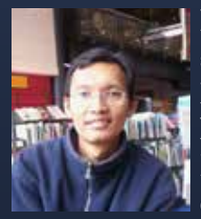
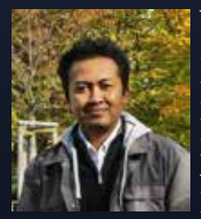
APLIKASI PEMODELAN RASCH
PADA ASSESSMENT PENDIDIKAN

Peningkatan kualitas pendidikan dapat dilakukan dengan berbagai cara. Salah satu cara yang bisa banyak membantunya adalah melakukan penilaian dan evaluasi pendidikan yang sifatnya komprehensif. Untuk itu pemodelan rasch sangat efektif digunakan. Hal ini karena pemodelan racsh mengubah data skor mentah menjadi data dengan interval yang sama sehingga menghasilkan skala pengukuran yang linier, presisi dan mempunyai satuan. Pemodelan rasch dapat digunakan untuk analisis kualitas soal, mengetahui tingkat abilitas siswa dan tingkat kesulitan soal, sampai kepada deteksi miskonsepsi, adanya bias dalam soal, ataupun kemungkinan mengetahui adanya siswa-siswa yang mencontek. Hal ini menunjukkan bahwa rasch model bisa membantu guru (ataupun dosen) untuk mengetahui kualitas pembelajaran yang dia lakukan, dimana perbaikan perlu dilakukan dan seperti apa karakteristik soal dan siswa/mahasiswa yang diajarnya. Sehingga upaya peningkatan kualitas pendidikan dapat difasilitasi dengan lebih mudah, ilmiah dan tepat.



Bambang Sumintono, Ph.D. saat ini merupakan dosen kebijakan pendidikan di Institute of Educational Leadership, Universiti Malaya, di Kuala Lumpur, Malaysia. Dia menyelesaikan S1 di Universitas Terbuka dalam bidang pendidikan kimia. Studi S2 dalam bidang administrasi pendidikan diselesaikan di Flinders University, Adelaide, Australia; dan mendapat gelar doktor (S3) dalam bidang kebijakan pendidikan di Victoria University of Wellington, Wellington, New Zealand. Dia dapat di kontak melalui email: deceng@gmail.com



Wahyu Widhiarso, MA. adalah dosen Fakultas Psikologi UGM di Bagian Pendidikan & Psikometri. Mengampu mata kuliah konstruksi tes, penyusunan skala psikologi dan psikometri. Minat penelitiannya di bidang pengukuran dengan pendekatan Rasch, Teori Respons Butir (IRT) dan Pemodelan Persamaan Struktural (SEM). Sedang melanjutkan studi S3 di Department of Methodology & Evaluation Research, Institute of Psychology, Friedrich-Schiller-University Jena di Jerman.
email: wahyu_psy@ugm.ac.id

Trim Komunikata Publishing House
Jalan Cihanjuang No. 155 Cimahi 40153
Telepon 022-86617757
trim_komunikata@yahoo.com
www.trimkomunikata.com

ISBN 13: 978-602- - -

APLIKASI PEMODELAN RASCH PADA ASSESSMENT PENDIDIKAN
Bambang Sumintono
Wahyu Widhiarso

APLIKASI PEMODELAN RASCH
PADA ASSESSMENT PENDIDIKAN

Bambang Sumintono & Wahyu Widhiarso

APLIKASI PEMODELAN RASCH PADA ASSESSMENT PENDIDIKAN

Bambang Sumintono & Wahyu Widhiarso

Aplikasi Pemodelan Rasch: pada Assessment Pendidikan

Penyunting : Bambang Trim
Penata Letak : Den Binikna
Perancang Kover : Den Binikna

Cetakan I, September 2015
ISBN 978-

Penerbit Trim Komunikata
Anggota Ikapi No. 248/JBA/2013
Jalan Cihanjuang No. 155
Cimahi 40513
Telp. 022-86617757
tri3muvi@gmail.com

KATA PENGANTAR

Bismillah. Setelah terbitnya buku *Aplikasi Rasch Model* pada September 2013 lalu, banyak aktivitas yang menyertainya. Permintaan untuk mengadakan pelatihan/ *workshop* pemodelan Rasch berturut-turut datang dari berbagai universitas negeri dan swasta yang sebagian besar di Pulau Jawa. Ini tentu hal yang menggembirakan bagi kami berdua. Salah satu perkembangannya adalah kemunculan edisi revisi buku tersebut pada November 2014 untuk mengakomodasi permintaan banyak peserta pelatihan.

Saat yang sama pada beberapa *workshop*, guru dan dosen menanyakan aplikasi pemodelan Rasch dalam penilaian dan evaluasi pendidikan (*educational assessment and evaluation*) secara khusus. Untuk hal ini memang perlu buku baru ditulis. Oleh karena itu, buku ini memang dirancang khusus untuk membantu guru, dosen, atau mahasiswa jurusan kependidikan untuk mengaplikasikan Rasch Model dalam analisis ujian yang biasa mereka lakukan. Pendekatan ini penting karena kemampuan analisis yang dihasilkan dengan Rasch Model bisa memberikan informasi yang lebih kaya dibandingkan teori tes klasik (pendekatan skor) yang biasa dilakukan.

Terima kasih kami ucapkan atas kebaikan hati Prof. Dr. Mark Wilson dari University of California, Berkeley, yang telah memberikan izin penggunaan konsep *four building blocks*-nya di Bab 2, buku ini. Melalui email, beliau mendukung upaya penyebarluasan konsepnya melalui buku ini.

Naskah buku dirancang dan ditulis dengan memanfaatkan teknologi informasi dan komunikasi di dua tempat berbeda (Kuala Lumpur dan Yogyakarta), sesuai dengan domisili kami berdua. Besar harapan kami bahwa pemodelan Rasch akan banyak diterapkan di sekolah dan perguruan tinggi kita sehingga banyak membantu meningkatkan kualitas pembelajaran dan penilaiannya pada siswa dan mahasiswa.

Juli 2015,
Bambang Sumintono & Wahyu Widhiarso

DAFTAR ISI

BAB 1
PENILAIAN PENDIDIKAN DAN UJIAN

1.1 Pengantar

Jika kita akan melakukan perjalanan darat dengan mobil dari Bandung menuju Jakarta, tentu kita dengan mudah mengetahui ketika sudah tiba di tempat tujuan. Misalnya, dengan melihat rambu di jalan yang menunjukkan kita sudah berada di Jakarta, melihat secara langsung tugu Monas, ataupun dengan mengecek aplikasi GPS yang ada di telepon genggam. Sepanjang perjalanan dari Bandung ke Jakarta tersebut, kita akan mendapat berbagai informasi apakah jalan yang kita tempuh memang di jalur yang tepat. Contohnya jika kita lewat jalan tol Purbaleunyi, ketika melewati Purwakarta, kita di arah yang tepat; akan berbeda jika ternyata kita menyadari sedang berada di jalur tol yang justru ke arah timur kota Bandung sehingga kita harus berbalik arah.

Pendidikan pada dasarnya tidak jauh berbeda dengan perjalanan. Terdapat berbagai tujuan yang harus dicapai, tentu juga terdapat berbagai jalur yang bisa ditempuh mencapai tujuan tersebut. Apabila kita mau menuju Jakarta dari Bandung, bisa dilakukan melalui tol Purbaleunyi kemudian lanjut ke tol Cikampek; ataupun bisa melewati kota Cianjur terus ke jalur Puncak, menuju Bogor dan akhirnya masuk tol Jagorawi.

Di sekolah, pada umumnya guru yang menentukan tujuan dan jalur yang harus ditempuh siswanya berdasarkan kurikulum yang ditetapkan. Guru bisa memilih buku teks mana yang dijadikan rujukan, media dan alat belajar apa yang digunakan, dan bagaimana mengajarkan pokok bahasan tertentu pada siswa. Tentu yang paling penting adalah, baik guru maupun siswa perlu mengetahui dengan jelas apakah perjalanan belajar mereka mencapai tujuannya atau tidak. Untuk tahu secara pasti maka penilaian kemajuan belajar suatu hal yang esensial untuk dilakukan.

Penilaian (*assessment*) pendidikan adalah proses yang tidak terpisahkan dari pendidikan itu sendiri. Proses belajar mengajar di sekolah selalu melibatkan penilaian pendidikan sebagai hal yang sangat penting dilakukan. Bab ini akan menjelaskan ruang lingkup penilaian mulai tujuannya, aspek yang perlu dinilai, ujian (tes) sebagai bagian dari penilaian serta mengetahui bagaimana analisis ujian dapat dilakukan secara lebih tepat.

1.2 Penilaian (Assessment) Pendidikan

Penilaian dalam pendidikan sebagai satu disiplin ilmu relatif adalah sesuatu yang baru; dalam lingkup pendidikan formal baru pada abad ke-19 lah hal ini dipraktekkan di sekolah-sekolah. Sebelumnya, ujian yang merupakan bagian penilaian, baru dikenal menurut catatan sejarah adalah melalui seleksi untuk mendapatkan pegawai yang berkualitas, dan ini mulai terjadi di Tiongkok pada tahun 260 SM saat Dinasti Han berkuasa. Untuk mendapatkan birokrat kerajaan yang berkualitas tersebut mereka melakukan semacam tes standar bagi calon pegawainya. Sistem ujian ini kemudian diketahui oleh misionaris Eropa pada abad ke-16, dan menggunakannya untuk seleksi pegawai di Eropa, baik untuk pegawai pemerintah maupun perusahaan swasta seperti *English East India Company*.

Ujian pada bidang pendidikan yang kita kenal sekarang, baru mulai dilakukan di Eropa pada 1800-an. Saat itu adalah Revolusi Industri ketika tuntutan perbaikan program pendidikan dan sosial banyak disuarakan, dan ujian menjadi instrumen untuk mengetahui ketercapaian perbaikan tersebut. Menurut catatan sejarah juga, perkembangan tentang ujian ini baru terjadi pada tahun 1845 di Boston, Amerika Serikat, ketika ujian sekolah yang biasa dilakukan secara lisan (oral) kemudian diganti dengan ujian tertulis. Hal itu dilakukan untuk bisa memfasilitasi pencapaian prestasi dalam perbandingan antar-sekolah.

Jenis Penilaian dalam Pendidikan

Definisi penilaian pendidikan sangat beragam, namun biasanya hal itu menyebutkan bahwa penilaian adalah cara untuk menempatkan pembelajar dalam konteks yang dapat menyatakan apa yang dia ketahui dan mampu dia lakukan, di samping juga

menjelaskan apa yang belum dia tahu dan belum mampu dia lakukan. Definisi penilaian pendidikan seperti ini memang sangat luas yang mengindikasikan bahwa untuk mengetahui kemajuan belajar seseorang bisa dilakukan baik secara formal maupun informal, kapan saja, dan dalam waktu jangka waktu yang tidak harus dibatasi.

Dalam aktivitas kegiatan belajar mengajar di sekolah, yang lebih dikenal secara luas dalam penilaian pendidikan disebut sebagai **penilaian formatif** dan **penilaian sumatif**. Penilaian formatif adalah kegiatan penilaian oleh guru terhadap siswa yang dalam hal ini tujuannya lebih pada memberikan informasi yang bermanfaat sehingga pembelajaran berikutnya kualitasnya lebih baik lagi. Hal itu berimplikasi bahwa pada penilaian formatif guru mengumpulkan informasi dan melakukan interpretasi dari bukti hasil belajar yang ada, tentang apa yang perlu diketahui lebih lanjut oleh siswa, serta melakukan adaptasi pengajarannya sesuai dengan kebutuhan siswa. Dalam bahasa yang populer ini juga disebut sebagai *assessment for learning*.

Penilaian sumatif adalah penilaian yang dilakukan untuk mengetahui apa yang sudah diketahui pelajar atau yang bisa dia lakukan, pada periode akhir masa belajar yang ditetapkan. Tujuannya memang untuk memberikan informasi apa prestasi yang telah dicapai; dalam istilah populernya disebut *assessment of learning*. Pada jenis penilaian ini, siswa selalu berada dalam situasi ketika mereka harus menampilkan segala yang telah dikuasai selama waktu tertentu yang menunjukkan prestasi belajarnya, misalnya dalam Ujian Akhir Nasional (UAN).

Namun, tujuan penilaian pendidikan tidak hanya memusatkan pada penilaian formatif dan sumatif. Penilaian pendidikan meliputi berbagai aspek yang melingkupi aspek di dalam dan luar sekolah yang juga menjadi penting sebagai bukti akuntabilitas kegiatan pengembangan sumber daya manusia. Secara lebih lengkapnya, penilaian pendidikan paling tidak meliputi lima tujuan (Musial et. al, 2009) sebagai berikut.

a. Memberikan umpan balik (*feedback*). Satu tujuan utama dari penilaian adalah memberikan umpan balik kepada pembelajar/siswa. Beberapa ahli menganggap tujuan ini adalah hal utama dalam penilaian karena fokusnya memenuhi pada kebutuhan dan harapan siswa yaitu siswa memahami informasi dan menggunakannya untuk langkah belajar berikutnya. Informasi yang diberikan tidak sekadar skor atau persentase jawaban yang benar. Namun penilai harus menginterpretasikannya untuk memahami kekuatan dan keterbatasan siswa. Ini tidak lain ciri dari penilaian formatif yang dalam hal ini bentuknya bisa beragam, seperti penyelenggaraan kuis, ujian, pengamatan, catatan kegiatan, diskusi dengan siswa dan lainnya, yang bertujuan mengumpulkan informasi dan memberikan umpan balik.

b. Menentukan apa yang dipelajari selanjutnya. Penilaian ini masih bersifat formatif, namun lebih terfokus. Hal ini biasa digunakan untuk materi bahasan

yang sifat pemahamannya bertahap. Misalnya, pada pelajaran matematika, jika seorang siswa dapat menyebutkan urutan bilangan asli (1, 2, 3…), itu tanda dia bisa diajarkan tentang pertambahan; hal yang sama apabila siswa mampu mengurutkan bilangan secara terbalik (10, 9, 8….), dia bisa dilatih melakukan operasi pengurangan. Demikian juga jika siswa sudah mampu memahami operasi pertambahan, mengajarkan perkalian adalah tahapan berikutnya.

c. **Diagnosis kesulitan belajar dan miskonsepsi**. Penilaian dalam bentuk formatif juga bisa dilakukan guru untuk mendeteksi adanya kesulitan belajar siswa, dan mendeteksi adanya miskonsepsi terhadap materi belajar. Miskonsepsi dapat menyebabkan kesalahpahaman tentang suatu konsep. Misalnya dalam fisika, siswa menganggap berat jenis suatu benda bergantung pada besarnya sehingga benda yang besar dianggap berat dan akan tenggelam dalam air. Akan tetapi, hal itu tidak berlaku untuk gunung es karena es seberapa pun besarnya, berat jenisnya lebih kecil dari air. Diagnosis terhadap miskonsepsi akan sangat berguna bagi guru untuk bisa membantu siswa secara lebih tepat, yang dilakukan dengan cara mengajarkan kembali hal yang sama dan melakukan koreksi terhadap miskonsepsi yang terjadi.

d. **Menentukan kemajuan belajar dan mengetahui perkembangannya**. Sistem pendidikan formal seperti sekolah didesain berdasarkan kemampuan atau usia secara bertingkat dengan berbagai persyaratan prestasi tertentu yang harus dicapai siswa di setiap tingkatannya. Penilaian pendidikan digunakan dalam hal ini untuk mengetahui seberapa jauh kemajuan belajar siswa dibandingkan siswa lain di peringkat yang sama pada daerah atau negara lain (misalnya seperti ujian TIMSS dan PISA). Ini contoh penilaian sumatif, sekaligus bersifat normatif. Penilaian tidak dimaksudkan untuk memberikan informasi spesifik tentang apa yang siswa tahu, namun memberikan gambaran dan info perbandingan prestasi di tahapan pendidikan tertentu antarsiswa dalam satu negara misalnya. Penilaian sumatif seperti ini bisa dilakukan secara diagnostik dengan sampel yang besar maupun dalam bentuk kegiatan nasional seperti ujian akhir nasional.

e. **Sebagai alat evaluasi dan akuntabilitas program**. Kegunaan penilaian lainnya adalah untuk memperbaiki kualitas pengajaran di sekolah dibanding dengan mengetahui kebutuhan belajar individu siswa. Setiap sistem pendidikan memerlukan informasi seberapa bagus kegiatan pembelajaran yang sudah berlangsung, misalnya pada negara kita dalam bentuk Ujian Akhir Nasional. Hal ini tidak lain di samping untuk mengetahui kualitas pengajaran secara nasional, informasi perbandingan prestasi antara daerah, juga untuk memenuhi aspek akuntabilitas karena telah menggunakan dana negara yang tidak sedikit. Penilaian seperti ini jelas adalah bersifat sumatif karena siftanya yang massal dan luas maka jenis penilaian yang dilakukan biasanya adalah tes standar.

Diperlukan bukti-bukti pendukung untuk merealisasikan pencapaian tujuan di atas. Bukti tersebut dikumpulkan dan diinterpretasikan dengan menggunakan prosedur yang ilmiah agar umpan balik yang diberikan kepada siswa juga memiliki nilai ilmiah. Hal ini membawa kita kepada pemahaman pada penerapan prinsip-prinsip penilaian pendidikan yang baik.

Terdapat tiga hal yang harus dipenuhi yaitu (a) mengumpulkan bukti hasil belajar yang berhubungan dengan kegiatan belajar dan mengajar; (b) melakukan interpretasi bukti tersebut secara tepat; dan (c) memahami dimensi-dimensi utama dalam pembelajaran yang berkaitan.

Oleh karena itu, bagaimana siswa mengetahui sesuatu dan bagaimana siswa belajar merupakan hal inti yang secara mendasar perlu kita ukur. Tentunya bukan hanya sekadar menampilkan bukti hasil belajar melainkan lebih mendalam lagi yaitu, materi pelajaran apa yang memang sangat berharga untuk dipelajari bagi generasi muda sehingga kita memang perlu mengajarkannya serta melakukan penilaian untuk mengetahui kemajuannya. Di samping itu pengetahuan apa yang memengaruhi prestasi belajar siswa juga perlu diketahui dengan baik oleh guru (lihat **Kotak 1**).

Kotak 1. Faktor-Faktor yang Memengaruhi Prestasi Belajar di Sekolah Berdasarkan Hasil Riset

a. Faktor-faktor di dalam sekolah
   - Ketatnya kurikulum
   - Pengetahuan dan keterampilan guru
   - Pengalaman dan kehadiran guru
   - Ukuran besarnya kelas
   - Ketersediaan teknologi yang membantu pengajaran
   - Keamanan di sekolah

b. Faktor-faktor di luar sekolah
   - Kurangnya berat badan saat kelahiran
   - Keracunan timbal
   - Kelaparan dan nutrisi yang tidak bergizi
   - Kebiasaan membaca pada usia anak-anak
   - Banyaknya waktu digunakan menonton televisi
   - Keberadaan dan partisipasi orang tua
   - Mobilitas siswa

Para ahli pendidikan sepakat bahwa pengetahuan yang berharga untuk dipelajari melibatkan tiga dimensi utama yaitu (a) aspek kognitif yang menyangkut pemahaman terhadap materi pelajaran; atau yang biasa disebut menguasai produk pengetahuan, (b) aspek psikomotor yang menyangkut keterampilan dalam keahlian tertentu (*skills*), dan (c) aspek afektif yang menyangkut adaptasi nilai, sikap dan norma yang positif.

Meski terdapat tiga dimensi utama pembelajaran yang penting tersebut, faktanya yang selalu terjadi, penilaian pada dimensi kognitif mendominasi penilaian dalam pendidikan, misalnya penilaian terhadap pemahaman materi pelajaran. Hal yang tidak bisa dihindari adalah karena penilaian dimensi ini lebih mudah dilakukan dan disetarakan dibanding yang lainnya. Aspek keterampilan misalnya, penilaian dapat dilakukan melalui pengamatan dan bertanya secara langsung kepada siswa saat dia diberikan tantangan dalam konteks yang membuat dia menunjukkan kemampuan berpikir strategis dan menunjukkan keterampilannya. Pada aspek sikap pun guru bisa menempatkan siswa dalam kondisi yang membuat dia harus bekerja secara disiplin dan jujur dalam melaksanakan satu tugas, yakni guru bisa mengamati dan menilai kerja yang dilakukan siswa. Namun, kedua hal terakhir ini memang tidak semudah melakukan penilaian materi pelajaran yang bisa dilakukan secara serentak dengan bahan serta prosedur yang sama.

1.3 Penilaian Pendidikan Melalui Ujian (Tes)

Ujian atau tes adalah bagian yang tidak terpisahkan dalam penilaian; keduanya seolah tidak bisa dibedakan. Namun, pembedaan kedua hal ini perlu dilakukan untuk kejelasan aktivitas yang dilakukan. Penilaian pendidikan lebih luas cakupannya dibandingkan ujian (tes), yang memang lebih fokus. **Ujian** adalah prosedur evaluasi yang dilakukan oleh seorang guru terhadap pengetahuan dan keterampilan siswa untuk mengetahui kinerjanya dengan menggunakan instrumen tertentu. Adapun yang disebut **instrumen** pun beragam, bisa berbentuk set soal yang harus dikerjakan maupun tugas menghasilkan suatu produk tertentu. Ujian bisa dilakukan dalam berbagai bentuk, dimaksudkan untuk memberikan pengukuran yang objektif dari kegiatan pembelajaran yang telah dilakukan.

Bentuk ujian atau tes yang paling umum dipakai oleh guru dalam menguji siswanya di kelas adalah tes tertulis. Namun, bentuk ujian lain juga bisa dilakukan seperti ujian pada mata pelajaran olahraga, yang menguji keterampilan berenang dengan patokan waktu yang diperlukan dalam jarak tertentu dengan menggunakan gaya bebas misalnya; ataupun keterampilan melakukan kegiatan eksperimen dengan alat dan bahan laboratorium pada pelajaran ilmu pengetahuan alam. Apapun bentuk,

konteks, atau mata pelajarannya, ujian yang dilakukan selalu dimaksudkan untuk melakukan evaluasi terhadap sesuatu yang ingin dinilai.

Hasil ujian yang didapatkan biasanya digunakan dalam berbagai cara. Satu hasil ujian matematika seorang siswa sekolah dasar kelas tiga misalnya, menunjukkan pengetahuan keterampilan berhitung (aritmatika) seperti proses penambahan, pengurangan, pengalian, dan pembagian di bawah jumlah angka seratus. Skor yang didapat oleh siswa dalam ujian matematika tadi bisa menunjukkan seberapa bagus prestasinya dibanding temannya di kelas, ataupun prestasi yang telah dia raih sebelumnya di kelas yang sama. Secara lebih lengkap, hasil ujian ini dapat digunakan oleh guru untuk (a) menentukan abilitas siswa relatif terhadap siswa lain dalam tes yang sama; (b) menunjukkan perkembangan kemampuan siswa dalam suatu jangka waktu dalam pengetahuan, dan keterampilan tertentu; (c) menunjukkan bukti pemahaman akan satu materi pelajaran, pengetahuan atau ide tertentu; dan (d) meramalkan kinerja siswa di masa depan.

Jika kita mendapatkan informasi mengenai hasil ujian dalam volume yang lebih besar, seperti dari semua siswa di satu kelas atau satu sekolah, hal tersebut bisa digunakan untuk melihat pencapaian yang didapat dibandingkan kelas lain atau sekolah lain yang berdekatan, ataupun indikasi prestasi kelas/sekolah tadi di tingkat kabupaten/kota. Untuk hal terakhir ini, maka suatu tes standar perlu dirancang dengan baik sebelum digunakan secara luas.

Suatu bentuk ujian yang bisa menginformasikan prestasi antarsiswa atau antar sekolah bahkan antar daerah memerlukan tes standar yang baik. Bentuk tes standar yang umum dipakai adalah jenis soal/aitem pilihan ganda dengan satu pilihan jawaban yang benar dan siswa disediakan beberapa pilihan jawaban; ataupun soal berbentuk uraian/esai, yakni siswa harus menjawab secara tertulis jawabannya secara benar, dan lengkap. Supaya hasil tes bisa dipercaya dan tepat untuk digunakan, maka aspek validitas dan reliabilitas instrumen adalah hal esensial yang harus dipenuhi.

1.4 Validitas

Suatu tes haruslah valid, artinya tes tersebut mengukur sesuatu yang harus diukur. Walaupun konsep ini terlihat sederhana, para guru biasanya melupakan hal ini. Misalnya, soal-soal ujian yang disusun baru dibuat pada saat akhir pengumpulan soal yang waktunya terbatas. Akibatnya, pokok bahasan yang diberikan secara lengkap dan mendalam di bagian awal pelajaran misalnya tidak diakomodasi atau malah terlewat dijadikan soal pada ujian yang isinya cenderung berisi bagian akhir materi pelajaran saja. Pada kasus lain, jika hasil pelajaran yang diinginkan meliputi perubahan dalam pengetahuan, keterampilan, dan sikap, maka soal-soal yang dibuat pun haruslah mencakup ketiga hal tersebut.

Validitas adalah masalah proses pembuktian yang berkelanjutan, mengacu pada sejauh mana bukti dan teori mendukung interpretasi terhadap skor tes sesuai tujuan tes. Proses validasi melibatkan proses pengumpulkan bukti untuk memberikan dasar ilmiah untuk interpretasi skor tes. Validitas adalah masalah interpretasi terhadap nilai tes, bukan tes itu sendiri, karena validitas tidak seberapa terkait dengan bentuk atau jenis tes, tetapi interpretasi terhadap skor tes. Oleh karena itu, ketika skor tes digunakan atau ditafsirkan lebih dari satu cara, setiap cara interpretasi dimaksudkan harus divalidasi.

Tipe Validitas

Banyak literatur yang membagi validitas menjadi tiga tipe, yaitu validitas isi, validitas kriteria, dan validitas konstruk. Saat ini pembagian berdasarkan tipe validitas mulai jarang dan digantikan dengan tipe validitas berdasarkan tipe pembuktian yang dipakai. Hal ini disebabkan validitas lebih menyangkut pembuktian apakah skor atau keputusan yang kita buat berdasarkan skor tersebut sudah tepat. Pada perkembangan terkini, validitas lebih menyangkut masalah pembuktian. Setidaknya ini yang dipakai oleh asosiasi pendidikan dan psikologi di Amerika Serikat (AERA atau American Educational Research Association dan APA atau American Psychological Association). Berikut ini ada bukti yang dapat dipakai untuk menunjukkan validitas tes yang kita susun berdasarkan standar dari AERA dan APA.

a. **Bukti berdasarkan isi tes.** Bukti ini menyangkut hubungan antara isi tes dan konstruk yang diukur. Isi tes memuat spesifikasi isi (misalnya, kisi-kisi, cetak biru tes) serta properti tes seperti jenis butir (contohnya, jenis soal pilihan ganda), instruksi, tugas yang diberikan, dan penulisan butir. Pembuktian validitas tipe ini dapat dilakukan melalui analisis logis atau empiris terhadap domain isi tes, landasan teori yang dipakai, bahasa dalam penulisan butir, serta properti-properti tes lainnya. Berikut ini beberapa contoh bahan yang dapat menjadi bukti validitas tes berdasarkan isinya:
   - persetujuan pakar atau praktisi terhadap kisi-kisi dan butir-butir tes;
   - persetujuan ahli bahasa yang mengevaluasi daya keterbacaan pernyataan di dalam butir;
   - pendapat siswa yang mengatakan bahwa mereka dapat memahami dengan baik semua pernyataan di dalam butir.

b. **Bukti berdasarkan proses respons.** Bukti ini menyangkut bagaimana siswa merespons butir. Misalnya, kita membuat butir berbentuk soal cerita untuk mengukur penalaran matematika (*mathematical reasoning*), maka jawaban benar yang diberikan oleh siswa pada soal ini haruslah berdasarkan kemampuan penalaran matematika, dan bukan kemampuan dia dalam memahami bacaan. Bukti-bukti mengenai proses respons dapat dikumpulkan melalui wawancara

untuk mendapatkan informasi mengenai proses kognitif apa yang dibutuhkan siswa dalam menjawab soal. Jika proses tersebut relevan dengan kemampuan yang diukur, tes atau butir yang dikaji memiliki bukti kevalidan.

c. **Bukti berdasarkan struktur internal.** Struktur internal tes dapat menunjukkan sejauh mana butir dan komponen tes sesuai dengan konstruk yang diukur. Tes dapat memiliki satu komponen atau beberapa komponen. Antara satu komponen tes dan komponen lainnya dapat berkaitan atau tidak. Jika komponen-komponen tersebut memiliki keterkaitan yang rendah, tes akan cenderung bersifat multidimensi. Jika keterkaitannya tinggi, tes tersebut bersifat unidimensi. Struktur ini dapat dibuktikan melalui analisis faktor yang mengindikasikan struktur faktor tes dan reliabilitas yang mengindikasikan homogenitas butir-butir tes. Beberapa analisis statistik juga menunjukkan struktur internal, misalnya bobot faktor (*factor loading*) dan indeks ketepatan butir-model (misalnya, koefisien *infit* dan *outfit* dalam pemodelan rasch, akan dijelaskan kemudian).
d. **Bukti berdasarkan keterkaitan dengan variabel lain.** Variabel eksternal dapat berupa (a) kriteria yang diharapkan diprediksi oleh skor tes dan (b) konstruk lain yang memiliki kesamaan dengan konstruk yang diukur oleh tes. Termasuk dalam bukti ini adalah bukti konvergensi dan diskriminasi. Hubungan antara skor tes yang kita kembangkan dan skor tes lain yang mengukur konstruk yang sama memberikan bukti konvergensi, sedangkan hubungan antara skor tes yang kita susun dan tes lain yang mengukur konstruk berbeda memberikan bukti diskriminansi.
e. **Bukti berdasarkan konsekuensi tes.** Bukti ini terkait dengan konsekuensi dari pengambilan keputusan yang didasarkan oleh skor tes. Keputusan tersebut dapat berupa penentuan kelulusan atau kenaikan kelas, penempatan pada kelas khusus, atau ada tidaknya kebutuhan khusus. Jika konsekuensi yang dialami siswa itu memberikan manfaat yang positif, tes yang kita kembangkan valid. Namun sebaliknya, jika tes yang kita susun menunjukkan bahwa si A memiliki kebutuhan khusus padahal sebenarnya tidak, tes yang kita kembangkan memiliki validitas yang rendah.

Catatan Mengenai Validitas

a. **Bukti kualitatif juga dapat menunjukkan validitas.** Selama ini banyak yang beranggapan bahwa validitas itu bersifat formal yang hanya menekankan pada bukti kuantitatif yang ditunjukkan dengan koefisien validitas. Dari beberapa bukti yang dipaparkan di muka, validitas banyak bersumber dari penggalian informasi tambahan, baik sebelum maupun setelah tes diberikan. Kebanyakan

informasi ini adalah informasi yang bersifat kualitatif yang dihasilkan melalui proses observasi atau wawancara.

b. **Validitas dapat berada pada level tes dan butir.** Meskipun berisi butir-butir yang valid, belum tentu tes memiliki validitas yang tinggi. Hal ini akan bergantung pada bagaimana hasil tes tersebut diinterpretasikan atau difungsikan. Misalnya, sebuah tes kosakata berisi sepuluh butir yang valid. Tes ini menjadi tidak valid jika hasilnya dipakai untuk menunjukkan seberapa jauh kemampuan penalaran siswa.

c. **Validitas bukan hal yang diskrit.** Validitas merupakan derajat kontinum, bukan sesuatu yang diskrit, antata valid dan tidak valid. Tipe validitas juga tidak bersifat diskrit, misalnya validitas isi dan validitas konstruk bukanlah sesuatu yang diskrit dan terpisah, melainkan saling mendukung satu sama lain.

d. **Daya diskriminasi butir.** Seringkali ditemui peneliti yang salah kaprah dalam memahami koefisien validitas. Misalnya, banyak yang melaporkan daya diskriminasi butir yang dihitung melalui korelasi butir-total sebagai koefisien validitas. Daya diskriminasi memang merupakan salah indikator yang perlu, tetapi belum cukup kuat untuk menunjukkan kevalidan (*necessary but not sufficient*). Namun, perlu dicatat bahwa statistik ini bukan koefisien validitas. Singkatnya seperti sering dikatakan, bahwa "*validity is not about number, it is about argument*".

1.5 Reliabilitas

Selain validitas, suatu tes yang diberikan ke siswa juga harus reliabel atau ajek, yang bermakna pengukuran dengan ujian yang dilakukan mendapatkan hasil yang konsisten. Misalnya, ujian yang diberikan hari ini kepada siswa oleh seorang guru, seharusnya memberikan nilai yang tidak jauh berbeda apabila diberikan esoknya (karena tidak ada aktivitas pembelajaran atau lupa pada jangka waktu yang hanya satu hari). Kecilnya reliabilitas dapat terjadi karena set soal ujian yang tidak baik (butir soal yang membingungkan) ataupun tidak adanya konsistensi dalam pemberian skor. Kedua hal tersebut adalah tanggung jawab guru untuk menghindarinya.

Ada tiga terminologi yang menggambarkan reliabilitas pengukuran, yaitu stabilitas (*stability*), ekuivalensi (*equivalency*), dan konsistensi internal (*internal consistency*). Reliabilitas sebagai koefisien stabilitas menunjukkan hasil yang sama didapatkan dari pengulangan tes, ekuivalensi menunjukkan seberapa jauh dua tes yang paralel akan menghasilkan skor tes yang sama, dan konsistensi internal menunjukkan seberapa konsisten hasil skor tiap butir dalam satu tes. Reliabilitas dapat diestimasi jika ada yang dibandingkan. Perbandingan antar-waktu yang diturunkan menjadi pendekatan

reliabilitas tes ulang, perbandingan antar-bentuk tes yang diturunkan menjadi **reliabilitas tes paralel**, dan perbandingan antar-komponen tes yang diturunkan menjadi pendekatan **konsistensi internal**.

a. **Pendekatan Tes Ulang.** Relabilitas tes ulang didapatkan dari korelasi antara skor dari tes yang sama. Jika tes diberikan kepada siswa dengan populasi yang sama, diharapkan koefisien reliabilitas yang mendekati 1. Pada tipe ini, koefisien reliabilitas didapatkan melalui korelasi skor tes antarwaktu. Ada dua jenis koefisien korelasi yang dipakai, pertama adalah korelasi Pearson (*product moment*) dan korelasi intrakelas (*interclass correlation*/ICC).
b. **Pendekatan Tes Paralel.** Tipe ini disusun untuk mengatasi permasalahan yang ada pada tipe reliabilitas tes paralel berkaitan dengan isu efek bawaan atau kontaminasi. Reliabilitas tes paralel disebut juga dengan reliabilitas form pengganti (*alternate form*). Sama seperti reliabilitas tes ulang, harga reliabilitas didapatkan dari korelasi antara skor dari kedua tes yang paralel.
c. **Pendekatan Konsistensi Internal.** Reliabel dalam pengertian konsistensi internal menunjukkan bahwa antara satu bagian tes dan bagian lainnya menghasilkan pengukuran yang konsisten. Konsistensi intenal diindikasikan oleh tingginya korelasi antara belahan tes. Belahan ini dapat berupa butir maupun komponen tes. Karena itu, dalam pendekatan konsistensi internal dikenal konsistensi dua belahan tes yang biasa dihitung dengan koefisien Spearman-Brown, atau tiga belahan yang biasa dihitung dengan koefisien Feldt, atau yang dihitung dengan menggunakan koefisien alpha.

Catatan Mengenai Reliabilitas

a. **Keterbatasan estimasi reliabilitas teori klasik.** Harga reliabilitas yang diestimasi dengan menggunakan teori klasik memiliki kelemahan masalah ketergantungan pada sampel, skor mentah yang non-linear, adanya pembatasan dalam rentang skor, dan harganya bisa berarah negatif.
b. **Reliabilitas pada IRT/Rasch memiliki cara penafsiran berbeda.** Berbeda dengan reliabilitas dalam teori klasik yang memiliki harga tunggal, reliabilitas dalam konteks IRT/Rasch antara satu tingkat kemampuan dengan kemampuan berbeda-beda. Sebuah tes yang sama akan menghasilkan reliabilitas pengukuran yang berbeda ketika diberikan pada individu dan kemampuan sangat tinggi dan sangat rendah. Harga reliabilitas tunggal yang dilaporkan oleh beberapa *software* IRT/Rasch (misalnya, Ministeps) merupakan rangkuman umum dari reliabilitas per tingkat kemampuan individu yang diukur.
c. **Reliabilitas bukan properti instrumen.** Sebuah tes tidak dapat diestimasi reliabel tidaknya karena reliabitas bukan atribut untuk instrumen akan tetapi untuk skor atau pengukuran. Banyak literatur yang menyarankan menggunakan

istilah "reliabilitas skor yang dihasilkan oleh Instrumen A" atau "reliabilitas pengukuran yang dihasilkan dari sampel X".

d. Paradoks antara reliabilitas dan validitas. Reliabilitas (konsistensi internal) dapat ditingkatkan harganya dengan cara memilih butir yang memiliki inter korelasi dengan butir lainnya yang sangat tinggi. Upaya ini di satu sisi memang dapat meningkatkan reliabilitas, tetapi pada sisi lainnya akan menurunkan validitas. Rendahnya validitas ini ditunjukkan rendahnya validitas isi yang ditandai dengan terbatasnya domain ukur yang dijangkau oleh instrumen. Keterbatasan ini disebabkan oleh instrumen yang berisi butir-butir yang isinya tumpang tindih (*overlap, redundant*).

e. Reliabilitas dalam Pemodelan Rasch. Nilai reliabilitas dalam pemodelan Rasch ditunjukkan dengan nilai separasi individu (*person separation*) dan separasi butir (*item separation*). Separasi individu menunjukkan seberapa baik seperangkat butir di dalam tes menyebar sepanjang rentang atau kontinum kemampuan *logit* (akan dijelaskan kemudian). Semakin besar harga separasi individu, semakin baik tes yang disusun karena butir-butir di dalamnya mampu menjangkau individu dengan kemampuan di tingkat tinggi sampai ke yang rendah. Separasi butir menunjukkan seberapa besar sampel yang dikenakan pengukuran tersebar sepanjang skala interval linier. Semakin tinggi nilai separasi butir, semakin baik pengukuran yang dilakukan. Indeks ini berguna untuk mendefinisikan kebermaknaan konstruk yang kita ukur.

1.6 Analisis Hasil Ujian (Tes)

Prosedur analisis dimulai dari proses mendapatkan informasi mengenai siswa dari hasil ujian yang berupa skor. Terdapat berbagai cara untuk mendapatkan skor yang menunjukkan kemampuan siswa. Cara yang umum dilakukan adalah menjumlahkan skor jawaban yang benar. Skor ini menunjukkan kemampuan siswa. Analisis lebih lanjut adalah dengan melakukan prosedur statistik sederhana untuk bisa menjelaskan lebih jauh tentang kualitas soal, kualitas siswa, maupun perbandingan atribut yang diukur.

Pendekatan yang banyak dipakai saat ini dalam analisis hasil ujian adalah pendekatan teori tes klasik (*classical test theory* atau CTT). Teori tes klasik bisa digunakan untuk melakukan prediksi tentang hasil dari suatu ujian (tes). Prediksi ini dilakukan dengan mempertimbangkan beberapa parameter, seperti kemampuan siswa dan tingkat kesulitan soal. Charles Spearman mengemukakan teori tes klasik ini pada tahun 1904 dan banyak diaplikasikan dalam bidang pendidikan, khususnya penilaian pendidikan. Asumsi dasar yang dipunyai oleh teori tes klasik ini adalah, skor yang

didapat dilambangkan dengan X, tidak lain adalah terdiri dari skor murni (T) dan eror pengukuran (E), sehingga persamaannya:

$$X = T + E$$

Artinya, di dalam skor hasil ujian yang didapat satu siswa misalnya, terkandung skor murni dan eror pengukuran. Hal yang perlu dicatat adalah, skor tampak (X) bersifat nyata (muncul dalam data secara langsung), sedangkan skor murni (T) dan eror pengukuran (E) bersifat tersembunyi (*latent*) atau tidak bisa diamati secara langsung. Keduanya muncul dalam data setelah melalui proses estimasi. Asumsi lain yang perlu diketahui adalah eror pengukuran (E) dalam CTT bersifat acak dan tidak berkorelasi dengan X maupun T, dan korelasi yang diharapkan muncul adalah 0 (nol). Teori tes klasik (CTT) hanya menekankan skor tampak dari satu ujian, yang biasanya disimpulkan sebagai kemampuan (abilitas) seseorang dari ujian yang diikuti.

Dari skor mentah ini maka berbagai analisis dan interpretasi bisa dihasilkan sesuai dengan keperluan yang dilakukan di antaranya sebagai berikut.

a. Statistik deskriptif. Tiga statistik yang dipakai dalam konteks ini adalah mengenai tendensi sentral (misalnya, rata-rata), ukuran keragaman (misalnya, varian), dan tabel frekuensi. Ketiganya akan memberikan informasi secara langsung butir soal mana yang berguna dan mana yang tidak. Misalnya, keragaman skor antarsiswa yang rendah menunjukkan rendahnya kualitas soal-soal dalam tes.

b. Tingkat kesulitan. Tingkat kesulitan menunjukkan proporsi siswa yang dapat mengerjakan soal secara benar dari satu ujian. Tingkat kesulitan mempunyai titik terendah sebesar 1,0, artinya semua siswa dapat menjawab dengan betul soal tes. Angka 1 menunjukkan 100% individu bisa mengatasi suatu soal. Titik tertinggi tingkat kesulitan adalah 0,0, menunjukkan tidak ada satupun (0%) individu yang bisa menjawab dengan benar. Butir soal yang memiliki titik ekstrem seperti kedua contoh di muka tidak banyak berguna karena tidak bisa membedakan kemampuan individu, dengan kata lain soal itu tidak bagus kualitasnya. Karena itu, tingkat kesulitan 0,50 (yaitu 50%) dari anggota kelompok yang diuji lulus, merupakan tingkat kesulitan optimal, yakni soal tersebut mempunyai tingkat pembedaan kemampuan tertinggi untuk peserta tes.

c. Daya diskriminasi. Daya diskriminasi butir menunjukkan seberapa jauh sebuah soal mampu membedakan individu yang memiliki kemampuan yang tinggi dan rendah. Sederhananya, jika siswa berkemampuan tinggi dan rendah dapat mengatasi soal nomor 10, soal ini memiliki daya diskriminasi butir yang rendah. Sebaliknya, jika siswa berkemampuan tinggi dapat mengatasi soal

nomor 10 sedangkan yang berkemampuan rendah tidak dapat mengatasi, butir nomor 10 memiliki daya diskriminasi yang tinggi. Daya diskriminasi dapat dihitung dengan menggunakan indeks D dan korelasi butir total.

d. Pembobotan butir soal. Umumnya, dalam konteks CTT skor untuk tiap butir soal diberikan sama (misal, 1 untuk jawaban betul; 0 untuk jawaban salah), pembobotan skor diberlakukan apabila satu soal yang diberikan mempunyai bobot yang berbeda untuk menghasilkan total skor mentah. Terdapat banyak cara untuk memberikan pembobotan, misal melalui reliabilitas soal, dalam hal ini soal dengan reliabilitas tinggi memiliki bobot lebih besar. Pembobotan juga dapat dilakukan dengan menggunakan nilai korelasi butir-total, regresi, ataupun dengan analisis faktor.

Catatan Mengenai Teori Skor Klasik

Teori skor klasik bukan satu-satunya pendekatan dalam psikometri. Ada beberapa pendekatan lain yang merupakan alternatif dari pendekatan teori klasik. Pada dasarnya penggunaan skor mentah/*raw score* sebagai ukuran prestasi memiliki beberapa kelemahan, di antaranya sebagai berikut.

a. Skor mentah pada dasarnya bukanlah hasil pengukuran. Lebih tepatnya skor mentah adalah jumlah jawaban benar dari soal yang dikerjakan siswa.

b. Skor mentah adalah informasi awal. Skor mentah juga biasanya dinyatakan dalam persentase (%) yang tidak lain hanyalah ringkasan data berupa angka, tetapi tidak memberikan data dari suatu pengukuran.

c. Skor mentah memiliki makna kuantitatif yang lemah. Makna kuantitatif dari skor mentah yang didapat akan berbeda, bergantung pada banyaknya soal, sedangkan persentase jawaban betul selalu bergantung pada tingkat kesulitan soal.

d. Skor mentah tidak menunjukkan kemampuan seseorang terhadap tugas tertentu. Skor mentah juga tidak bisa banyak menjelaskan tingkat kesulitan soalnya.

e. Skor mentah dan persentase jawaban benar tidak selalu bersifat linier. Dalam sebuah tes yang bersifat linier, siswa yang memiliki skor 15 (skala 0 hingga 100) selalu memiliki kemampuan lebih tinggi dibanding yang memiliki skor 10. Namun, secara empirik terkadang keduanya memungkinkan memiliki kemampuan yang sama. Untuk yang terakhir ini, pernyataan dari Choppin (1990, h. 7) misalnya menyebutkan "*in many applications, raw scores on test are unsatisfactory measures (because of their linear nature, and the uncertainity as to the meaning of zero and perfect scores)*".

Untuk bisa menjelaskan lebih jauh, misalnya terdapat dua orang siswa, yaitu Abdul dan Budi, yang sama-sama mempunyai skor mentah 10 (artinya 10 jawaban betul) dari 20 soal pelajaran Bahasa Indonesia yang mereka kerjakan. Dengan nilai skor yang sama tersebut, kita tidak tahu secara pasti mana yang prestasinya lebih baik, apakah Budi atau Abdul? Sebab, memang tidak ada informasi lain yang tersedia. Jika kita mendapat informasi bahwa Budi lebih banyak mengerjakan dengan benar soal-soal yang susah dibandingkan Abdul, dengan info ini kita bisa simpulkan bahwa abilitas Budi lebih baik dibanding Abdul. Dalam konteks teori klasik juga, apabila ada satu siswa bernama Catur dan skornya adalah 0 (tidak ada satupun jawaban betul), bagaimana memaknai hal ini? Apakah ini bermakna bahwa Catur tidak bisa berbahasa Indonesia? Apakah Catur tidak mempunyai abilitas dalam mengerjakan soal tadi? Hal ini tidak lain menunjukkan suatu ketidakpastian makna dari skor yang didapat. Karena itu pembuatan suatu skala dari data mentah yang didapat menjadi hal yang fundamental untuk bisa mendapatkan informasi yang lebih akurat dan dapat diperbandingkan (lihat **Kotak 2**).

Karena itu, pendekatan yang berbeda dengan penggunaan skor mentah sangat diperlukan dalam konteks penilaian pendidikan. Hal ini khususnya untuk bisa memberikan informasi yang sangat lengkap dari abilitas yang dimiliki oleh peserta didik dan pada saat yang sama juga menentukan kualitas soal yang diberikan. Pendekatan yang dimaksud adalah dengan mengaplikasikan pengukuran pemodelan rasch (*rasch model measurement*) pada data mentah hasil ujian, yang tujuan utamanya menghasilkan suatu skala pengukuran dengan interval yang sama yang nantinya bisa memberikan informasi secara akurat tentang peserta tes maupun kualitas soal yang dikerjakan. Skala yang dihasilkan dengan pemodelan rasch kualitasnya menyamai pengukuran pada dimensi fisik dalam fisika, seperti mengukur panjang dengan mistar sentimeter ataupun mengukur berat dengan neraca kilogram, dalam hal ini hasil yang didapat bisa dibandingkan karena mempunyai satuan yang sama, bersifat linier, dan mempunyai interval yang sama.

Kotak 2. Rekomendasi untuk Mengembangkan Penilaian

Pada tahun 2001, The National Research Council membentuk satu badan bernama Committee on the Foundations of Assessment, yang bertugas melakukan pemeriksaan dan sintesis berbagai hasil riset terbaru dalam ilmu kognitif, serta melakukan studi atas implikasi hal ini pada penilaian pendidikan. Beberapa rekomendasinya dalam hal praktek penilaian sebagai berikut.

1. *Pengembang instrumen penilaian/ujian baik di kelas maupun dalam skala yang besar harus memberikan perhatian yang eksplisit terhadap tiga elemen penilaian, yaitu aspek kognitif, ujian, dan interpretasinya, termasuk koordinasi dari ketiganya.* Ketiga elemen tersebut harus didasarkan pada pengetahuan modern tentang bagaimana siswa belajar dan bagaimana belajar sebaiknya diukur. Waktu dan usaha harus dilakukan pada desain berdasarkan teori dan proses validasi sebelum penilaian dilakukan.

2. *Pengembang kurikulum pendidikan dan penilaian kelas harus membuat alat yang membantu guru untuk melaksanakan pembelajaran berkualitas tinggi dan praktek penilaian, yang hal ini konsisten dengan pemahaman modern bagaimana siswa belajar dan bagaimana belajar tersebut diukur.* Penilaian dan bahan pengajaran tambahan harus melakukan interpretasi temuan riset dalam ilmu kognitif yang akan berguna bagi guru-guru. Pengembang harus menggunakan keunggulan teknologi untuk menilai bagaimana siswa belajar secara mendetail, dengan frekuensi yang selayaknya dalam cara yang terintegrasi dengan pembelajaran.

3. *Pengambil kebijakan didesak untuk mempertimbangkan keterbatasan penilaian yang ada saat ini dan mendukung pengembangan suatu sistem baru dengan melibatkan penilaian beragam yang akan meningkatkan kemampuan mereka dalam membuat keputusan tentang program pendidikan dan alokasi sumber daya.* Keputusan penting dari seseorang tidak selayaknya hanya berdasarkan satu hasil tes saja. Pengambil kebijakan seharusnya mengembangkan sistem penilaian yang menggunakan berbagai pengukuran kinerja siswa. Penilaian di kelas maupun dalam skala yang besar selayaknya berkembang berbagi pengetahuan tentang sifat alami belajar itu sendiri.

4. *Kesetimbangan antara kekuasaan dan sumber daya harus dipindahkan dari penekanan pada penilaian oleh lembaga luar kepada penekanan tes formatif di kelas yang dirancang untuk membantu pembelajaran.*

Bagian berikutnya adalah menjelaskan empat komponen utama pengukuran (*four building blocks*) yang menjelaskan tentang konteks ujian sebagai alat pengukuran yang objektif.

BAB 2
EMPAT KOMPONEN UTAMA PENGUKURAN

2.1 Pengantar

Di bagian sebelumnya sudah dijelaskan konteks tentang penilaian pendidikan dan juga dibahas perbedaannya dengan ujian atau tes, serta pengolahan hasilnya yang tidak lain memberikan gambaran tentang kegiatan pengukuran. Pada bagian ini akan dijelaskan analisis hasil ujian dengan pendekatan pemodelan Rasch (*Rasch model*) dengan menggunakan kerangka konseptual yang dibuat oleh Mark Wilson (2005), yaitu konsep empat komponen utama pengukuran (*four building blocks*) dari bukunya yang berjudul *Constructing Measures: an Item Response Modeling Approach*. Konsep dari Mark Wilson ini digunakan karena memberikan penjelasan yang komprehensif