sebagian besar untuk konteks dua kategori saja (seperti berhasil atau gagal). Suatu program uji nasional terkemuka, National Assessment of Educational Progress (NAEP), mengembangkan program pengujian yang menyatakan *performance* dengan rangkaian tingkatan *performance* dinilai: *Basic*, *Proficient*, dan *Advanced* (Cizek, 2004:31). *Standard setting* menjadi penting sejalan dengan perubahan pendidikan berbasis standar, hal ini tumbuh dari tuntutan publik karena mereka juga ikut bertanggung-jawab terhadap pendidikan.

Pendekatan metode *standar setting* yang mudah dijelaskan dan prosedurnya mudah dilaksanakan adalah metode modifikasi Angoff dan metode Bookmark (Goodwin, 1996:253-254), walaupun terdapat perbedaan kecil dari nilai rerata antara kelompok Angoff dan *Bookmark* pada putaran berikutnya, standar deviasi metode *Bookmark* menurun dibandingkan dengan metode Angoff (Buckendahl, 2002:260). Presentasi data berdasarkan norma dalam modifikasi Angoff dapat meningkatkan reliabilitas *inter-judge* dan keefektivan metode *Bookmark* dapat direkomendasikan pada bidang subyek yang berbeda. Kedua metode *standard setting* tersebut dapat saling mengisi kelemahan di lain pihak.

Profil *rater* sebagai *judgment* ahli adalah pengambil keputusan dalam diskusi panel (dengan pendekatan teknik Delphi dan FGD=*focus Groups Discussion*) yang akan dilibatkan untuk *standard-setting*. Beberapa parameter dan strategi spesifik

disyaratkan oleh Bennett (1998:9-11) bila menggunakan prosedur *standard-setting* HSC (*Higher School Certificate*) yang terkait dengan syarat *rater* adalah sebagai berikut: a) Profil *rater* yang terlibat dalam tugas *standard-setting* harus sangat berpengalaman dalam hal mengajar, dan b) Jumlah *rater* dalam tim, disarankan kira-kira 6 orang dalam satu tim, karena memungkinkan terjadinya diskusi, pada tim yang besar diskusi akan menjadi dangkal. Antar *rater* mempunyai pemahaman cukup dan bidang keahlian dalam semua aspek menyangkut materi yang sedang diuji. Selanjutnya Bennet menyarankan penggunaan pendekatan kompensasi, artinya *score* dalam suatu pengujian merupakan penjumlahan yang diperoleh pada tiap-tiap materi. Selain itu penggunaan pendekatan kompromi dalam suatu sistem berbasis standar dapat dikalibrasi maka perkembangan dari tahun ke tahun harus ditetapkan untuk mengetahui keterkaitannya dengan pola kemampuan seseorang.

Metode Extended Angoff

Metode Angoff (selanjutnya disebut metode Angoff tradisional) dan metode Angoff yang dimodifikasi (selanjutnya disebut metode modifikasi Angoff atau *extended Angoff*) sering digunakan oleh para ahli psikometri, ahli pengukuran, dan peneliti. Kedua metode yang diperbandingkan ini dipercaya karena reliabilitas asesmen *intra* dan *inter-judgment* tinggi (Kane, 1994b dalam Goodwin, 1996:252-253). Kelemahan metode Angoff tradisional dibanding dengan metode modifikasi Angoff adalah pada a) pertimbangan batas *score* (merupakan rerata dari para *judge*), dan penentuan keputusan didasarkan pada *judgement* (Ricker, <http://education.ualberta.ca/educ/psych/crame/files/RickerSCCE2003.pdf> diambil 27 Oktober 2005:29); b) kesulitan para *judge* dalam hal memberikan pertimbangan secara konsep batasan dan meminimalkan *borderline-group* atau kandidat yang berkualitas (Goodwin 1996:254); dan c) ketergantungan pada situasi sosial dan politik bagi pengembang standar dalam menentukan *cut-off score*. Kelemahan metode Angoff tradisional ini kemudian dimodifikasi dengan yang disebut metode modifikasi Angoff. Kelebihan modifikasi Angoff yaitu lebih mudah mengestimasi proporsi dan dapat mengurangi *standard error* melalui diskusi antar *judge* (untuk meningkatkan reliabilitas *inter-judge*). Prosedur

metode modifikasi Angoff menunjukkan secara tepat hal yang spesifik dari modifikasinya.

Prosedur perluasan Angoff yang paling sering dimodifikasi adalah: 1) menyediakan pelatihan luas untuk penilai (*judgment* ahli) dalam proses mengidentifikasi target testee dan mengestimasi butir-butir kinerja testee, 2) menyediakan data kinerja nyata kepada penilai dan dengan dampak proporsi berhasil atau tidak berhasil yang dikaitkan dengan penilaian *cut score*, dan 3) memberikan lebih dari satu kesempatan untuk mengestimasi kinerja testee (Plake, 1998 dalam Impara, 2002:3). Waktu antara putaran digunakan untuk diskusi di antara anggota panel. Tujuan ~~tentang~~ diskusi adalah untuk meningkatkan persetujuan antar *judgment* ahli yaitu mengurangi kesalahan standar. Busch dan Jaeger, 1990 dalam Ricker (2003:6) menemukan suatu proses *iterative* untuk mengurangi variabilitas antar *judgment* ahli untuk estimasi butir dan *cut-off score* secara keseluruhan. Nilai perkataan berulang-ulang yang benar untuk pengaturan *cut-off score* terbatas oleh kemampuan ~~tentang~~ anggota *judgment* ^{dalam} membuat pertimbangan mandiri, dan moderator menjaga selama proses diskusi. Presentasi dari data berdasarkan norma dalam modifikasi Angoff ditunjukkan untuk meningkatkan reliabilitas *inter-judge*.

Secara umum metode *standard setting* Angoff (1971) memerlukan panelis ahli untuk 1) merumuskan kadidat yang memiliki kepedulian (berkualifikasi minimal) dan 2) mengestimasi performansi item sebenarnya untuk para kandidat tersebut. Dalam studi *standard setting*, suatu diskusi panel *judgment* ahli diminta membuat suatu prediksi tentang seberapa kemampuan minimal peserta tes untuk menjawab dengan benar suatu item tes. Panelis selaku *rater* adalah orang-orang yang ahli di bidang yang diujikan. Jaeger (1991) mengidentifikasi delapan kualifikasi seorang profesional *judgment* ahli untuk studi *standard setting*, yakni 1) terbaik dalam bidang spesialisasinya; 2) memiliki konsep yang/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; (bekerja keras dan terampil melakukan evaluasi diri; dan 8) mempunyai daya ingat semantik

yang lebih kompleks. Di samping itu, Jaeger (1991) mengemukakan 2 kualifikasi tambahan untuk diasumsikan, yaitu 1) memiliki *skill* yang cukup untuk mengkonseptualisasi kelompok kandidat guru yang memiliki kemampuan minimal dengan mengidentifikasi *level skill* dan prestasi khusus dari grup kandidat ini; dan 2) memprediksi seberapa besar proporsi individu mampu menjawab dengan benar setiap item tes.

Metode Bookmark

Metode Standard setting *Bookmark* dikembangkan untuk mengatasi kekurangan pada prosedur Angoff dalam menentukan *cutscores* (Lewis, Mitzel, & Green, 1996 dalam Buckendahl, Smith, Impara, dan Plake 2002:254). Satu kelebihan penting dari metode ini adalah menggunakan prosedur IRT (*Item Response Theory*) untuk memetakan butir sesuai tingkat kesulitan. Artinya metode ini mengatur butir-butir dari tingkat kesukaran butir yang paling rendah sampai ke yang paling sulit. Selain mengatur tingkat kesukaran butir dapat memetakan butir pada distribusi keahlian (θ) sesuai standar *cutoff score* yang ditentukan (Karantonis dan Sireci, 2006:5). Walaupun ada perbedaan pada penerapan metode *standard-setting*, namun satu kelebihan penting metode *Bookmark* yang didukung oleh Karantonis dan Sireci (2006:4), adalah pada OIB (*Ordered Item Booklet*).

OIB ini berisi satu rangkaian butir tersusun dengan tingkat kesulitan dari yang paling mudah sampai yang paling sulit. Butir-butir yang sulit ditentukan secara empiris, khususnya yang menggunakan IRT (seperti parameter b , untuk butir yang diskor secara diskotom, dan parameter ambang untuk masing-masing *score point* pada butir skor politomus). Kelebihan metode *Bookmark* yang lain adalah 1) mengakomodasi respon yang dibangun seperti halnya butir tes *selected-response* (SR); 2) lebih efisien dalam mengakomodasikan berbagai *cutscores* dan berbagai format tes; dan 3) mengurangi kompleksitas kognitif *judgment ahli*. Para pengembang *Bookmark* menyarankan untuk menggunakan model 3PL (3-parameter logistik) untuk item-item SR, dan model kredit partial 2PL (2-parameter logistik) untuk item-item CR (*criterion-response*). Lewis, Cizek, 2004 dalam Karantonis dan Sireci, 2006:5, menunjuk bahwa model (Rasch) 1PL (1-parameter logistik) untuk item CR dan SR yang kadang-kadang digunakan.

Reckase (2006:6) dalam penelitiannya melibatkan teori psikometri dalam penentuan standar dengan mengevaluasi keberfungsian variasi metode *standard setting* yaitu metode modifikasi Angoff dan *Bookmark*. Kerangka konsep teori psikometri tersebut digunakan untuk memperoleh tiga kriteria-kriteria dalam mengevaluasi proses *standard setting*. Dalam konteks ini teori psikometri yang diterapkan pada *standard setting* secara kasar dianalogikan pada teori *true score* untuk tes prestasi Reckase (2006:4). Tujuan proses *standard setting* ini adalah untuk mengkonversi standar-standar yang diharapkan panelis untuk menuju kepada satu skala skor tes.

Reckase mengembangkan dua prosedur yang berbeda terkait dengan ICS (*Intended Cut Score*). Prosedur pertama bahwa konsep ICS digunakan untuk mengembangkan kriteria statistik dalam mengevaluasi implementasi prosedur *standard setting*. Dari perspektif ini ICS sebagai konsep-konsep *true score* atau kemampuan (θ) yang digunakan untuk mengembangkan metodologi psikometri, seperti estimasi tingkat kemampuan atau model-model IRT. Prosedur kedua bahwa cara berpikir ICS adalah sebagai kontrak hipotetis yang menjadi target pada proses *standard setting*. ICS merupakan operationalisasi ideal estimasi panelis terhadap kebijakan dan uraian kompetensi. Perspektif ini memerlukan jenis yang sama dari studi-studi validasi yang diperlukan untuk tes yang mendukung untuk mengukur satu kontrak hipotetis. Bukti harus dikumpulkan untuk mendukung validitas standard. Kerangka konsep ICS dari Reckase (2006:5) mengarahkan pada tiga kriteria yang dapat digunakan untuk proses *standard setting*. Tiga kriteria tersebut adalah: 1) ICS panelis diperoleh dari proses estimasi *standard setting* secara konsisten dengan ICS; 2) *error* yang dihitung atas pertimbangan panelis untuk proses *standard setting* didasarkan percobaan pikiran hipotetik tanpa memperhitungkan sesi sebelumnya; dan 3) simpangan baku distribusi hipotetik dibuat sekecil mungkin.

Hasil kesepakatan ahli merupakan data tingkat kesepakatan dari metode *standard setting* dengan format berbeda pada tiap metode (*Extended Angoff* dan *Bookmark*). Data metode *Extended Angoff* berbentuk skala tingkat kesulitan (5 skala), tingkat relevansi indikator (4 skala) dan pembobotan (%) subkompetensi dari kompetensinya, sedangkan data metode *Bookmark* menggunakan skala tingkat kesulitan yaitu sukar, sedang dan mudah.

METODE PENELITIAN

Uji kesepakatan visi pada kompetensi dilakukan dengan teknik Delphi melalui angket yang diisi oleh sejumlah 8 (delapan) ahli di bidang ketekniksipilan. Pertimbangan pendekatan menuju kesepakatan yang digunakan adalah menggunakan metode Angoff dan Bookmark. Kesepakatan dari kedua metode pendekatan diatur sesuai persyaratan dan prosedur metode. Komponen kompetensi dalam konstruksi tes berupa kesepakatan bentuk subkompetensi, kesesuaian indikator dan deskriptor serta isi dan jumlah butir tes. Sebagai contoh dalam penelitian ini kesepakatan 8 (delapan) orang maka dinyatakan bahwa 100% ahli sangat sepakat, bila terjadi kesepakatan 7 (tujuh) orang maka dinyatakan bahwa 87,5% ahli sepakat, bila terjadi kesepakatan 6 (enam) orang maka dinyatakan bahwa 75% ahli cukup sepakat, namun bila terjadi kesepakatan kurang dari 62,5% atau kurang dari 5 (lima) orang maka subkompetensi, indikator, deskriptor atau isi butir tes harus diperbaiki atau digugurkan.

Uji kelompok kecil dan saran para ahli dilakukan terutama terhadap aspek pemahaman konsep dasar, prinsip bidang studi ketekniksipilan dengan maksud untuk mendapatkan informasi tentang (1) estimasi tingkat kompetensi profesional, (2) hal-hal yang tidak jelas dari isi tes yang terdapat didalamnya, serta (3) hambatan-hambatan teknis yang mungkin ada. Masing-masing dari mereka yang terpilih diminta untuk mengerjakan tes, kemudian dimintakan komentar dan pendapatnya melalui angket yang dikembangkan untuk itu. Uji coba pada kelompok kecil bertujuan untuk mendapatkan informasi tentang kelayakan, adaptabilitas, dan keberfungsian penerapan tes kompetensi professional menurut kondisi lapangan. Lebih khusus uji coba pada tahap ini ditujukan untuk mendapatkan data uji validitas, reliabilitas dan kesetaraan dari alat ukur.

HASIL PENELITIAN

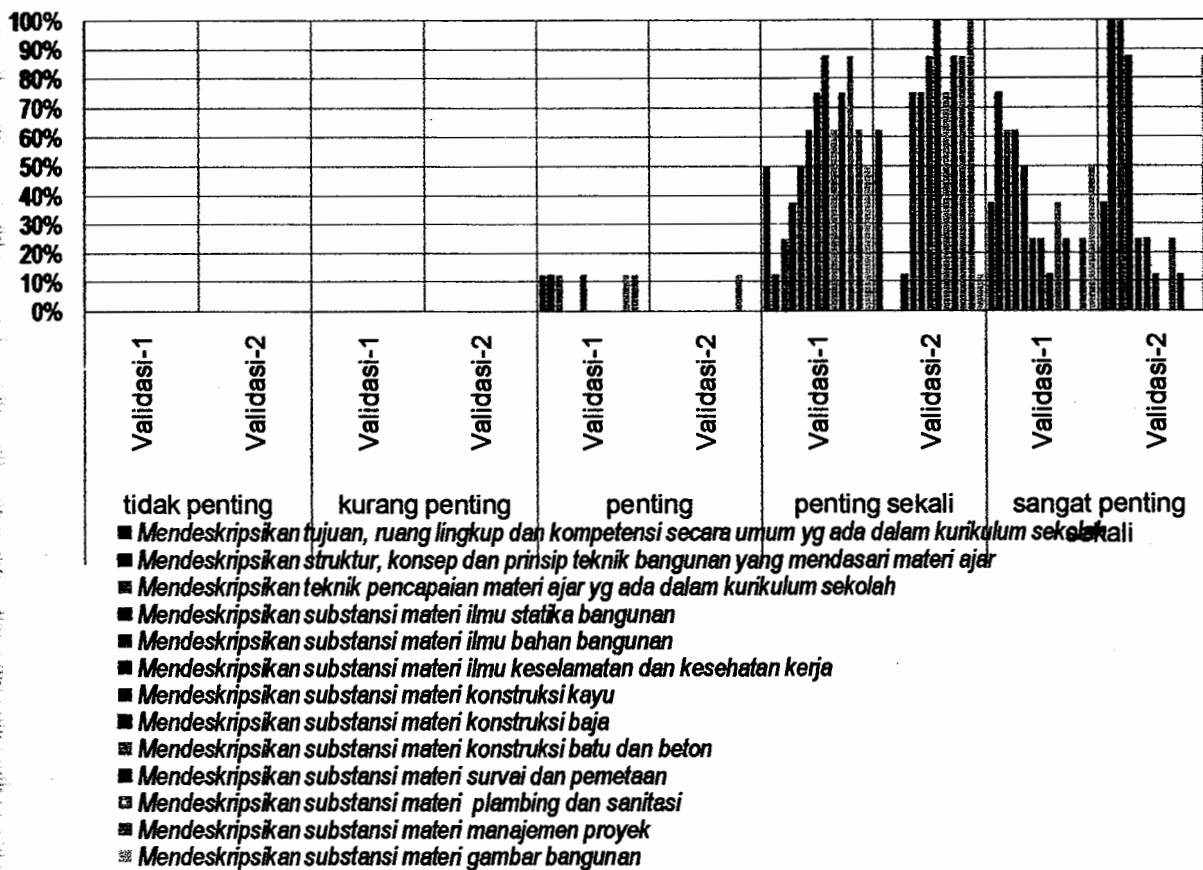
Pada tahap awal, Instrumen pengumpul data dirakit menjadi konstruksi tes (kisi-kisi tes) dengan mempertimbangkan standar kompetensi minimal yang harus dimiliki oleh seorang guru, sehingga atas dasar konsep pokok di atas dirumuskanlah 3 standar kompetensi profesional bidang teknik bangunan (ketekniksipilan). Tiap standar kompetensi kemudian dikembangkan menjadi kompetensi-kompetensi dasar yang lebih

spesifik. Rumusan kompetensi spesifik dibuat berdasarkan kata kerja operasional, yaitu kata kerja yang dapat terukur (teramati) dilengkapi dengan deskripsinya.

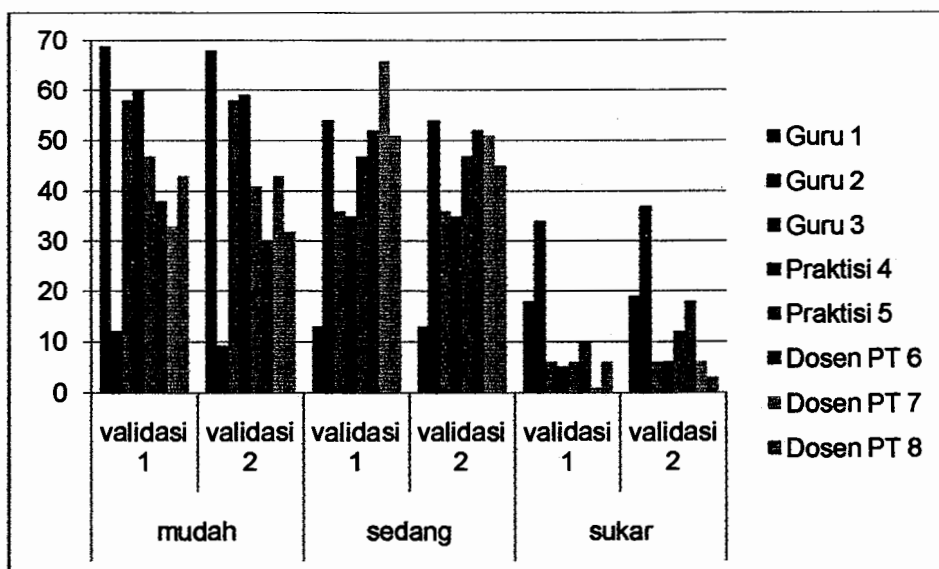
Prosentase kesepakatan pada metode Angoff berkisar antara 3-7 (tiga sampai tujuh) orang guru atau 37,5-87,5% pada putaran pertama, dan 4-8 (empat sampai delapan) orang guru atau 50-100% pada putaran kedua (Gambar 1). Identifikasi kesepakatan ahli tersebut digunakan untuk mendapatkan kata sepakat konten indikator pada saat pembuatan deskriptor (tiap butir soal). Hasil uji kesepakatan materi/bidang studi (digunakan alpha Cronbach SPSS versi 15.0) pada putaran pertama adalah $\alpha=0,953$ dan putaran kedua $\alpha=0,987$. Hal ini membuktikan bahwa reliabilitas (Koefisien alpha Cronbach) atau konsistensi penilaian antar *rater* adalah sangat tinggi dan meningkat pada putaran kedua.

Kesepakatan pada metode *Bookmark* ditandai dengan bergesernya butir-butir soal prosentase kesepakatan *rater* antara 69% (putaran ke 1) dan 92% (putaran ke 2) pada butir mudah yaitu butir nomor 2, 72, dan 94; antara 92% (putaran ke 1) dan 89% (putaran ke 2) pada butir sedang yaitu butir nomor 1, 32, dan 72; dan antara (putaran ke 1) 100% dan 93% (putaran ke 2) pada butir sukar yaitu butir nomor 95. Hasil uji kesepakatan materi/bidang studi (digunakan alpha Cronbach SPSS versi 15.0) pada putaran pertama adalah $\alpha=0,621$ dan putaran kedua $\alpha=0,728$. Hal ini membuktikan bahwa reliabilitas (koefisien alpha Cronbach) atau konsistensi penilaian antar *rater* adalah cukup tinggi.

Tahap rencana pembuatan konstruksi instrumen kompetensi profesional terdiri dari 3 indikator, 13 deskriptor dan 100 butir instrumen, namun selanjutnya melalui pertimbangan *judgment* ahli (metode Angoff dan *Bookmark*) dipilih 60 butir dengan proporsi 17,50% sukar, 60% sedang, dan 22,50% mudah. Pemilihan butir-butir kompetensi profesional ini berdasarkan pada prinsip dasar dalam memahami bidang studi ketekniksipilan. Hasil penelitian Angoff dan *Bookmark* ditunjukkan pada Gambar 1 dan Gambar 2.



Gambar 1. Identifikasi Kesepakatan Ahli Metode Angoff



Gambar 2. Identifikasi Kesepakatan Ahli Metode Bookmark

Pada tahap ini konstruksi instrumen telah memenuhi syarat dan dapat memberikan tingkat performan deskriptor (PLDs = *performance-level descriptors*) yang tepat sebagai dampak besar di lokasi *cutoff score* (Perie, 2008).

DISKUSI

Kata sepakat yang ditunjukkan dengan konsistensi internal (Alpha Cronbach) adalah reliabilitas yang didapatkan dari pengujian konsistensi antar item (antar bagian) bahwa konsistensi respon subyek dari alat ukur pada satu kali pengukuran. Koefisien reliabilitas ini dapat langsung diketahui meskipun tes diberikan satu kali saja. Pada metode Angoff tampak pada koefisien Alpha pada putaran pertama meningkat sampai putaran kedua, demikian juga pada metode *Bookmark*. Hal ini menunjukkan bahwa pelatihan sangat diperlukan untuk memahami 1) konsep-konsep yang berbeda minimal kemahiran atau penguasaan, 2) interpretasi yang berbeda mengenai konten diukur dengan ujian, dan (3) kelalaian dan/atau pergeseran dalam penggunaan skala nilai selama proses rating item.

Selain reliabilitas internal, konsistensi antar *judgment rater* menggunakan koefisien korelasi antar kelas (*Intraclass Correlation Coefficients*, ICC) untuk mengestimasi reliabilitas antar *rater* jika ada beberapa orang *rater* menilai individu baik melalui instrumen *rating* yang menghasilkan data ordinal. Reliabilitas antar *rater* dihitung dengan menggunakan koefisien korelasi antar kelas atau ICC. ICC menunjukkan perbandingan antara variasi yang diakibatkan atribut yang diukur dengan variasi pengukuran secara keseluruhan. Semakin banyak kemiripan hasil penilaian antara satu *rater* dan *rater* lainnya maka koefisien reliabilitas yang didapat akan semakin tinggi. Tabel terlampir memperlihatkan kesepakatan penilaian antar *rater* sangat tinggi pada putaran pertama dan meningkat pada putaran kedua.

Pada metode Angoff menunjukkan bahwa reliabilitas semakin tinggi artinya para ahli yang bertindak sebagai *rater* menunjukkan kata sepakat yang sangat tinggi ($\alpha=0,953 \sim \alpha=0,987$), demikian pula pada metode *Bookmark* kata sepakat yang tinggi ($\alpha=0,621 \sim \alpha=0,728$).

PENUTUP

Kemampuan *judgment* ahli dalam mempertimbangkan masalah konsep standar minimal yang terkait dengan masalah kompeten atau tidak berkompotennya seseorang memerlukan pemikiran yang serius. Hal ini sangat berpengaruh kepada kebijakan, potensi dan nasib seseorang serta dana yang tidak sedikit.

Penelitian yang terkait dengan pengaturan standar seharusnya divalidasi dengan membandingkan hasilnya dengan data empiris.

PUSTAKA

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff C.W., Smith R.W., Impara J.C., & Plake B.S. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement*, 39 (3) p.253–263.
- Baker G.J., Bunch Michael B. & Koons Heather. 2004. Setting Performance Standards: Contemporary Methods. *Educational Measurement, Issues and Practice*. Washington: *Winter 2004*. Vol.23, No.4; pp. 31-50.
- Becker, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.
- Brown, M.T. (2002a). Practice-Based standards Setting. *The Bar Examiner*, August 2002.
- Chantolis, A., & Sireci S.G. (2006). The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement: Issues and Practice*. Vol.25 (1). Spring 2006, 4-12.
- Che Marianne. 2008. A Guide to Understanding and Developing Performance-Level Descriptors, *National Center for the Improvement of Educational Assessment*.
- Cramer, M.D. 2006. A Conceptual framework for a psychometric theory for standards setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*. Vol.25 (1). 2006, 4-12.
- Cramer Kathryn L.. *Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods*.
[http://education.ualberta.ca/educ/psych/cramer/files/RickerSCCE2003 .pdf](http://education.ualberta.ca/educ/psych/cramer/files/RickerSCCE2003.pdf)
diambil tanggal 27 November 2005.

AMPIRAN:

Alpha Cronbach
.953

Tabel 1. Hasil Analisis Kesepakatan Angoff Validasi 1

Intraclass Correlation Coefficient

	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.719 ^b	.535	.881	21.434	12.0	84	.000
Average Measures	.953 ^c	.902	.983	21.434	12.0	84	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

- Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
- The estimator is the same, whether the interaction effect is present or not.
- This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Tabel 2. Hasil Analisis Kesepakatan Angoff Validasi 2

Intraclass Correlation Coefficient

	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.902 ^b	.812	.963	74.680	12.0	84	.000
Average Measures	.987 ^c	.972	.995	74.680	12.0	84	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

- Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
- The estimator is the same, whether the interaction effect is present or not.
- This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Tabel 3. Hasil Analisis Kesepakatan Bookmark Validasi 1

Intraclass Correlation Coefficient

	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.170 ^b	.110	.247	2.637	99.0	693	.000
Average Measures	.621 ^c	.497	.724	2.637	99.0	693	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

- Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
- The estimator is the same, whether the interaction effect is present or not.
- This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA

SERTIFIKAT

No.: 05/HEPI/01/2010

Diberikan kepada:

DRA. NANIK ESTIDARSANI, M.PD

Atas partisipasinya sebagai PESERTA/PEMAKALAH pada Kongres dan
Seminar Nasional HEPI yang Diselenggarakan pada Tanggal 29 – 30 Januari 2010
di Jakarta

Jakarta, 30 Januari 2010

Ketua,



Prof. Djemari Mardapi, Ph.D