Ujian Nasional (UN) yang Akurat dan Adil untuk Sekolah-sekolah di Indonesia yang Sangat Bervariasi Kondisi dan Kualitasnya

Oleh Hari Setiadi[*)]

Abstrak: *Salah satu fungsi Ujian Nasional (UN) adalah sebagai alat ukur yang akurat untuk mengetahui kualitas atau mutu pendidikan di Indonesia secara menyeluruh di dalam wilayah NKRI. Bagaimana membuat Ujian Nasional (UN) tersebut hasil pengukurannya akurat dan adil untuk sekolah-sekolah di Indonesia yang sangat bervariasi kondisi dan kualitasnya? Pertama agar hasil pengukuran secara nasional akurat adalah data yang dikumpulkan diambil secara menyeluruh (sensus, bukan hanya sekedar survei). Selain itu, alat ukur atau ujian yang digunakan juga harus baik kualitasnya. Untuk mendapatkan alat ukur yang kualitasnya baik, soal-soalnya seharusnya diambil dari Bank Soal, karena semua soal-soal yang ada di dalam Bank Soal sudah jelas kualitas dan karakteristik. Kunci ke dua, karena kondisi sekolah-sekolah dan kemampuan siswa di Indonesia sangat beragam, maka tingkat kesukaran soal-soal dalam UN seharusnya berbeda, sesuai dengan kemampuan peserta tesnya. Dengan menggunakan soal-soal yang ada di dalam Bank Soal yang sudah dikalibrasi dengan konsep Item Response Theory (Calibrated Item Bank), hal ini memungkinkan peserta UN mengerjakan perangkat tes yang berbeda sesuai dengan kemampuannya, tetapi tetap hasilnya dapat dibandingkan.*

Kata Kunci: *Ujian Nasional, National Education Quality Control, Item Response Theory (IRT), Kalibrasi, Calibrated Item Bank.*

[*)] *Hari Setiadi adalah pegawai Balitbang Depdiknas yang diperbantukan di Kantor Kementerian Koordinator Bidang Kesra, Ketua Program Pascasarjana PEP UHAMKA, dan mengajar di Program Pasca Sarjana Psikologi, Kekhususan Psikometri, Universitas Indonesia (UI).*

1. Pendahuluan

Tidak bisa disangkal bahwa kondisi dan kualitas sekolah-sekolah yang berada di Indonesia saat ini sangat beragam. Sekolah-sekolah yang lokasinya di pedalaman, misalnya di pedesaan di Irian Jaya, umumnya juga lain kondisi dan kualitasnya dengan sekolah-sekolah yang berada di Jakarta. Perbedaan ini juga bisa sangat tajam seperti dalam hal kualitas guru-gurunya, sarana/prasarana belajarnya seperti laboratorium, perpustakaan, alat peraga, dan sarana/prasarana lainnya, juga Status Sosial Ekonomi (SSE) orang tua siswanya.

Orang tua siswa yang kehidupan sosial ekonominya rendah umumnya tidak memberikan fasilitas dan dukungan sepenuhnya baik moral maupun material kepada anaknya dalam bersekolah. Pada daerah-daerah tertentu di pedesaan, anak-anak mereka umumnya masih diminta untuk membantu orang tuanya mencari nafkah, dengan kata lain anaknya tidak didukung seratus persen hanya untuk bersekolah saja. Hal ini tentunya akan berpengaruh pada konsentrasi dan prestasi belajar anaknya di sekolah.

Sebaliknya orang tua siswa yang hidup di kota, umumnya mereka sangat mendukung belajar anak-anaknya, bahkan tidak sedikit orang tua di kota yang meminta anaknya untuk secara teratur menambah jam belajarnya di luar jam belajar di sekolah atau belajar tambahan *(private lesson* atau bimbingan belajar). Tentunya hal ini juga akan sangat berpengaruh terhadap prestasi belajar anaknya di sekolah.

Sedangkan di lain pihak kita harus memperhatikan bahwa kita hidup dalam Negara Kesatuan Republik Indonesia (NKRI) yang menginginkan semua warga negaranya mendapatkan layanan pendidikan yang sama. Seperti yang tertera di dalam Pembukaan Undang-Undang Dasar 1945, secara tegas dinyatakan bahwa salah satu tujuan negara Republik Indonesia adalah mencerdaskan kehidupan bangsa. Selanjutnya pada Bab XIII Pasal 31 ayat (1) dinyatakan bahwa "tiap-tiap warga berhak mendapatkan pengajaran" dan pada ayat (2) dinyatakan bahwa "Pemerintah mengusahakan dan menyelenggarakan satu sistem pengajaran nasional, yang diatur dengan Undang-undang".

Jadi jelas bahwa amanat Undang-Undang negara kita mengatakan bahwa setiap warga negara Indonesia berhak menda-

patkan pelayananan pendidikan yang layak. Tetapi pelayanan pendidikan di negara kita saat ini masih sangat beragam seperti yang dijelaskan di muka. Karena itu pemerintah saat ini, baik pemerintah pusat maupun pemerintah daerah, sedang giat-giatnya berusaha untuk membantu melengkapi sarana/ prasarana belajar bagi sekolah-sekolah yang membutuhkan dengan berbagai macam program bantuan. Selain sarana dan prasarana belajar, diperhatikan pula kuantitas dan kualitas guru-gurunya yang mengajar di sekolah-sekolah, terutama sekolah-sekolah yang berada di pedalaman/pedesaan atau di desa terpencil.

Tetapi tentunya program pemerintah untuk memperbaiki kondisi sekolah-sekolah di Indonesia tersebut tidak bisa selesai dalam waktu singkat. Sementara itu yang lebih penting lagi adalah bahwa dengan kondisi yang beragam tersebut, bagaimana kualitas sekolah atau mutu pendidikannya secara nasional di Indonesia? Bagaimana bisa diketahui mutu pendidikan atau kualitas lulusan sekolah-sekolah di Indonesia tersebut? Diharapkan mutu pendidikan sekolah-sekolah di wilayah Indonesia dari Sabang sampai Merauke relatif tidak terlalu berbeda. Artinya untuk membandingkan mutu pendidikan atau kualitas lulusan sekolah-sekolah yang ada di dalam wilayah NKRI diperlukan satu alat ukur, yang sekarang kita namakan Ujian Nasional (UN).

Masalahnya kemudian adalah bagaimana membuat alat ukur atau Ujian Nasional (UN) yang dapat mengukur kualitas sekolah-sekolah di wilayah Negara Kesatuan Republik Indonesia (NKRI) yang kondisi dan keadaan sekolahnya sangat beragam? Apakah bisa digunakan suatu alat ukur atau tes yang sama (dalam hal ini tingkat kesukaran soal-soalnya sama), sedangkan peserta tesnya bervariasi tingkat kemampuannya? Misalnya saja, apakah bisa digunakan alat ukur (tes) yang sama untuk mengukur murid-murid yang bersekolah di dalam kota Jakarta dengan murid-murid yang bersekolah di desa terpencil di Irian Jaya? Bagaimana agar hasil pengukuran tersebut adil dan akurat, sehingga dapat digunakan untuk pengambilan kebijakan peningkatan mutu pendidikan secara nasional?

2. Tinjauan Kepustakaan

Dalam teori pengukuran *(measurement)*, apabila kita ingin mengukur suatu objek (kelompok orang)

tertentu, seharusnya kita sudah mempunyai target orang atau kelompok orang yang akan kita ukur kemampuannya (Wright dan Stone, 1979).

Setelah ditentukan target orang yang akan diukur kemampuannya, kemudian instrumen pengukuran atau tes dibuat spesifik berdasarkan target tersebut. Agar hal ini dapat dibuat sistematik, maka kita harus jelas mendefinisikan apa yang kita maksud dengan target pengukuran tersebut, seperti misalnya dimana seharusnya kita memperkirakan kemampuan rata-rata dari kelompok orang yang akan diukur, berapa kemampuan terendah dan kemampuan tertinggi, dan bagaimana kemampuan kelompok di antara dua kemampuan ekstrim (terendah dan tertinggi)?

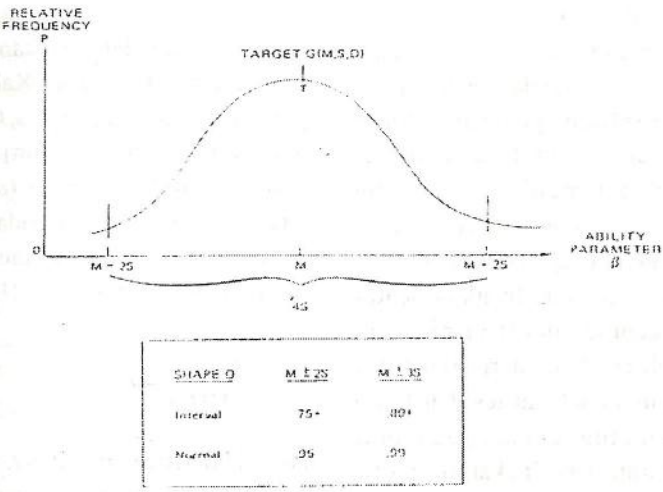
2.1 Target Pengukuran

Kadang-kadang secara eksplisit kita mempunyai pengetahuan pendahuluan tentang target kemampuan yang akan diukur (Setiadi dan Hayat, 1997). Misal kita mempunyai informasi pendahuluan (*prior information*) tentang kemampuan pengikut tes dari hasil sebelumnya sehingga kita dapat memperkirakan daerah atau lokasi rata-rata kemampuan serta rentang kemampuan berdasarkan pengalaman sebelumnya.

Dengan menggunakan soal-soal yang telah dikalibrasi dengan konsep *Item Response Theory (IRT),* kita dapat menentukan bahwa soal-soal tertentu tepat diberikan untuk kelompok target peserta tes, soal-soal lain terlalu sulit, dan soal-soal lainnya terlalu mudah untuk kelompok target (Setiadi dan Hayat, 1997). Untuk menyatakan spesifikasi target pengukuran, digunakan tiga (3) variabel, yaitu: (1) titik tengah target (*target center*); (2) dispersi target (*target dispersion*); dan (3) distribusi target (*target distribution*) (Wright dan Stone1979).

Jadi bisa disimpulkan bahwa pengetahuan pendahuluan kita tentang target kelompok yang akan diukur dengan tiga target parameter, yaitu: M, S, dan D (seperti terlihat pada Gambar 1).

Bisa dinyatakan bahwa target G dengan pernyataan G (M,S,D), dimana M adalah perkiraan terbaik kita tentang titik tengah lokasi/lokasi target (*center*), S adalah perkiraan terbaik kita tentang dispersi atau sebaran target (*dispersion*), dan D adalah perkiraan terbaik kita tentang distribusi skor target.



SHAPE D	M ± 2S	M ± 3S
Interval	75+	89+
Normal	95	99

Gambar 1: Kurva Target Tes

Memperkirakan target lokasi M bisa dinyatakan secara tepat berdasarkan pengalaman data yang lalu. Tetapi memperkirakan target dispersi S dan distribusi skor D lebih sulit karena memaksa kita untuk berpikir tentang perbedaan antara target perorangan dengan kelompoknya.

Jika kita memperkirakan target distribusinya adalah normal, maka dapat diharapkan 95% dari hasil pengukuran kemungkinan terletak dalam interval M±2S, dan hampir keseluruhan (99%) pengukuran terletak dalam interval M±3S. Dan apabila kita tidak yakin bahwa target kita adalah normal, maka kemungkinan lain adalah interval. Apabila interval, berarti bahwa 75% dari hasil pengukuran kemungkinan terletak dalam interval M±2S dan hampir keseluruhan (89%) pengukuran terletak dalam interval M±3S.

Setelah kita menetapkan karakteristik target, kita harus menetapkan seberapa akurat kita akan menolerir kesalahan hasil pengukuran. Dalam hal target satu orang, kita menginginkan *Standard Error of Measurement* (SEM) lebih kecil dari sebaran target atau *target dispersion* (S).

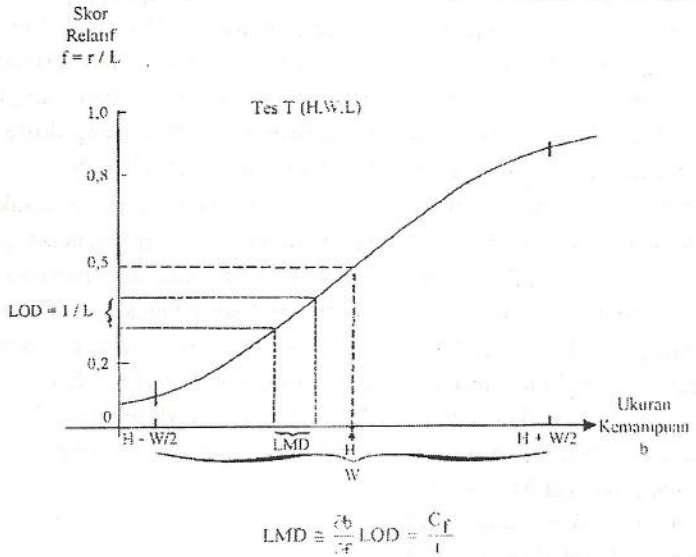
2.2 Karakteristik Tes

Untuk mendapatkan hasil pengukuran yang akurat dan reliabel, maka langkah selanjutnya, apabila target pengukuran sudah ditentukan, instrumen pengukuran atau tes dibuat spesifik berdasarkan target pengukuran yang telah ditentukan. Gambar 2 memperlihatkan bentuk kurva suatu tes, dimana terlihat dua (2) buah ciri khas dari sebuah tes, yaitu Tinggi sebuah tes dan Lebar atau rentang kemampuan atau penyebaran dari tingkat kesukaran soal.

Dua ciri di atas, Tinggi dan Lebar, masih belum lengkap untuk menjadi ciri-ciri sebuah tes. Kalau kita perhatikan Gambar 2 di atas, terlihat ada skor yang menunjukkan pertambahan kemampuan (*ability*) terkecil yang kita ukur pada suatu tes tertentu yang dinamakan *Least Measurable Difference* (LMD), dimana C_f

$$LMD = \frac{C_f}{L}$$

adalah koefisien error dan L adalah panjang tes atau banyaknya soal dalam sebuah tes. *Standard Error of Measurement* (SEM) juga tergantung pada banyaknya soal



$$LMD \cong \frac{\partial b}{\partial f} LOD = \frac{C_f}{1}$$

Gambar 2: Kurva Operasi Sebuah Tes

dalam sebuah tes ($SEM=LMD^{1/2}$), sehingga panjang suatu test (L) atau banyaknya soal dalam sebuah tes juga termasuk dalam ciri-ciri sebuah tes.

Kita bisa menggambarkan suatu desain tes dengan pernyataan T (H,W,L), dimana:

H= Tinggi tes atau rata-rata tingkat kesukaran soal yang diambil dari Bank Soal yang sudah dikalibrasi dengan konsep IRT

W=Lebar tes, atau rentang (*range*) dari tingkat kesukaran soal tersebut

L= Panjang tes, atau banyaknya soal dalam tes

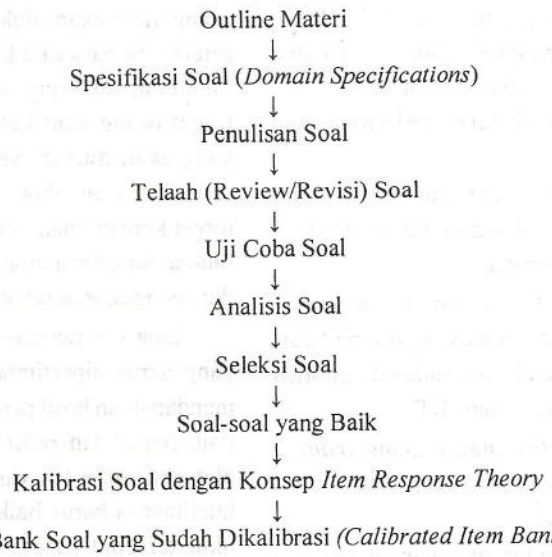
2.3 Desain Tes yang Baik (dengan Konsep *Item Response Theory*)

Untuk mendesain tes atau ujian yang baik adalah mengusahakan agar karakteristik tes T(H,W,L) yang kita buat dapat memenuhi karakteristik target pengukuran G(M,S,D), sehingga *Standard Error of Measurement* atau kesalahan pengukurannya dapat sekecil mungkin di daerah variabel yang akan diukur.

Jadi apabila kita akan mendesain tes, seharusnya pertama-tama kita berusaha memperkirakan kemampuan target orang atau kelompok orang yang akan diukur. Kemudian setelah itu baru kita berusaha untuk membuat tes yang sesuai dengan target orang atau kelompok orang yang akan diukur. Semakin karakteristik tes tersebut sesuai dengan target kemampuan orang yang akan diukur, semakin akurat nantinya hasil dari pengukuran tersebut.

Dua (2) permasalahan pokok yang harus dipertimbangkan untuk mendapatkan hasil pengukuran yang baik (*valid* dan *reliable*) yaitu: (1) alat ukur atau tes yang digunakan kualitasnya harus baik, dan (2) alat ukur tersebut harus sesuai dengan target objek (kemampuan peserta tes) yang akan diukur.

Untuk mendapatkan alat ukur atau tes yang baik, soal-soalnya seharusnya diambil dari Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory (IRT)*, karena soal-soal yang ada di dalam Bank Soal tersebut sudah diseleksi melalui beberapa tahap (Setiadi, 1996), yaitu:



Terlihat bahwa apabila soal-soal yang digunakan dalam suatu tes diambil atau berasal dari Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory* (IRT), maka kualitas soal-soalnya sudah teruji dengan dukungan data empirik. Artinya soal-soal tersebut sudah melalui tahap uji coba soal dan setelah dianalisis baru diseleksi untuk mendapatkan soal-soal yang baik.

2.4 Kelemahan Bank Soal yang Belum Dikalibrasi Dengan IRT

Paling tidak ada dua (2) kelemahan, dalam mendesain tes atau ujian, apabila soal-soalnya diambil dari Bank Soal yang belum dikalibrasi dengan konsep *Item Response Theory (IRT)* atau masih menggunakan Teori Tes Klasik.

Pertama, soal-soal yang belum dikalibrasi dengan konsep IRT (masih menggunakan teori tes klasik), tingkat kesukaran soal-soalnya sangat tergantung pada kelompok peserta tes yang diuji. Artinya, apabila suatu soal diberikan kepada kelompok peserta tes yang pintar, maka soal itu kelihatannya mudah. Kemudian apabila soal yang sama tersebut diberikan kepada kelompok peserta tes yang kurang pintar, maka sebaliknya soal itu menjadi kelihatannya sukar.

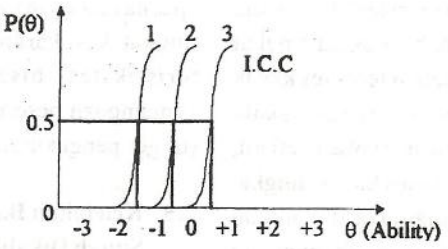
Kelemahan teori tes klasik yang ke dua adalah bahwa skala tingkat kesukaran soal dalam teori tes klasik berbeda definisinya dengan skala domain kemampuan siswa (Setiadi, 1998). Dalam teori tes klasik, tingkat kesukaran soal biasanya dilaporkan dalam skala [0,1] dan dihubungkan dengan populasi pengikut tes atau proporsi banyaknya pengikut tes yang menjawab soal tertentu dengan benar. Sedangkan domain kemampuan siswa, walaupun juga dilaporkan dalam skala [0,1], dihubungkan dengan populasi soal dalam tes. Jadi di dalam teori tes klasik, definisi populasi tingkat kesukaran soal adalah kumpulan orang-orang peserta tes, misalnya berapa persen orang atau peserta tes yang dapat menjawab suatu soal dengan benar. Sedangkan definisi populasi domain kemampuan siswa adalah kumpulan materi (topik atau sub-topik yang dites), misalnya berapa persen topik-topik dalam tes yang dapat dijawab oleh peserta tes.

Artinya secara konsep di dalam Teori Tes Klasik, selain tingkat kesukaran soalnya sangat tergantung pada kelompok peserta yang dites (*sample dependent*), juga tidak ada hubungan antara tingkat kesukaran soal dengan kemampuan peserta tes, sehingga tidak bisa atau sangat sukar apabila kita ingin membuat tes dimana tingkat kesukaran soalnya (karakteristik tes) bisa sesuai dengan kemampuan peserta yang akan dites (target pengukuran).

2.5 Kelebihan Bank Soal yang Sudah Dikalibrasi dengan Konsep IRT

Konsep *Item Response Theory* (IRT) sangat berguna untuk memecahkan masalah-masalah dalam penyeleksian soal-soal untuk mendesain suatu perangkat tes tertentu. Salah satu keunggulan utama *Item Response Theory* (IRT) dibandingkan dengan Teori Tes Klasik adalah bahwa dalam konsep *Item Response Theory* (IRT), statistik soal, khususnya tingkat kesukaran soal, terletak dalam skala yang sama dengan kemampuan siswa yang diukur (Hambleton *et al.*, 1991). Sebagai ilustrasi dapat dilihat pada Grafik 3.

Pada Grafik 3 terlihat bahwa ada tiga buah *Item Characteristic Curve* (ICC) untuk satu-parameter IRT Model. Jadi dalam konsep *Item Response Theory* (IRT) setiap soal diwakili oleh satu *Item Characteristic Curve* (ICC). *Item Characteristic Curve* adalah pernyataan matematika yang berhubungan dengan probabilitas keberhasilan peserta tes (dalam menjawab suatu soal dengan



Grafik 3: Tiga Buah *Item Characteristic* Curve (ICC) *Rasch Model*

benar) dengan kemampuannya (*ability*). Sedangkan ke tiga *Item Characteristic Curve* (ICC) yang terlihat pada Grafik 3 di atas adalah ICC untuk satu-parameter IRT model (*Rasch Model*) dan dapat dinyatakan dengan persamaan sebagai berikut.

$P_i(\theta)=e^{(\theta-bi)}/1+e^{(\theta-bi)}i=1, 2, ...n$

dimana:

$P_i(\theta)$ adalah peluang (*probability*) seorang pengikut tes dengan kemampuan (θ) menjawab soal i dengan benar (dalam Grafik 3 di atas terletak pada garis vertikal)

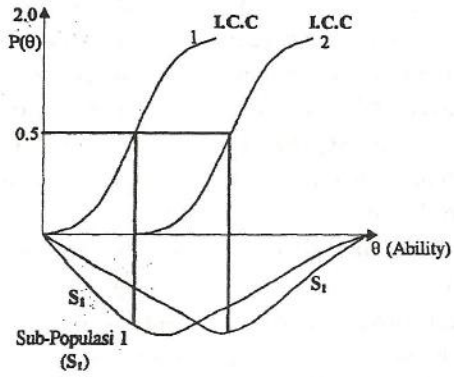
b_i adalah parameter tingkat kesukaran soal

n adalah jumlah soal dalam tes

e adalah bilangan exponential

Berdasarkan persamaan di atas, b_i parameter atau tingkat kesukaran soal adalah titik pada skala kemampuan (garis horizontal dalam Grafik 3) dimana peluang (*probability*) menjawab benar soal tersebut adalah 0,5. Jadi parameter ini juga menunjukan posisi ICC yang berhubungan dengan skala kemampuan (θ). Semakin besar nilai b_i parameter, semakin besar kemampuan yang dibutuhkan oleh peserta tes untuk menjawab peluang 50% dengan benar, karena itu semakin sukar juga soal tersebut. Semakin sukar soal tersebut, letak atau posisi ICCnya semakin ke kanan, dan semakin mudah soal tersebut, letak atau posisi ICCnya semakin ke kiri pada skala kemampuan (θ) dalam Grafik 3. Jadi terlihat bahwa skala tingkat kesukaran soal dalam konsep *Item Response Theory* (IRT) terletak dalam skala yang sama dengan skala kemampuan peserta tes.

Selain itu yang penting juga pada konteks ini adalah bahwa tingkat kesukaran soal dalam konsep *Item Response Theory* (IRT) tidak berubah atau tidak tergantung pada sampel peserta tes yang mengerjakan tes, dan hal ini dapat diilustrasikan dalam Grafik 4.



Grafik 4: Dua ICC dan Distribusi Frekuensi untuk Sub Populasi 1 dan Sub Populasi 2

Pada Grafik 4 di atas terlihat bahwa ada dua (2) buah *Item Characteristic Curve* (ICC), yaitu untuk soal nomor 1 dan soal nomor 2. Terlihat bahwa soal nomor 2 lebih sukar dari pada soal nomor 1, karena ICC soal nomor 2 letaknya lebih ke kanan dari pada ICC nomor 1 dalam skala kemampuan. Terlihat pula ada dua (2) buah distribusi frekuensi skor, yaitu distribusi frekuensi skor untuk Sub Populasi 1 dan distribusi frekuensi skor untuk Sub Populasi 2. Dari Grafik 4 terlihat bahwa banyak orang yang mendapat skor rendah pada Sub Populasi 1, dan sebaliknya banyak orang yang mendapat skor tinggi pada Sub Populasi 2.

Jadi terlihat bahwa tingkat kesukaran soal nomor 1 dan tingkat kesukaran soal nomor 2 nilainya tetap tidak berubah apakah akan dikerjakan oleh Sub Populasi 1 ataupun dikerjakan oleh Sub-Populasi 2. Yang berbeda adalah bahwa apabila soal nomor 1 yang tingkat kesukaran soalnya lebih mudah dari pada soal nomor 2 diberikan kepada Sub Populasi 1, maka jumlah peserta ujian yang menjawab soal nomor 1 dengan benar pada Sub Populasi 1 akan lebih banyak daripada apabila diberikan kepada Sub Populasi 2, karena memang banyak orang di Sub-Populasi 1 yang kemampuannya rendah yang sesuai untuk diberikan soal yang tingkat kesukaran soalnya mudah. Demikian pula halnya untuk soal nomor 2 yang mempunyai tingkat kesukaran soalnya lebih sukar,

apabila diberikan kepada Sub Populasi 1, maka jumlah peserta ujian yang menjawab soal nomor 2 tersebut dengan benar akan lebih sedikit dari pada apabila diberikan kepada Sub Populasi 2, karena memang banyak orang di Sub Populasi 2 yang kemampuannya lebih tinggi dari pada di Sub Populasi 1.

Jadi dengan Grafik 3 dan Grafik 4 di atas terlihat bahwa dalam konsep *Item Response Theory* (IRT) skala tingkat kesukaran soal terletak pada skala yang sama dengan skala kemampuan siswa, dan juga tingkat kesukaran soalnya tidak tergantung pada sampel orang yang mengerjakan soal tersebut. Dengan kelebihan tersebut di atas, maka soal-soal yang ada di dalam Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory* (IRT) dapat digunakan untuk mendesain perangkat tes atau menyeleksi soal-soal dimana tingkat kesukaran soal-soalnya dapat disesuaikan dengan kemampuan (*ability*) siswa yang hendak diukur atau dites.

3. Pembahasan

3.1 Bagaimana mengukur kemampuan siswa di sekolah-sekolah yang sangat berbeda kondisi dan kualitasnya?

Apabila kita ingin mengukur suatu objek, secara logika tidak tepat menggunakan alat ukur yang sama, misalnya saja menggunakan alat ukur timbangan yang sama apabila objek atau benda yang hendak diukur beratnya juga sangat berlainan. Apabila kita ingin mengukur berat buah semangka, tidak bisa kita menggunakan timbangan yang sama dengan apabila kita mengukur berat anting emas atau cincin emas.

Demikian pula dengan ujian atau tes sebagai alat ukur untuk mengukur kemampuan peserta tes. Apabila kita ingin mengukur kemampuan siswa di sekolah-sekolah yang berbeda sekali kondisi dan kualitas kemampuan siswa-siswanya, tentunya juga tidak tepat apabila digunakan ujian atau tes yang sama tingkat kesukaran soalnya. Selain hasilnya nanti tidak akan akurat, juga tidak adil sebagai alat ukur atau tes.

Seperti yang dijelaskan pada teori pengukuran di atas, bahwa memang seharusnya apabila kita ingin mengukur dua objek yang sangat berbeda, maka alat ukur yang

digunakan juga harus berbeda. Jadi memang tepat pula apabila kita ingin mengukur peserta tes yang sangat berbeda kemampuannya, maka ujian atau tes yang digunakan juga seharusnya berbeda tingkat kesukaran soalnya Bisa jadi kemampuan rata-rata peserta tes yang berada di pedalaman Irian Jaya sangat berbeda dengan rata-rata umumnya kemampuannya peserta tes yang berada di Jakarta. Dua tes atau ujian tersebut dibuat tergantung dari perkiraan terbaik target pengukurannya. Target pengukuran kemampuan murid-murid di dalam kota Jakarta akan lain dengan target pengukuran kemampuan murid-murid di pedalaman Irian Jaya.

Jadi supaya adil dan hasilnya akurat, maka ke dua kelompok tersebut di atas (kelompok murid yang berada di pedalaman Irian Jaya dengan kelompok murid yang berada di Jakarta) seharusnya menggunakan dua tes (misalnya Ujian Nasional) yang berbeda tingkat kesukaran soalnya.

Permasalahan pokoknya kemudian adalah apabila digunakan ujian atau tes yang tidak sama atau berbeda tingkat kesukaran soalnya, karena disesuaikan dengan tingkat kemampuan peserta tesnya, apakah bisa dibandingkan hasilnya?

3.2 Bagaimana membandingkan hasil ujian yang berasal dari tes yang berbeda tingkat kesukarannya?

Misalnya kelompok A mengerjakan ujian atau tes yang tingkat kesukaran soalnya lebih sulit dari kelompok B. Apakah mungkin dua kelompok A dan B tersebut dibandingkan hasilnya?

Calibrated Item Bank artinya soal-soal di dalam Bank Soal tersebut sudah dikalibrasi. Melaksanakan kalibrasi soal sama dengan kita mentera suatu alat ukur. Misalnya apabila kita ingin mentera alat ukur timbangan, artinya kita menyamakan skalanya atau menyamakan titik nolnya dan juga menyamakan satuan-satuan unitnya.

Demikian pula apabila kita mengkalibrasi soal di dalam Bank Soal, artinya semua soal di dalam Bank Soal tingkat kesukaran soal-soalnya disamakan skalanya, atau disamakan titik nolnya dan satuan-satuan unitnya. Apabila soal-soalnya di dalam Bank Soal sudah disamakan skalanya dengan konsep *Item Response Theory* (*Calibrated Item Bank*), maka semua perangkat tes yang berbeda dimana soal-soalnya diambil atau berasal dari Bank Soal tersebut hasilnya dapat diban-

dingkan, apabila dikerjakan oleh peserta tes yang berbeda.

Dengan menggunakan konsep *Item Response Theory* (IRT) seperti yang telah diterangkan di bagian terdahulu, maka masalah tersebut di atas dapat dipecahkan. Apabila semua soal-soal dalam ujian atau tes tersebut sudah diambil atau berasal dari Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory* (IRT) *(Calibrated Item Bank)*, maka walaupun kelompok A mengerjakan ujian atau tes yang berbeda tingkat kesukaran soalnya dengan kelompok B, hasil ujian ke dua kelompok tersebut tetap dapat dibandingkan hasilnya.

4. Simpulan

Dari uraian di atas dapat disimpulkan bahwa:

- Untuk mengukur suatu objek (kemampuan peserta tes tertentu) harus digunakan alat ukur atau tes yang sesuai dengan objek (kemampuan peserta tes) yang hendak diukur tersebut.
- Untuk mengukur kemampuan peserta tes yang sangat beragam kemampuannya di Indonesia, seperti Ujian Nasional (UN), seharusnya digunakan juga ujian atau tes yang berbeda tingkat kesukaran soalnya supaya adil dan juga akurat hasilnya.
- Peserta tes atau ujian (seperti Ujian Nasional) yang mengerjakan tes atau ujian yang berbeda tingkat kesukaran soalnya tetap bisa dibandingkan kemampuannya, asalkan soal-soal dalam ujiannya tersebut berasal atau diambil dari Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory* (IRT).
- Jadi Ujian Nasional (UN) untuk peserta tes di pedalaman Irian Jaya misalnya, tetap dapat dibandingkan hasilnya dengan peserta tes di Jakarta walaupun peserta tes yang tinggal di Jakarta menempuh ujian yang lebih sukar tingkat kesukaran soalnya dibandingkan dengan peserta tes yang tinggal di pedalaman Irian Jaya.
- Apabila soal-soalnya diambil dari Bank Soal yang belum dikalibrasi dengan konsep *Item Response Theory* (IRT), maka dua kelompok yang menempuh dua tes yang berbeda tingkat kesukarannya tidak dapat dibandingkan kemampuannya.
- Ujian Nasional (UN) juga menjadi adil dan hasil pengukurannya menjadi lebih akurat, karena

peserta tes menempuh ujian dimana tingkat kesukaran soal-soal di dalam tes sesuai dengan tingkat kemampuan peserta ujian.

5. Saran

Dalam rangka memperoleh sistem ujian berskala nasional yang memiliki keadilan dengan tetap memperhitungkan tingkat validitas dan reliabilitas, disarankan agar pangambil kebijakan mengambil langkah-langkah berikut.

- Semua soal di dalam Ujian Nasional (UN) harus diambil dari Bank Soal yang sudah dikalibrasi dengan konsep *Item Response Theory* (*Calibrated Item Bank*). Alasannya adalah karena selain soal-soalnya sudah melalui beberapa tahap seleksi dan uji coba soal, dengan *Calibrated Item Bank* Ujian Nasional (UN) dapat dibuat adil dan hasilnyapun dapat akurat, karena tingkat kesukaran soalnya dapat disesuaikan dengan kemampuan peserta tes yang sangat beragam.
- Untuk ke depan, perlu dipersiapkan secara lebih serius lagi Sumber Daya Manusia (SDM) di daerah-daerah di Indonesia untuk mengembangkan Bank Soal yang dikalibrasi dengan konsep *Item Response Theory* (IRT), sehingga soal-soalnya nanti dapat dikembangkan di masing-masing daerah.
- Soal-soal yang berasal dari Bank Soal yang dikalibrasi di daerah-daerah bisa lebih disesuaikan tingkat kesukaran soal-soalnya dengan kemampuan peserta tes di daerah dan *content* (isi)nya juga dapat disesuaikan dengan kebutuhan dan kondisi masing-masing daerah.
- Agar tetap dapat dibandingkan hasilnya secara nasional, terutama untuk keperluan *National Education Quality Control* di dalam Negara Kesatuan Republik Indonesia (NKRI), walaupun peserta tes di pusat menempuh tes yang berbeda dengan peserta tes di daerah-daerah, maka Bank Soal yang ada di daerah-daerah harus dihubungkan (*link*) dengan Bank Soal yang ada di pusat dengan menggunakan konsep *Item Response Theory* (IRT).

Pustaka Acuan

Hambleton, R.K., Swaminathan, H., dan Rogers, H.J. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Setiadi, H. 1996. *Item Banks to Improve Assessment Practice*. Laboratory of Psychometric and Evaluation Research Report, University of Massachusetts, USA.

Setiadi, H. 1998. Bank Soal yang Dikalibrasi dengan Konsep IRT Memecahkan Permasalahan pada Ujian-ujian Sistematik yang Diadakan pada Periode-periode Tertentu. *Kajian*, Balitbang Depdiknas, 013 (4), 4-19.

Setiadi, H. dan Hayat, B. 1997. *Mendesain Tes dengan Menggunakan Rasch Model*. Paper dipresentasikan pada Pelatihan Tingkat Nasional untuk Guru Sekolah Lanjutan Tingkat Pertama di Indonesia.

Wright, R.D. dan Stone, M.H. 1979. *Best Test Design*. Chicago: MESA Press.