KEPUTUSAN SELEKSI DALAM *HIGH STAKE EXAMS*: WACANA PSIKOMETRIS

UNIVERSITAS GADJAH MADA

Pidato Pengukuhan Jabatan Guru Besar dalam
Bidang Psikometri pada
Fakultas Psikologi
Universitas Gadjah Mada

Oleh:

Prof. Dr. Saifuddin Azwar, MA

Bismillahi rrohmaani rrohiim

Yang saya hormati
Ketua, Sekretaris, dan Anggota Majelis Wali Amanat, Universitas Gadjah Mada
Ketua, Sekretaris, dan Anggota Majelis Guru Besar, Universitas Gadjah Mada
Ketua, Sekretaris, dan Anggota Senat Akademik, Universitas Gadjah Mada
Rektor, Wakil Rektor Senior, dan Wakil Rektor, Universitas Gadjah Mada
Para Dekan dan Wakil Dekan di lingkungan Universitas Gadjah Mada
Dekan dan para Wakil Dekan Fakultas Psikologi Universitas Gadjah Mada
Para tamu hadirin sekalian

Assalamu alaykum warohmatullahi wa barokatuh,

Syukur alhamdulillah, dengan rahmat dan barokah Allah azza wa jall, kita semua dianugerahi kesempatan untuk berkumpul dalam ruangan ini untuk mengikuti rapat terbuka Majelis Guru Besar Universitas Gadjah Mada yang pada hari ini berkenan memberi saya kesempatan guna menyampaikan pidato pengukuhan saya sebagai Guru Besar di bidang Psikometri pada Fakultas Psikologi Universitas Gadjah Mada. Pidato pengukuhan ini berjudul "Keputusan Seleksi dalam *High Stake Exams* : Wacana Psikometris".

Hadirin yang saya hormati,

Di Indonesia, selain penyelenggaraan Ujian Nasional (UN), Ujian Masuk Perguruan Tinggi (UMPT) pun merupakan bentuk penyelenggaraan ujian atau tes yang ditanggapi sangat serius oleh mereka yang terlibat secara langsung atau tidak langsung. Kedua ujian tersebut merupakan bentuk pengukuran kompetensi kognitif yang menghasilkan predikat keberhasilan atau kegagalan, namun dalam UMPT keberhasilan tersebut lebih dikaitkan dengan faktor seleksi.

Uraian berikut akan membahas sisi psikometris penggunaan skor sebagai dasar pengambilan keputusan seleksi dalam UMPT .

Paling tidak terdapat empat alasan mengapa perguruan tinggi harus bersifat selektif dalam penerimaan mahasiswa baru. Alasan pertama adalah bahwa perguruan tinggi merupakan ajang penyiapan calon pemimpin masyarakat di masa yang akan datang, yang karena itu diperlukan semacam 'kepastian' bahwa para mahasiswa di perguruan tinggi sebagai calon pemimpin itu benar-benar memiliki kualitas yang diperlukan yang tidak semua calon mahasiswa memilikinya. Alasan yang ke dua adalah langkanya kesempatan untuk belajar di perguruan tinggi, terutama di negara-negara yang sedang berkembang seperti di Indonesia, sehingga perguruan tinggi menginginkan peluang yang kecil itu diberikan kepada calon yang paling berkualitas bukan asal lulusan SLTA saja. Ketiga, dengan adanya sistem seleksi dimungkinkan terjaringnya *human talent* yang berharga, sehingga penyia-nyiaan potensi manusiawi termaksud dapat dihindari. Keempat adalah kenyataan bahwa pendidikan tinggi merupakan upaya yang sangat mahal yang harus dimanfaatkan secara efisien oleh mereka (calon mahasiswa) yang paling besar kemungkinannya untuk berhasil dalam belajar (Suryabrata, 1989).

Dalam kaitannya dengan keempat hal tersebut di atas, suatu sistem seleksi penerimaan mahasiswa baru yang ideal haruslah diselenggarakan dengan mempertimbangkan sekurang-kurangnya empat aspek, yaitu (a) efektivitas ekonomik, (b) insentif belajar-mengajar, (c) efektivitas prediksi, dan (d) ekuitas (Departemen Pendidikan dan Kebudayaan, 1990). Pertimbangan tersebut langsung membawa kepada permasalahan penyediaan sistem dan alat seleksi, yang secara umum dikenal dengan ujian/tes masuk perguruan tinggi (*college admission tests*). Salah-satu bentuk ujian atau tes masuk perguruan tinggi yang sudah sejak lama digunakan adalah bentuk ujian tertulis yang pada umumnya berupa perangkat tes yang terdiri dari beberapa mata uji.

Pertimbangan efektivitas prediksi menuntut terpenuhinya fungsi ujian masuk sebagai prediktor keberhasilan mahasiswa setelah diterima untuk belajar di perguruan tinggi. Artinya, mereka yang dinyatakan diterima dikarenakan berhasil memperoleh skor tertinggi dalam ujian masuk perguruan tinggi memang ternyata kemudian

memperlihatkan keberhasilan akademik yang memuaskan (yang secara operasional sering dinyatakan dalam bentuk indeks prestasi kumulatif yang tinggi). Landasan pertimbangan ini adalah bahwa calon yang paling besar kemungkinannya untuk berhasil harus diterima, karena mereka inilah *human talent* yang lebih berhak untuk memperoleh kesempatan menikmati pendidikan tinggi yang terbatas itu.

Bersamaan dengan pertimbangan efektivitas prediksi, pertimbangan ekuitas menuntut bukan hanya daya prediksi UMPT yang tajam tetapi juga fungsi tes yang tidak bias dalam memprediksi sehingga mampu menjadi landasan pengambilan keputusan yang tidak merugikan kelompok tertentu dikarenakan faktor-faktor yang tidak relevan dan pengaruhnya terjadi secara sistematis. Ekuitas mengandung makna adanya kesempatan atau peluang yang setara (*equal opportunities*) bagi subjek peserta tes, yang berkaitan dengan dua isu yaitu pertama adalah isu mengenai bias pada tes (*test bias*) yang merupakan karakter tes itu sendiri dan yang ke dua adalah isu mengenai *unfairness* yang merupakan aspek interpretasi dan aplikasi hasil tes. Klitgaard menerjemahkan terminologi ekuitas sebagai *group representativeness* (Klitgaard, 1987 dalam Suryabrata, 2005), suatu istilah yang tidak mudah untuk didefinisikan kecuali bagi kelompok yang jelas dasar pembedaannya seperti jenis kelamin. Bagi sebagian ahli, memang penggunaan istilah *fairness* dan ekuitas dalam kaitannya dengan tes hasil belajar, seringkali tidak dibedakan (Uwakwe, 2004).

Keterbatasan kapasitas tempat atau daya tampung perguruan tinggi telah menjadi alasan praktis timbulnya kebutuhan untuk melakukan seleksi terhadap calon mahasiswa yang kemudian memacu perkembangan prosedur dan sistem seleksi yang semakin canggih, efisien, dan valid, termasuk perkembangan tes tulis yang digunakan sebagai bagian dari prosedur seleksi yang sangat penting.

Di sebagian negara yang lebih maju seperti Amerika, Canada, Cina, Australia, dan beberapa negara di Eropa sistem seleksi calon mahasiswa perguruan tinggi berakar pada sejarah pengukuran yang panjang dan telah mengalami perkembangan pesat sedangkan di sebagian negara lain masih dalam taraf perkembangan untuk mencari bentuk yang tepat. Bagaimanapun, skor ujian tulis telah menduduki posisi utama sebagai pertimbangan dalam keputusan seleksi.

Hadirin yang saya hormati,

Kelayakan keputusan yang diambil berdasarkan interpretasi skor tes sangatlah ditentukan oleh kualitas pengukuran dan ketepatan interpretasinya. Oleh karena itu sangat dapat dimengerti mengapa para pakar pengukuran menuntut terpenuhinya syarat-syarat validitas, reliabilitas, dan objektivitas pada penggunaan tes sebagai alat ukur. Di antara ketiga hal tersebut, validitas merupakan kondisi utama yang harus ada pada setiap pengukuran. Kondisi mutlak yang harus terpenuhi agar deskripsi atribut atau kesimpulan yang diambil merupakan kebenaran adalah bahwa pengukuran harus menghasilkan data yang valid.

Di sisi lain, harus disadari bahwa subjek tes adalah manusia. Karena itu persoalan tes dan pengukuran bukan sekedar masalah keberhasilan mendeskripsikan atribut dalam diri manusia ke dalam bentuk angka dan label interpretasinya. Masalah yang lebih penting adalah akibat yang dapat ditimbulkan oleh hasil tes yang bahkan dapat menjangkau bukan saja subjek pengukuran itu saja melainkan juga orang-orang lain yang ikut berkepentingan (*social consequences of test*).

Mudah dimengerti bahwa berkaitan dengan hasil tes seleksi masuk perguruan tinggi, maka konsekuensi dari kesalahan dalam pengambilan keputusan yang diakibatkan oleh informasi dari skor tes yang tidak akurat akan dapat membawa akibat sosial yang buruk bagi yang bersangkutan, ancaman terhadap harga diri (*self-esteem*), bahkan juga kehilangan masa depan. Inilah yang disebut oleh para ahli sebagai *'high stake' exams* yang hasil ukurnya menjadi landasan pengambilan keputusan yang dapat mengubah kehidupan. Karena itulah, evaluasi terhadap kualitas tes yang digunakan dalam berbagai tes seleksi masuk perguruan tinggi mestinya tidak lagi hanya terbatas pada analisis aitem serta estimasi validitas prediktif saja melainkan sudah harus dipertajam pada aspek penggunaan skor dalam pengambilan keputusan seleksi yang didasari oleh interpretasi skor yang *fair*.

Fairness dalam interpretasi skor tes tidak dapat diharapkan dari suatu tes yang berfungsi bias. Suatu tes disebut bias bila dua kelompok subjek (pria dan wanita, misalnya) yang memiliki tingkat

kemampuan setara cenderung memperoleh skor yang berbeda (Childs, 1990). Adanya bias tes (*test bias*) merupakan persoalan alat ukur atau pengukuran (*measurement problem*) yang mengakibatkan subjek kelompok tertentu memiliki peluang yang lebih besar untuk memperoleh skor tinggi dibandingkan subjek dari kelompok lain yang sebetulnya memiliki kemampuan setara. Bias terjadi karena adanya eror sistematik yang berasal dari karakteristik subjek yang tidak relevan dengan tujuan tes namun ikut mempengaruhi skor. Para ahli mengatakan bahwa bias tes adalah *'a systematic error that disadvantages the test performance of one group compared to another'* (Shepard, 1981) atau '*systematic under-or overestimation of a population parameter by a statistic'* (Jensen, 1980) dan '*constant or systematic error, as opposed to chance or random error, in the estimation of some value'* (Reynolds, 1982 h. 199). Schumacker (2005) mengatakan bahwa bila suatu tes lebih menguntungkan salah-satu kelompok subjek maka tes itu dianggap bias dan melanggar prinsip *fairness*.

Hadirin yang saya hormati,

Pertanyaan besar mengenai *test fairness* di Amerika mulai mengedepan di akhir tahun 60an dan awal tahun 70an, sekalipun tidak memperoleh jawaban yang cukup memuaskan di masa itu (Cole & Zieky, 2001). Perhatian terhadap masalah *fairness* dipicu pula antara lain oleh kontroversi yang hebat akibat artikel Jensen dalam salah-satu nomor terbitan Harvard Educational Review berjudul 'How Much Can We Boost IQ and Scholastic Achievement?' tentang pertimbangan komponen genetik di antara berbagai penyebab perbedaan performans antara kelompok kulit putih dan kulit berwarna di Amerika (Jensen, 1969). Hanson dkk. (1973) kemudian menerbitkan artikel yang berisi rumusan beberapa strategi guna mengatasi masalah bias dalam admisi perguruan tinggi khususnya bias jenis kelamin.

Di penghujung Tahun 1980an mulailah era kebangkitan kesadaran akan kompleksitas dan pentingnya permasalahan tersebut sehingga para peneliti mulai memusatkan perhatian mereka pada wacana *fairness* sebagai aspek dari validitas. Semenjak itu kemudian riset-riset mengenai validitas diferensial (*differential validity*) dan

prediksi diferensial (*differential prediction*) terus dilakukan secara ekstensif terutama dalam masalah yang berkaitan dengan fungsi prediktif ujian masuk perguruan tinggi (Young & Kobrin, 2001).

Di Indonesia, yang sistem seleksi masuk perguruan tingginya masih terus dalam pengembangan, studi tentang validitas UMPT pada umumnya masih terbatas pada permasalahan validitas prediktif dan sama sekali belum menyentuh aspek yang lebih dalam seperti permasalahan validitas diferensial dan bias prediksi. Sudah saatnya untuk mengangkat wacana kesetaraan peluang untuk memasuki perguruan tinggi bagi calon yang memiliki potensi yang sama. Pada lingkup akademis, sudah waktunya bagi peneliti untuk memperluas objek studi mereka sehingga mencakup ranah kajian bias tes.

Pertanyaan mengenai *fairness* ini, terutama dalam konteks jenis kelamin, tentu relevan untuk diangkat ke permukaan. Memang belum terdengar adanya 'protes' dari calon yang tidak diterima yang mempertanyakan apakah tes seleksi yang digunakan dalam seleksi masuk perguruan tinggi sama akuratnya bagi pria dan wanita. Sejauh ini kebanyakan dari mereka yang gagal masuk mengatribusikan kegagalannya pada kekalahan bersaing karena calon lain memang lebih pintar, atau karena program studi tertentu memang lebih cocok bagi jenis kelamin tertentu, atau karena memang belum beruntung. Belum muncul keinginan untuk mengetahui apakah ada faktor lain pada tes yang mungkin secara sistematis merugikan atau bahkan menguntungkan bagi sebagian mereka.

Hadirin yang saya hormati,

Seleksi masuk perguruan tinggi menggunakan, antara lain, bentuk tes tertulis yang dirancang untuk memprediksikan peluang keberhasilan calon mahasiswa bila diberi kesempatan belajar lebih lanjut. Artinya, makna skor tes masuk perguruan tinggi menjadi penting karena dijadikan sebagai acuan yang mendasari pengambilan keputusan penerimaan atau penolakan calon mahasiswa, bukan sekedar merupakan deskripsi kuantitatif mengenai kemampuan mereka. Dalam konteks inilah bias tes berimplikasi terhadap *fair* atau tidaknya penggunaan skor tes tersebut bagi peserta tes yang berasal dari kelompok yang berbeda. Pertanyaan yang lebih jauh daripada

sekedar validitas prediktif, yaitu apakah fungsi prediksi itu tidak merugikan kelompok tertentu dan apakah keputusan yang diambil dalam menerima mahasiswa adalah *fair* bagi semua kelompok, menjadi sangat relevan khususnya dalam konteks perbedaan jenis kelamin yang selama ini belum tersentuh oleh penelitian kualitas ujian masuk perguruan tinggi. Permasalahan validitas diferensial yang berimplikasi pada bias dan *fairness* memang agaknya belum banyak memperoleh perhatian dari kalangan masyarakat awam sebagai subjek yang paling berkepentingan, padahal permasalahan ini di luar negeri sudah merupakan isu yang sangat krusial sebagaimana digambarkan dalam prakata buku 'Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions': *The role of standardized tests in this sorting process has been one of the principal flashpoints in discussions of its fairness. Tests have been cited as the chief evidence of unfairness in lawsuits over admissions decisions, criticized as biased against minorities and women, and blamed for the fierce competitiveness of the process. Yet tests have also been praised for their value in providing a common yardstick for comparing students from diverse schools with different grading standards* (Beatty, et al., 1999).

Hadirin yang saya hormati,

Sebagai instrumen ukur yang akan menghasilkan data untuk dijadikan landasan pengambilan keputusan penerimaan atau penolakan calon mahasiswa, tes didesain agar secara optimal dapat memberikan informasi mengenai kemampuan yang diperlukan agar sukses dalam belajar lebih lanjut di perguruan tinggi (tes berfungsi sebagai prediktor). Pemahaman ini berangkat dari asumsi adanya hubungan korelasional linier antara hasil ukur kemampuan saat ini dengan hasil ukur keberhasilan belajar di waktu yang akan datang.

Oleh karena itu, secara teoretis, dalam pengembangan tes yang mengungkap potensi sebagai prediktor maka validitas tes harus teruji secara konstrak (*construct-valid*) sedangkan dalam pengembangan tes yang mengungkap hasil belajar sebagai prediktor maka validitas tes harus harus teruji secara logis (*content-valid*). Namun demikian, guna mengetahui apakah asumsi adanya hubungan korelasional antara tes

dan kriterion keberhasilan belajar lebih lanjut didukung oleh data empiris, diperlukan pengujian validitas prediktif (*criterion-related*). Tingginya koefisien korelasi antara tes sebagai prediktor dengan indikator performans belajar menunjukkan bahwa tes memiliki kemampuan memprediksi keberhasilan dalam belajar. Skor tes semacam ini, apabila digunakan dalam pengambilan keputusan untuk menerima atau menolak calon mahasiswa, akan menghasilkan keputusan yang tepat.

Keputusan seleksi yang tepat sebagai buah dari prediksi tes yang valid adalah ketika calon yang diterima kemudian terbukti memang berhasil menunjukkan performans yang tinggi sedangkan calon yang ditolak, andai diberi kesempatan, memang tidak mampu memperlihatkan prestasi yang baik. Sebaliknya, keputusan yang tidak tepat akan mengorbankan potensi calon yang sesungguhnya dapat berprestasi baik andaikan diterima atau mengorbankan kesempatan yang semestinya dapat diberikan kepada calon yang potensial yang terlanjur ditolak.

Sebenarnya, memperlihatkan bahwa suatu koefisien validitas lebih tinggi dari yang lainnya atau menunjukkan bahwa koefisien validitas itu signifikan tidak banyak membantu untuk memahami seberapa besar tes yang bersangkutan mampu meningkatkan ketepatan prediksi. Sifat statistik yang *sample-dependent* dan adanya *sampling error* memungkinkan diperolehnya koefisien validitas prediktif berbeda antara kelompok subjek yang satu dan kelompok subjek yang lain. Bahkan dari kelompok subjek yang sama sekalipun, koefisien yang berbeda dapat terjadi bila pengukuran dilakukan pada waktu yang berbeda atau pada ukuran sampel yang tidak sama. Perbedaan yang mungkin timbul semacam ini sudah sangat dimaklumi oleh para ahli pengukuran sebagai salah-satu kelemahan aplikasi pendekatan teori tes klasik.

Hadirin yang saya hormati,

Persoalannya menjadi lain apabila koefisien validitas yang berbeda antara satu subkelompok dengan subkelompok lainnya itu diakibatkan oleh faktor-faktor yang tidak relevan dengan tujuan tes namun secara sistematis mempengaruhi besarnya koefisien validitas.

Tes menjadi bias terhadap satu subkelompok bila validitas bagi subkelompok tersebut menjadi berbeda semata-mata karena subjeknya berasal dari kelompok tertentu (validitas diferensial). Pada sisi lain, faktor yang tidak relevan tersebut dapat bekerja mempengaruhi fungsi prediksi pada satu subkelompok tapi tidak pada subkelompok yang lain, sehingga apabila tes digunakan untuk memprediksi berdasar persamaan regresi yang berbeda tersebut, maka dapat timbul kondisi yang disebut prediksi diferensial yang dapat membawa kepada ketidakadilan keputusan (*unfairness*) dalam seleksi.

Banyak ahli yang membedakan antara pengertian sifat tes yang tidak bias dengan *fairness* dalam tes, yaitu bahwa bias mengacu pada ciri intrinsik suatu tes sedangkan *fairness* mengacu pada masalah etikal mengenai penggunaan tes. Sandifer mengatakan bahwa bias dalam tes mengacu pada keuntungan yang tidak *fair* yang diperoleh oleh satu kelompok subjek, baik pada satu aitem mau pun pada seluruh tes. Bias dapat dinyatakan secara kuantitatif atau secara kualitatif (Sandifer, 2001). Ahli yang lain juga mengatakan bahwa *fairness* bukanlah kualitas dari tes sebagaimana bias, melainkan mengacu pada masalah cara pemakaian hasil tes (Banicky & Foss, 2000), bahkan bias dapat didefinisikan sebagai salah-satu bentuk ketidakvalidan tes (Green, 1975; Reynolds, 1982; dalam Shepard, 1982). Karena itu Shepard mengatakan bahwa dalam konteks seleksi, bias dipandang sebagai validitas prediktif diferensial (Shepard, 1982).

Hasil tes masuk perguruan tinggi sering mengungkapkan perbedaan rata-rata skor yang substansial di antara subkelompok etnis, seks, dan sosioekonomis. Seringkali perbedaan tersebut dianggap sebagai bukti yang cukup untuk mengatakan adanya bias tes, namun dari perspektif psikometri, *test's fairness* tidak dapat lepas dari aspek validitasnya. Menurut Cole dan Moss, adanya bias tes (yaitu pelanggaran terhadap *fairness*) terjadi 'bila skor tes memiliki makna atau implikasi bagi suatu kelompok tertentu yang berbeda dari makna atau implikasi bagi kelompok lain'. Bias adalah validitas diferensial dari interpretasi skor tes (Cole & Moss, dalam Zwick, 2007). Dapat dikatakan dengan tegas bahwa pertanyaan mengenai ekuitas dan *fairness* langsung berkaitan dengan pertanyaan mengenai validitas, karena argumen yang kuat mengenai validitas interpretasi skor tes harus memperlihatkan bahwa interpretasi tersebut dapat

digeneralisasikan pada subkelompok subjek yang berbeda dan bahwa skor tes dapat digunakan secara *fair* (Young, 2008).

Jauh sebelumnya, Thorndike telah mengatakan bahwa ada-tidaknya perbedaan skor rata-rata di antara dua kelompok atau perbedaan variabilitas tidak langsung memberi informasi apapun mengenai *fairness* penggunaan tes. Memang adanya perbedaan lebih menarik untuk diperhatikan, namun perbedaan skor itu sendiri bukanlah bukti *unfairness*. Korelasi antara perbedaan tersebut perlu diuji. Bahkan masing-masing korelasi perlu diuji lebih dahulu untuk mengetahui sejauhmana inferensi dapat diberlakukan pada kelompok subjek yang bersangkutan. Bila tidak terdapat korelasi yang berarti maka tidak ada dasar yang kuat untuk melakukan inferensi apapun mengenai masalah *fair* atau *unfair* suatu tes sehingga, sebagai konsekuensinya, persoalannya bukan lagi mengenai tes yang *unfair* melainkan mengenai tes yang *irrelevant*, yaitu yang digunakan tidak sesuai dengan tujuannya. Dalam hal ini, tidak ada alasan apapun yang dapat membenarkan penggunaan tes yang irelevan guna melakukan inferensi terhadap kelompok manapun (Thorndike, 1971).

Hadirin yang saya hormati,

Akar dari adanya bias tes adalah terdapatnya aitem-aitem dalam tes yang tidak berfungsi sama terhadap kelompok yang berbeda, lepas dari fakta bahwa kedua kelompok tersebut sebenarnya memiliki kemampuan yang setara. Bentuk bias aitem yang dikenal dengan nama *differential item functioning* (DIF) ini terjadi bilamana dua orang yang memiliki tingkat kemampuan setara tapi berasal dari kelompok yang berbeda tidak memiliki peluang yang sama untuk memilih jawaban yang benar (Gieri, et al., 2003). Dorans menyebut DIF dengan nama *unexpected differential item performance* dan menyamakannya dengan *item bias* atau *item unfairness* (Dorans, 1989). Ahli lain mengatakan bahwa sekalipun DIF merupakan syarat untuk terjadinya bias aitem, namun bias aitem dan bias tes tidak ditentukan semata-mata oleh adanya DIF (Schumacker, 2005).

Semestinya, subjek akan merespon soal secara kognitif sesuai dengan kapasitas mental yang dimilikinya. Artinya, proses mental yang terlibat dalam penyelesaian masalah atau penemuan jawaban

semata-mata merupakan kinerja kemampuan kognitif yang sedang diukur oleh tes. Ketika dalam aitem terdapat faktor-faktor yang tidak relevan dengan kemampuan yang sedang diukur namun berkaitan dengan ciri tipikal subjek tertentu sehingga ikut mempengaruhi secara sistematik proses mental yang terjadi dan membuat respon yang diberikan subjek tidak lagi semata-mata tergantung hanya pada kemampuan kognitif yang sedang diukur, terjadilah bias. Sumber bias dalam aitem ini oleh Kopeikin (2000) disebut sebagai *content bias*.

Dalam konteks seks atau jenis kelamin *content bias* mencakup aitem (soal dan respon), bacaan, materi stimulus, gambar, grafik, peta, dan hal-hal lain yang berkaitan dengan tes yang memperlihatkan aktivitas, emosi, pekerjaan, sifat, dan/atau situasi stereotipikal jenis kelamin tertentu. Stereotipe pada isi aitem memicu respon yang berbeda antara pria dan wanita. Sebagai contoh, bagian kalimat dalam soal yang berbunyi "Seorang tukang bangunan …" akan secara langsung dan tidak disadari membawa stereotipe maskulinitas dan menempatkan subjek wanita di 'luar' konteks soal sehingga ia merasa seakan-akan dihadapkan pada persoalan di luar dunianya sedangkan subjek pria tidak merasa terganggu. Apabila kalimat soal tersebut sedikit diubah misalnya menjadi "Seorang tukang bangunan bernama Siti …" maka baik subjek pria dan wanita dapat mengalami disonansi kognitif dikarenakan stereotipe pekerjaan dan stereotipe nama yang tidak sesuai. Sangat mungkin subjek wanita justru akan menjadi tertarik pada soal seperti itu dan karenanya dapat merespon dengan lebih baik.

Bias juga dapat terjadi dikarenakan situasi pengetesan yang disebut oleh Kopeikin (2000) sebagai *atmosphere bias.* Dalam konteks perbedaan jenis kelamin *atmosphere bias* terutama berupa suatu situasi psikologis yang tercipta akibat berbagai tekanan dalam menghadapi tes seperti tekanan keterbatasan waktu pengerjaan tes yang bersifat *speeded*, ketidaksukaan terhadap bentuk-bentuk soal tertentu semisal bentuk soal esai yang memaksa subjek berekspresi secara tertulis, kecenderungan berspekulasi dalam menghadapi soal pilihan-ganda, atau keberanian mengambil resiko yang berbeda antara pria dan wanita. Sekalipun belum ditemukan penelitiannya, namun subjek pria dan wanita mungkin saja bereaksi berbeda terhadap pengawas tes yang berlawanan jenis kelamin dengannya.

Childs (1990) mengatakan bahwa bias jenis kelamin dapat bersumber dari (a) materi atau referensi yang ofensif terhadap pria atau terhadap wanita, (b) referensi objek dan gagasan yang lebih akrab bagi wanita dan kurang akrab bagi pria, atau sebaliknya, dan (c) representasi yang tak seimbang antara pria dan wanita sebagai aktor dalam aitem atau peranan gender yang bersifat stereotipe. Jadi timbulnya bias adalah sebagai reaksi subjek yang berbeda terhadap isi dan karakteristik aitem yang secara sistematis ikut berpengaruh terhadap peluang keberhasilan subjek dalam menjawab soal yang bersangkutan.

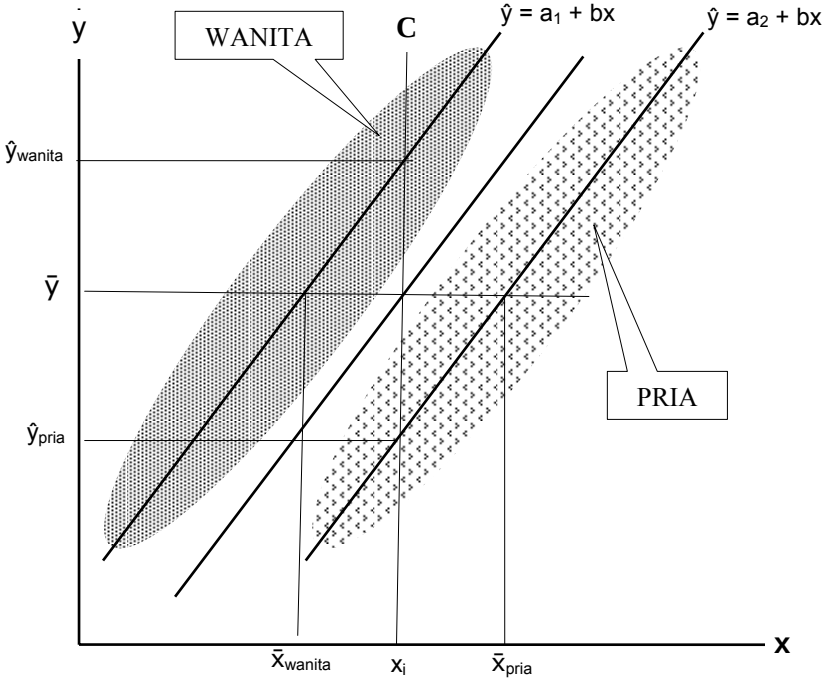
Hadirin yang saya hormati,

Dalam konteks tes sebagai prediktor terhadap performans di waktu yang akan datang maka validitas menyangkut fungsi prediktif tes, yaitu kemampuan skor tes untuk memprediksi. Karena fungsi prediktif skor tes berkaitan dengan akurasi dan kecermatan dalam pengambilan keputusan, apabila tes memiliki fungsi prediktif yang tidak sama bagi dua subkelompok yang berbeda maka akan terjadilah kondisi yang disebut prediksi diferensial atau disebut juga bias prediktif. Bias prediktif, pada gilirannya, akan menghasilkan keputusan seleksi yang tidak *fair*, yaitu sekalipun secara statistik mampu memprediksi sama efektifnya bagi kedua kelompok namun menghasilkan keputusan seleksi yang lebih menguntungkan salah-satu kelompok.

Adanya bias prediktif tidak dapat disimpulkan hanya dari perbedaan koefisien validitas saja melainkan menghendaki analisis perbandingan persamaan regresi di antara dua subkelompok (Linn & Werts, 1971). Cleary memberi ilustrasi grafis kondisi tersebut dalam Gambar 1. (Russell, 2000).

Dalam Gambar 1. terlihat bahwa kedua garis regresi memiliki *slope* regresi yang sama dan plot skater yang setara, namun memiliki *intercept* (titik potong garis regresi dengan sumbu Y) yang berbeda. Artinya kedua kelompok memiliki koefisien validitas prediktif yang identik. Apabila kedua garis regresi digunakan untuk seleksi dengan menarik garis pada titik batas kelulusan (C) yang sama, yaitu nilai x_i, maka tampaklah dua hal, yaitu pertama bahwa kelompok pria yang

lulus proporsinya lebih besar daripada kelompok wanita, dan ke dua bahwa pada nilai x yang sama, prediksi performans (ŷ) kelompok pria adalah lebih rendah dibanding prediksi performans (ŷ) kelompok wanita. Padahal bila mengikuti tujuan suatu proses seleksi, maka akan terjadi overprediksi terhadap calon dari kelompok pria dan dapat terjadi pula hilangnya kesempatan bagi calon dari kelompok wanita yang sebetulnya berpotensi tinggi.



Gambar 1. Tes yang Menghasilkan Keputusan Tidak Fair (diadaptasi dari Russell, 2000)

Jelaslah bahwa tes yang seharusnya digunakan dalam seleksi adalah tes atau perangkat tes yang hasil ukurnya tidak saja memiliki validitas tinggi namun juga dapat menghasilkan keputusan seleksi yang *fair* sehingga tidak merugikan atau menguntungkan salah satu kelompok subjek.

Hadirin yang saya hormati,

Young dan Kobrin menyajikan hasil reviu yang sangat komprehensif terhadap berbagai penelitian yang dilakukan sejak Tahun 1974 mengenai validitas diferensial dan prediksi diferensial tes masuk perguruan tinggi di Amerika berdasar kelompok etnis dan berdasar kelompok seks. Tigapuluh tujuh penelitian di antaranya adalah khusus mengenai validitas diferensial dan prediksi diferensial pada kelompok wanita dan kelompok pria. Hasil reviu dan analisis mereka dilaporkan dalam College Board Research Report No. 2001-6 berjudul 'Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis' (Young & Kobrin, 2001). Berikut adalah ringkasan dari hasil reviu mereka.

Young dan Kobrin menyimpulkan hasil reviu mereka bahwa secara umum hasil penelitian validitas diferensial menunjukkan besaran korelasi antara prediktor-prediktor dan berbagai kriterion nilai yang bemacam-macam secara konsisten agak lebih tinggi pada sampel wanita dibanding pada sampel pria (sekalipun hal ini tidak sepenuhnya berlaku pada kebanyakan perguruan tinggi yang selektif). Untuk hasil studi prediksi diferensial, dapat dikatakan bahwa temuan-temuan menunjukkan adanya underprediksi terhadap IP wanita. Pada beberapa perguruan tinggi yang sangat selektif di Amerika, underprediksi tersebut juga ditemukan sekalipun tingkatnya tidak begitu besar (Young & Kobrin, 2001). Young pernah melakukan studi replikasi mengenai validitas prediktif diferensial dalam memprediksikan performans akademik untuk membuktikan bahwa dugaan adanya validitas diferensial terjadi pula pada kelompok mahasiswa di universitas negara bagian (*state universities*) di Amerika. Hasil studinya mendukung fenomena yang telah ada sebelumnya bahwa dengan menggunakan analisis regresi prediktor tunggal ditemukan kecenderungan underprediksi terhadap performans kelompok mahasiswa wanita dan overprediksi terhadap performans kelompok minoritas. Bagi wanita, tidak bagi minoritas, perbedaan validitas prediktif tampaknya berkaitan dengan efek pemilihan mata pelajaran (Young, 1994).

Temuan bahwa performans wanita di perguruan tinggi cenderung diprediksi rendah oleh tes juga selaras dengan hasil

penelitian mengenai validitas tes standar seperti GRE yang digunakan dalam seleksi mahasiswa (Kuncel & Hezlett, 2007). Lewis dan Hoover menguji perbedaan prediksi Cognitive Aptitude Test (CogAT) terhadap prestasi pada The Iowa Tests of Basic Skills (ITBS) menurut jenis kelamin dengan sampel semua siswa kelas 2, 5, dan 8 yang mengikuti ITBS di musim gugur 1984 sejumlah kurang-lebih 5.000 anak laki-laki dan 5.000 orang anak perempuan di setiap tingkat kelas. Pengujian terhadap mean dan deviasi standar CogAT memperlihatkan bahwa anak laki-laki dan perempuan memiliki skor yang mirip di semua kelas. Namun bila skor ITBS diprediksi terpisah antara pria dan wanita maka ditemukan perbedaan yang signifikan pada *slope* dan *intercept* di antara persamaan prediksi pada beberapa subtes ITBS terutama pada subtes Membaca dan Ketrampilan Bahasa. Garis regresi untuk setiap subtes ITBS dan CogAT memperlihatkan overprediksi yang konsisten bagi anak laki-laki pada subtes Membaca, Materi Acuan, dan Ketrampilan Bahasa. Skor anak wanita secara konsisten diprediksi rendah pada subtes tersebut (Lewis & Hoover, 1987). Rudisill dan Morrison yakin mengenai adanya indikasi kuat bahwa perbedaan fisiologis berkaitan erat dengan kemampuan matematik yang menunjukkan bahwa pria lebih unggul daripada wanita. Bukti yang paling konklusif mengenai hal ini adalah pada bidang kemampuan visualisasi dan kemampuan spasial. Namun demikian, perbedaan fisiologis tidak dapat dipandang sebagai penyebab perbedaan prestasi karena hal itu dapat terjadi dikarenakan faktor pengalaman (Rudisill & Morrison, 1989). Berkaitan dengan faktor biologis, temuan lain menyimpulkan bahwa secara historis kelompok wanita cenderung mengungguli pria dalam menyelesaikan masalah yang bersifat verbal sedangkan kelompok pria cenderung mengungguli wanita dalam menyelesaikan soal-soal matematik (Tanner, 2001).

Rosser (1989) serta Gallagher dan De Lisi (1994) berpendapat bahwa format tes seperti SAT dapat merugikan peserta wanita. Rosser melakukan analisis yang mendetil terhadap jawaban per aitem untuk menentukan performans diferensial antara wanita dan pria. Ditemukan bahwa lebih banyak wanita dibanding pria yang membiarkan aitem tak terjawab dan bahkan lebih banyak lagi yang melewatkan lima pertanyaan terakhir pada tes Verbal dan sepuluh pertanyaan terakhir

pada tes Matematika SAT. Hal tersebut sangat mungkin dikarenakan wanita kurang berani mengambil resiko dan tidak mau menebak jawaban akibat peringatan adanya pengurangan skor bagi jawaban salah dalam SAT yang ditanggapi sangat serius oleh wanita. Diduga pula bahwa, dibanding pria, wanita lebih sulit menghadapi soal yang harus diselesaikan dalam tekanan waktu sangat terbatas seperti ketika menghadapi soal-soal matematika (Rosser, 1989). Linn dan Hyde (1995) juga menekankan adanya perbedaan pendekatan yang digunakan pria dan wanita dalam menjawab soal-soal SAT, sebagaimana diperlihatkan dalam studi Rosser (1989) tentang strategi menghadapi tes.

Belum ada jawaban yang konklusif mengenai penyebab terjadinya validitas diferensial dan prediksi diferensial pada tes masuk perguruan tinggi di Amerika dan mengenai fakta bahwa tes cenderung memprediksi rendah kemampuan wanita, namun beberapa hal berikut diduga ikut menjadi penyebab adanya kesenjangan performans antara pria dan wanita.

1. Soal tes yang bias
   Rosser (1989) menemukan bahwa soal tes yang jawaban benarnya berkaitan dengan perbedaan gender yang besar selalu menguntungkan pria sekalipun prestasi akademik wanita tinggi. Rosser juga menemukan bahwa umumnya wanita lebih baik dalam menjawab soal mengenai hubungan, estetika, dan humanitas sedangkan pria lebih berhasil dalam menjawab pertanyaan mengenai olah raga, IPA, dan bisnis. Kesimpulan ini didukung oleh temuan penelitian terdahulu oleh peneliti ETS (Educational Testing Service) Carol Dwyer yang meninjau kesenjangan gender dari perspektif historis. Dwyer menemukan bahwa di antara para penulis tes umumnya mengetahui bahwa perbedaan gender dapat dimanipulasi dengan sekedar memilih aitem-aitem tes yang berbeda. Sebagai contoh fakta diperlihatkan bahwa, di beberapa tahun awal pelaksanaan SAT, pria lebih tinggi skornya daripada wanita pada bagian Matematika tapi wanita selalu mengalahkan pria pada bagian Verbal. Pengambil kebijakan di ETS kemudian menetapkan perlunya penyeimbangan isi tes Verbal agar membantu pria dengan memperbanyak

pertanyaan yang berkaitan dengan politik, bisnis dan olah raga pada bagian Verbal. Sejak itu, pria mengungguli skor wanita baik pada bagian Matematika maupun pada bagian Verbal (Dwyer, 1976). Kenyataan tersebut berlanjut sebagaimana temuan Young dan Fisler yang meneliti sebanyak 69.284 siswa kelas akhir SLTA yang menempuh SAT di bulan November 1990, yaitu pada seksi Verbal ditemukan perbedaan rata-rata skor sebesar 4,68 poin dan pada seksi Matematika ditemukan perbedaan rata-rata skor sebesar 45,38 poin yang semua untuk keunggulan sampel pria dibanding sampel wanita (Young & Fisler, 2000).

2. Format pilihan-ganda
   Dari hasil penelitian yang diadakan bersama oleh ETS dan The College Board telah disimpulkan bahwa bentuk soal pilihan-ganda berpotensi bias terhadap wanita. Dalam suatu penelitian mengenai berbagai bentuk soal pada tes Advanced Placement (AP) yang dibuat oleh ETS untuk College Board, juga ditemukan bahwa kesenjangan skor antar gender berkurang atau hilang sama sekali pada semua bentuk soal lain (seperti jawaban-pendek, esai, dan respon terpola) kecuali pada bentuk pilihan-ganda. Hasil yang sama juga disimpulkan dari temuan pada tes California Bar Exam dan the SAT's English Composition Test with Essay.

3. Peluang menebak
   Tes pilihan-ganda dengan lima pilihan yang memberlakukan *guessing penalty*, yaitu mengurangi skor dengan seperempat bagi setiap jawaban yang salah dan memberi skor nol bagi pertanyaan yang tidak dijawab, dimaksudkan agar peserta tes yang tidak merasa pasti dengan jawabannya tidak membuat tebakan sembarang karena spekulasi dalam menjawab mengandung resiko kerugian skor. Penelitian menunjukkan bahwa pria cenderung lebih berani mengambil resiko dan akan menebak bila mereka tidak mengetahui jawaban, sedangkan wanita cenderung menjawab hanya bila mereka yakin betul bahwa jawaban mereka adalah benar dan cenderung tidak menebak. Ketidaksediaan melakukan tebakan pada ujian terbukti berdampak negatif terhadap skor tes.

4. Keterbatasan waktu
   Faktor lain yang ikut mempengaruhi adanya kesenjangan jenis kelamin adalah unsur keharusan bekerja cepat dalam merespon tes atau sifat *speeded* pada tes. Bukti menunjukkan bahwa wanita memiliki pendekatan pemecahan masalah yang berbeda dari pria. Pada umumnya wanita cenderung melihat problem secara menyeluruh, mempertimbangkan lebih dari satu kemungkinan jawaban yang benar dan memeriksa jawaban mereka. Pendekatan ini adalah baik untuk di sekolah dan di kehidupan sehari-hari namun akan merugikan sewaktu menghadapi soal ujian karena akan sangat menghabiskan waktu. Berbagai studi menemukan bahwa bila ujian diberikan tanpa tekanan keterbatasan waktu maka skor wanita akan meningkat tajam sedangkan skor pria tidak banyak berubah dibanding dengan ujian yang harus diselesaikan dalam waktu terbatas yang menimbulkan rasa tertekan (The National Center for Fair & Open Testing, 2007).

Apakah perbedaan performans antara pria dan wanita baik pada tingkat tes maupun pada tingkat aitem berarti bahwa tes merupakan ukuran yang bias terhadap kemampuan subjek ataukah skornya mencerminkan perbedaan yang sesungguhnya di antara kedua kelompok yang menempuh tes? Wanita ternyata memiliki performans yang lebih baik dalam menghadapi soal-soal konvensional dibanding ketika menghadapi soal inkonvensional, sedangkan subjek pria sebaliknya. Terdapat overlap yang berarti dalam strategi pemecahan masalah antara pria dan wanita namun wanita cenderung lebih mengandalkan strategi konvensional. Penggunaan strategi pemecahan masalah konvensional berkorelasi dengan sikap-sikap yang negatif seperti membenci matematika dan menganggapnya kurang penting, sedangkan sikap yang positif berkaitan dengan penggunaan strategi inkonvensional. Temuan ini mendukung pandangan bahwa perbedaan gender dalam skor tes, paling tidak sebagiannya, dikarenakan strategi pemecahan masalah (Gallagher & De Lisi, 1994). McCornack dan McLeod (1988) mengemukakan kemungkinan bahwa salah-satu sumber terjadinya bias gender adalah macam atau jenis mata pelajaran yang nilainya masuk dalam penghitungan GPA. Wanita cenderung memilih pelajaran-pelajaran yang standar penilaiannya tidak begitu

ketat. Wanita ternyata sangat pandai dalam memilih mata kuliah yang cocok dengan kemampuan mereka (Decore, 1984 dalam McCornack & McLeod, 1988).

Hadirin yang saya hormati,

Ratusan fakta dan bukti sebagai hasil penelitian mengenai bias tes dan implikasinya terhadap *fairness* dalam keputusan seleksi masuk perguruan tinggi di luar negeri, terutama di Amerika Serikat, telah dipublikasikan sejak lama dan telah menjadi domain publik. Hasil penelitian tersebut disikapi secara positif oleh pihak penerbit tes utama seperti The College Board, Educational Testing Service, dll. karena memberikan informasi yang sangat bermanfaat bagi usaha perbaikan dan peningkatan kualitas berbagai tes standar yang dibuat di negara maju. Di Indonesia, yang sistem pengujiannya masih dalam taraf perkembangan relatif awal dan jumlah pakar dalam bidang ilmu pengukurannya masih sangat terbatas, usaha evaluasi terhadap instrumen pengujian yang mencakup aspek bias dan dan validitas prediktif diferensial dapat dikatakan hampir belum ada.

Hadirin yang saya hormati,

Sebagai kesimpulan dari uraian ringkas ini adalah bahwa dengan berkaca pada kemajuan perkembangan sistem tes dan evaluasinya di negara-negara maju, sudah saatnya penelitian dan evaluasi terhadap kualitas tes yang digunakan dalam pengambilan keputusan seleksi dalam *high stake exams* seperti ujian masuk perguruan tinggi di Indonesia diperluas lebih dari sekedar evaluasi terhadap validas prediktifnya saja.

Secara lebih spesifik, dapat direkomendasikan kepada pihak pengambil keputusan di perguruan tinggi dan pihak-pihak yang berkepentingan dengan pengembangan perangkat tes tulis dalam seleksi masuk perguruan tinggi, untuk:

1. Memperluas kajian mengenai sifat prediktif diferensial tes tulis dalam pengambilan kebijakan seleksi mahasiswa baru, khususnya menyangkut kebijakan bagi calon mahasiswa wanita dan calon mahasiswa pria.

2. Mempertimbangkan kemungkinan adanya perbedaan daya prediktif masing-masing tes di program studi yang berbeda dalam penggunaan skornya sebagai dasar pengambilan keputusan seleksi. Disarankan untuk merumuskan kombinasi tes dan bobot skor yang sesuai bagi masing-masing program studi sehingga diperoleh batas penerimaan yang tepat.
3. Meneliti pelaksanaan sistem penilaian hasil belajar masing-masing program studi dan meningkatkan kualitas penilaian dengan menggiatkan pelatihan tehnik penulisan tes hasil belajar semester beserta evaluasinya bagi para dosen.
4. Membentuk tim penulis soal tes seleksi yang keanggotaannya tetap sehingga dapat dibina dengan pelatihan khusus guna meningkatkan ketrampilan menulis soal yang dapat meminimalkan potensi bias.
5. Membangun sistem *item banking* untuk soal tes seleksi yang dapat menyediakan informasi mengenai karakteristik aitem-aitem yang dilandasi oleh hasil analisis empiris yang menggunakan pendekatan *item response theory* (IRT).

Sekian dan terimakasih.

Wassalamu alaykum warohmatullahi wa barokatuh.

DAFTAR PUSTAKA

Banicky, L.A. & Foss, H.K. (2000). Assessing student learning. *Manual Document.* Delaware Education Research and Development Center, University of Delaware.

Beatty, A., Greenwood, M.R.C., & Linn, R.L. (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions.* Washington, DC: National Academy Press.

Childs, R.A. (1990). *Gender bias and fairness. ERIC Digest*. ERIC Clearinghouse on Tests Measurement and Evaluation, Washington DC: American Institutes for Research.

Cole, N.S. & Zieky, M.J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 4, 369-382.

Departemen Pendidikan dan Kebudayaan (1990): Model-model Sistem Seleksi Masuk Perguruan Tinggi Negeri. *Makalah,* Pusat Penelitian dan Pengembangan Sistem Pengujian.

Dorans, N.J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*, 3, 217-233.

Dwyer, C. (1976). *Fair test.* The National Center for Fair & Open Testing. http://www.fairtest.org/facts/genderbias.htm

Gallagher, A.M. & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test – Mathematics problem solving among high-ability students. *Journal of Educational Psychology*, *86*, 2, 204-11.

Gieri, M.J., Bisanz, J., Bisanz, G.L., & Boughton, K.A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*, 4, 281-306.

Hanson, G.R., Cole, N.S., & Lamb, R.R. (1973). Sex bias in selective college admissions. Dalam J. Pottker & A. Fishel (Eds.), *Sex Bias in The Schools: The Research Evidence.* Rutherford: Fairleigh Dickinson University Press.

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 1-123. Dalam *Citation Classics* Number 41, October 9, 1978.

Kopeikin, H.S. (2000). Test bias. *Psychology 121, Lecture 11.* University of California at Santa Barbara.

Kuncel, N.R. & Hezlett, S.A. (2007). Standardized tests predict graduate students' success. *Science, 315,* 5815, 1080-1081.

Lewis, J.C. & Hoover, H.D. (1987). Differential prediction of academic achievement in elementary and junior high school by sex. *The Journal of Early Adolescence*, *7*, 1, 107-115.

Linn, M.C. & Hyde, J.S. (1995). Gender, mathematics, and science. Dalam A.S. McDonald, P.E. Newton, C. Whetton, & P. Benefield, *Aptitude Testing for University Entrance: A Literature Review*. nfer.

Linn, R.L. & Werts, C.E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1, 1-4.

Mardapi, D. & Azwar, S. (1989). Equity pada sistem seleksi masuk perguruan tinggi negeri. *Paper,* Pusat Penelitian dan Pengembangan Sistem Pengujian, DepDikBud.

McCornack, R.L. & McLeod, M.M. (1988). Gender bias in the prediction of college course performance. *Journal of Educational Measurement*, *25*, 4, 321-331.

Reynolds, C.R. (1982). Methods for detecting construct and predictive bias. Dalam R.A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: The Johns Hopkins University Press.

Rosser, P. (1989). The SAT gender gap: Identifying the causes. *Gender Bias in College Admissions Tests*.

Rudisill, E.M. & Morrison, L.J. (1989). Sex differences in mathematics achievement: An emerging case for physiological factors. Dalam A.S. McDonald, P.E. Newton, C. Whetton, & P. Benefield, *Aptitude Testing for University Entrance: A Literature Review*. nfer.

Russell, C.J. (2000). The Cleary model: Test bias. *EEOC Uniform Guidelines on Employment Selection Procedures.* http://www.ou.edu/russell/whitepapers/Cleary_model.pdf

Sandifer, M. (2001). Testing, testing - Avoiding bias in testing. *The Bar Examiner*.

Schumacker, R.E. (2005). Test bias and differential item functioning. *Applied Measurement Associates*. http://www.appliedmeasurementassociates.com/White%20Papers/TEST%20BIAS%20AND%20DIFFERENTIAL%20ITEM%20FUNCTIONING.pdf

Shepard, L.A. (1982). Definitions of bias. Dalam R.A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: The Johns Hopkins University Press.

Suryabrata, S. (1989): Seleksi Masuk Perguruan Tinggi - Latar Belakang dan Kecermatan Prediksi, *Paper*. Pusat Penelitian dan Pengembangan Sistem Pengujian, DepDikBud.

Suryabrata, S. (2005): Pengembangan Sistem Seleksi Calon Mahasiswa Perguruan Tinggi yang Akurat dan Berkeadilan. *Rekayasa Sistem Penilaian dalam Rangka Meningkatkan Kualitas Pendidikan*, Yogyakarta: HEPI.

Tanner, D.E. (2001). *Assessing academic achievement.* Boston, MA: Allyn and Bacon.

The National Center for Fair & Open Testing (2007). *Gender bias in college admissions tests.* http://www.fairtest.org/facts/genderbias.htm

Thorndike, R.L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 2, 63-70.

Uwakwe, B.U. (2004): Building Fairness and Equity into Classroom-Based Assessment in the Nigerian Secondary School System. *Paper,* Departement of Guidance & Counselling, University of Ibadan, Nigeria.

Young, J.W. (2008). Differential validity and differential prediction. *Paper presented at the annual meeting of the National Council on Measurement in Education.* New York, March 27, 2008 dalam the NCME symposium "Different facets of test equity and fairness".

Young, J.W. & Kobrin, J.L. (2001). Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis. *College Board Research Report*, No. 2001-6. New York, NY: College Entrance Examination Board.

Young, J.W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement*, *54*, 4, 1022-1029.

Young, J.W. & Fisler, J.L. (2000). Sex differences on the SAT: An analysis of demographic and educational variables. *Research in Higher Education, 41,* 3, 401-416.

Zwick, R. (2007). *College admissions testing,* National Association of College Admission Counseling.