DIFFERENTIAL ITEM FUNCTIONING PADA TES MULTIDIMENSI

DISERTASI

Oleh:

Ali Ridho
10/306834/SPS/250

PROGRAM DOKTOR PSIKOLOGI
FAKULTAS PSIKOLOGI
UNIVERSITAS GADJAH MADA
2014

Abstrak

Terdeteksinya *differential item functioning* (DIF) pada sebuah aitem tes yang bersifat unidimensi sering diatribusikan pada asumsi multidimensionalitas yang terkandung didalamnya. Oleh sebab itu, menarik untuk membuktikan apakah DIF yang terkandung dalam aitem-aitem tes Potensi Akademik Seleksi Penerimaan Mahasiswa Baru Perguruan Tinggi Agama Islam Negeri (PA SPMB-PTAIN) tahun 2012 disebabkan oleh multidimensionalitas yang berbeda pada kelompok fokal dan referensi. Tujuan penelitian ini adalah: (1) mengevaluasi aitem-aitem PA SPMB-PTAIN tahun 2012 yang mengandung DIF berdasarkan kelompok gender (perempuan – lelaki, PL) dan sekolah (MAN – SMAN, MS); (2) mengevaluasi sejauh mana efek multidimensionalitas pada DIF tidaknya aitem; dan (3) mengevaluasi karakteristik psikometrik secara umum aitem-aitem yang terdeteksi DIF pada tes PA SPMB-PTAIN 2012. Hasil analisis terhadap data respons peserta (n = 14000 untuk PL; n = 10000 untuk MS) pada komponen Analogis, Logis, Analitis, Aritmetika, Komparasi, dan Geometri menunjukkan bahwa: (1) DIF multidimensi terjadi pada sembilan aitem dalam masing-masing komponen analogi, analitik, aritmetika dan geometri yang menguntungkan lelaki; tujuh aitem pada komponen aritmetika dan komparasi yang menguntungkan perempuan; dan lima aitem DIF pada komponen analogis, logis, dan aritmetika yang semuanya menguntungkan siswa lulusan SMAN; (2) Perbedaan multidimensionalitas antara kelompok fokal dan referensi tidak memicu munculnya DIF pada aitem; dan (3) aitem-aitem yang terdeteksi DIF multidimensi memiliki karakteristik sensitivitas daya beda yang tidak sama pada kelompok fokal dan referensi.

Kata kunci: *multidimensionalitas, item response theory*, DIF, *perempuan – lelaki*, *MAN – SMAN*, *potensi akademik*.

Abstract

The presence of differential item functioning (DIF) to an item assumed unidimensional often attributed to the assumption of multidimensionality. Therefore, it's interested to prove whether multidimensionality elicited items to be DIF between focal and reference group in tes Potensi Akademik (scholastic aptitude) Seleksi Penerimaan Mahasiswa Baru Perguruan Tinggi Agama Islam Negeri (PA SPMB-PTAIN). This study aimed at: (1) evaluate items of PA SPMB-PTAIN that contained DIF based on gender (female – male) and graduated (MAN – SMAN, MS); (2) evaluate the effect of multidimensionality to DIF; and (3) evaluate general psychometric characteristics of items that contain DIF. Analyzing to data response (n = 14000 for FM; n = 10000 for MS), the result showed that: (1) 9 items of analogy, analytic, arithmetic and geometry components were significantly contained multidimensional DIF favored Male; 7 items of arithmetic and comparative components contained multidimensional DIF favored Female; 5 items of analogy, logic, and arithmetic components contained multidimensional DIF favored student who graduated from SMAN; (2) the multidimensionality differences between focal and reference group did not elicit items to be DIF; (3) detected items as multidimensional DIF have sensitivity different of discrimination between focal and reference group.

Keyword: multidimensionality, item response theory, DIF, *female – male*, *MAN – SMAN*, *scholastic aptitude*.

Daftar Isi

Daftar Tabel

Daftar Gambar

Daftar Lampiran

BAB I

PENDAHULUAN

A. Latar Belakang

Item response theory (IRT) merupakan pemodelan matematik yang berusaha menjelaskan interaksi antara para peserta dengan aitem-aitem tes yang mengukur atribut laten sehingga menghasilkan suatu pola respons tertentu. IRT secara luas digunakan dalam dunia psikologi dan pendidikan (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; McKinley & Mills, 1985). Aplikasinya banyak bermanfaat dalam menyelesaikan persoalan-persoalan pengukuran; antara lain bank aitem, pengembangan tes baru, penyetaraan skor, dan *computerized adaptive testing* (CAT). Dalam IRT, secara teoritik, karakteristik aitem bersifat *group independent* dan skor yang mendeskripsikan peserta tes akan bersifat *item independent* (Hambleton & Swaminathan, 1985). Kondisi ini dapat terjadi bila interaksi antara para peserta dengan aitem-aitem tes menghasilkan respons yang memiliki kecocokan dengan model IRT yang dipilih. Dalam perkembangannya sampai dengan sekarang, IRT mampu mengakomodir respons peserta tes dalam model *unidimensional IRT* (UIRT) dan *multidimensional IRT* (MIRT).

Salah satu prinsip utama yang mendasari UIRT adalah unidimensionalitas yang berarti hanya terdapat satu atribut laten yang mendasari para peserta tes dalam menjawab aitem (Hambleton dkk., 1991; Lord, 1980). Sekumpulan aitem-aitem dalam tes dapat disebut unidimensional bila kinerja pada peserta tes dapat dijelaskan oleh sebuah atribut laten secara tunggal (Hambleton & Rovinelli, 1986). Jadi, tes atau subtes disebut unidimensi bila dalam menjawab benar aitem-aitem

tes para peserta membutuhkan sebuah atribut laten (θ) yang bersifat tunggal, tanpa atribut laten lain (θ lain) selain yang memang hendak diukur. Apabila terdapat atribut laten (θ) lebih dari satu yang dibutuhkan peserta untuk bisa menjawab benar sebuah aitem, hal ini menunjukkan bahwa aitem tersebut bersifat multidimensi.

Lebih jauh, pada tes pilihan ganda yang bersifat unidimensi misalnya, probabilitas menjawab benar pada sebuah aitem hanya dipengaruhi oleh parameter aitem, sebuah atribut laten θ, dan bukan yang lain. Inilah yang disebut dengan prinsip independensi lokal (*local independence*, LI) (Lord, 1980). Dengan demikian benarnya seorang peserta dalam menjawab sebuah aitem dipengaruhi oleh hanya sebuah atribut laten θ. Manakala sebuah θ belum mampu menjelaskan probabilitas benarnya peserta dalam menjawab sebuah aitem, ini berarti LI belum terpenuhi (Stout, 1984, 1989, 2002). Artinya, bila terdapat θ lain yang ikut mempengaruhi probabilitas menjawab benar, berarti LI belum terpenuhi.

Banyak tes yang didesain bersifat unidimensi, namun dalam kenyataannya sering kali kinerja peserta tes pada aitem tidak hanya dilandasi oleh atribut laten tunggal, namun lebih dari satu θ. Adanya θ lain diluar atribut laten yang dikehendaki akhirnya membuat aitem menjadi multidimensi (Oshima & Miller, 1992; Oshima, Raju, & Flowers, 1997; Roussos & Stout, 1996a). Dengan kata lain, asumsi unidimensional kadang bersifat problematik, yaitu ketika aitem-aitem tes didesain untuk mengukur satu atribut laten tertentu (bersifat unidimensi) namun ternyata para peserta memerlukan lebih dari satu θ dalam menjawab benar sebuah aitem. Meskipun aitem-aitem tes didesain untuk mengukur satu θ tertentu, dapat terjadi para peserta memerlukan lebih dari satu θ dalam menjawab benar sebuah aitem. Implikasinya, data respons sebenarnya bersifat multidimensi. Masalahnya

adalah manakala keadaan tersebut terjadi pada tes yang didesain bersifat unidimensi. Data respons yang bersifat multidimensi kemudian diperlakukan sebagai data unidimensi berarti menyisakan problematika pada asumsi unidimensionalitas yang dijadikan sandaran.

Memenuhi kebutuhan seleksi mahasiswa baru, sebagai pengelola Perguruan Tinggi Negeri (PTN) dengan karakteristik yang khas, Direktorat Pendidikan Tinggi Islam (Diktis) Kementerian Agama berinisiatif untuk membuat sistem seleksi secara bersama bagi seluruh Perguruan Tinggi Agama Islam Negeri (PTAIN). Tahun 2010 terealisir sistem seleksi pertama yang diberi nama Seleksi Penerimaan Mahasiswa Baru Perguruan Tinggi Agama Islam Negeri (SPMB-PTAIN). Salah mata uji dalam SPMB-PTAIN adalah Test Potensi Akademik (PA). Mata uji PA SPMB-PTAIN tahun 2012 terdiri dari dua subtes; yaitu subtes verbal yang memiliki komponen-komponen: analogis, logis, dan analitis; dan subtes kuantitatif terdiri dari komponen-komponen: aritmetika, komparasi, dan bentuk geometri. Kedua subtes diasumsikan bersifat unidimensi serta membentuk struktur potensi yang diukur.

Asumsi bahwa PA SPMB-PTAIN tahun 2012 bersifat unidimensi, sebenarnya menyisakan persoalan karena secara empirik data respons peserta tes PA menunjukkan adanya multidimensionalitas (Azwar & Ridho, 2012; Ridho, 2011). Kenyataan hasil dua penelitian tersebut menjadi dasar pentingnya memverifikasi bahwa data respons SPMB-PTAIN tahun 2012 juga bersifat multidimensi.

1. Pentingnya Evaluasi Dimensionalitas

Problematika asumsi unidimensionalitas pada tes yang didesain bersifat unidimensi telah memicu para ahli (misalnya Bolt, Wollack, & Suh, 2012; Hambleton dkk., 1991; Jang & Roussos, 2007; Tate, 2003; Zhang, 2012b) merekomendasikan untuk mengevaluasi dimensionalitas dan struktur tes dalam setiap tahapan pengukuran. Hartig dan Höhler (2009) menambahkan bahwa pertanyaan krusial sebelum menerapkan model pengukuran tertentu adalah apakah konstrak yang hendak diungkap bersifat satu atau banyak dimensi. Dengan melakukan penyelidikan tentang dimensi-dimensi yang melatarbelakangi peserta dalam menjawab benar aitem-aitem dalam tes, peneliti dapat mengaitkan antara interpretasi substantif dengan hasil analisis statistik demi pemahaman yang lebih komprehensif pada interaksi aitem-aitem dan peserta tes. Dengan demikian penyelidikan tentang dimensionalitas akan memberikan kontribusi pada bukti-bukti validitas skor yang dihasilkan tes. Mengacu pada Messick (1996) dalam aspek konsekuensi, hasil penyelidikan dimensionalitas ini akan berpengaruh terhadap bagaimana skor diperlakukan, apakah tiap dimensi berdiri sendiri atau sebagai komposit.

Penyelidikan dimensionalitas terkait dengan struktur dimensi. Penyelidikan tentang struktur dimensi yang diungkap oleh tes merupakan hal yang penting karena hal ini terkait secara langsung dengan proses kalibrasi parameter aitem dan atribut laten berdasarkan data respons peserta tes. Atribut laten peserta tes dapat disimpulkan melalui model psikometrik, misalnya model IRT satu, dua, atau tiga parameter. Pada model-model ini, asumsinya adalah aitem-aitem tes mengukur atribut laten (θ) tunggal sehingga bersifat unidimensi. Pada tes yang diklaim bersifat unidimensi, sampai sejauh mana penyimpangan data dari

unidimensionalitas merupakan hal yang penting untuk ditemukan terlebih dahulu sebelum melakukan kalibrasi. Apabila pada kenyataannya tes tersebut bersifat multidimensi, estimasi θ yang dihasilkan akan menjadi tidak akurat sehingga mengancam validitas skor.

Pada tes skala besar (*large scale testing*) semisal tes masuk perguruan tinggi, umumnya ada banyak atribut laten yang diungkap dalam tes tersebut sehingga akan mengakibatkan data respons para peserta tes bersifat multidimensional. Oleh sebab itu penting kiranya mengetahui struktur data respons peserta tes bila akan mengambil suatu kesimpulan berdasarkan hasil tes tersebut. Bahkan Stone dan Yeh (2006) mengatakan bahwa melalui penyelidikan struktur internal tes, peneliti akan dapat mengidentifikasi domain yang diukur, mengidentifikasi hubungan antar dimensi, menguatkan dugaan multidimensionalitas dan interpretasi skor, dan mengidentifikasi varians konstrak yang tidak relevan.

Selain itu, penyelidikan struktur internal tes akan mengarahkan peneliti pada temuan konstrak yang dikehendaki dan tidak dikehendaki oleh tes. Sebuah tes yang dikembangkan, ditujukan untuk mengukur atribut laten tertentu dan didasarkan struktur yang bersifat teoritik tertentu pula. Hal ini perlu diteliti dan dikonfirmasi. Evaluasi dimensionalitas memberikan bukti empirik dalam menyelidiki struktur internal konstrak yang didefinisikan oleh pengembang tes, dan menemukan sejauh mana struktur teoritik tersebut terbukti berdasarkan respons peserta tes. Oleh sebab itu penyelidikan tentang dimensionalitas sekaligus struktur internal tes melalui data respons merupakan hal esensial. Hal ini semestinya dilakukan sebelum tes digunakan secara luas sehingga pada saat digunakan struktur internal tes sudah teridentifikasi dengan baik.

Isu mengenai dimensi dalam tes penting untuk diteliti karena hal tersebut juga mempengaruhi sistem penskoran, analisis data dan laporan hasilnya (Abedi, 1997; Kahraman & Thompson, 2011). Tanpa melalui evaluasi dimensionalitas data respons, model pengukuran dan kesimpulan yang dihasilkan akan berpengaruh pada beberapa aspek dalam tes; diantaranya adalah hasil estimasi parameter aitem dan parameter peserta (atribut laten, θ) yang tidak akurat, biasnya aitem dan tes, dan kelirunya interpretasi skor (misalnya Ackerman, 1994; Ackerman & Evans, 1994; Beretvas, Cawthon, Lockhart, & Kaye, 2012; Walker & Beretvas, 2003; Yen & Walker, 2007).

Dalam konteks sebuah tes terdiri dari beberapa subtes, hal-hal yang perlu diperhatikan sebelum menginterpretasikan skor komposit tersebut, yaitu, pertama, perlu diketahui atribut laten komposit apa yang hendak diukur (Ackerman, 1994; Reckase & McKinley, 1991). Kedua, perlu dipastikan seluruh peserta tes diukur pada skala komposit yang sama melalui suatu model kombinasi atribut laten (Yao, 2011). Formula komposit ini akan tergantung pada struktur konstrak, bersifat sederhana atau kompleks (Yao, 2012).

Selain akurasi hasil estimasi parameter berdasarkan model psikometrik tertentu, penyelidikan dimensionalitas merupakan kegiatan yang penting dalam kaitannya dengan kelompok peserta tes. Hal ini menyangkut penjaminan komparabilitas skor pada kelompok-kelompok yang teridentifikasi, setelah disetarakan dalam metrik yang sama. Pada konteks administrasi tes dalam skala luas, komparabilitas skor antar kelompok dalam satu kali administrasi merupakan kondisi yang harus dapat ditegakkan. Komparabilitas juga penting dalam situasi kelompok antar waktu administrasi satu dengan administrasi yang lain agar perbandingan skor tes antar administrasi memiliki makna dalam kontinum yang

setara. Dengan kata lain, penegakan invariansi pengukuran alat tes guna menegakkan kesetaraan konstrak antar kelompok peserta dan waktu semestinya harus diselidiki agar skor antar kelompok betul-betul bersifat komparabel. Evaluasi dimensionalitas akan membantu proses penegakan prinsip invariansi pengukuran.

Bukti-bukti hasil evaluasi dimensionalitas akan membuat skor yang dihasilkan tes menjadi lebih bermakna. Sebagai contoh, dalam tes yang mengukur potensi akademik, evaluasi terhadap domain yang diukur yaitu kemampuan verbal, penalaran, dan kuantitatif dapat diverifikasi apakah data respons para peserta mengungkap dimensi verbal, penalaran dan kuantitatif. Dengan demikian, bila terbukti, kebermaknaan skor yang dihasilkan oleh tes mengandung interpretasi yang sejalan dengan tujuan ukurnya. Pelaporan hasil tes dapat dijustifikasi secara empirik pada level dimensi ataupun keseluruhan dimensi sebagai komposit.

2. Multidimensionalitas dan *Differential Item Functioning*

Menurut Haladyna (2004) *differential item functioning* (DIF) pada aitem-aitem tes yang bersifat unidimensi adalah aspek utama yang mencemari validitas, ditambah dengan adanya multidimensionalitas. Sebagaimana dikemukakan sebelumnya, identifikasi struktur dimensi dapat membantu menemukan varians konstrak yang tidak relevan sehingga terjadi multidimensionalitas. Seperti yang diindikasikan oleh Messick (1995), konstrak yang tidak relevan merupakan sumber utama biasnya interpretasi skor tes. Varians konstrak yang tidak relevan dapat memicu aitem berfungsi tidak sama antar kelompok sehingga aitem mengalami *differential item functioning* (DIF).

Untuk tes yang diasumsikan unidimensi, varians konstrak yang tidak relevan merupakan indikasi adanya multidimensionalitas dikarenakan adanya θ lebih dari

satu yang ikut berkontribusi pada kinerja peserta dalam menjawab aitem. Alat ukur yang pada kenyataannya bersifat multidimensi kemudian diasumsikan unidimensi akan memicu munculnya fungsi aitem yang berbeda pada kelompok yang berbeda lantaran dimensi ke dua diluar tujuan ukur ikut memberikan pengaruh kinerja peserta tes pada alat ukur (Angoff, 1993; Stone, Cook, Laitusis, & Frederick, 2010; Zumbo, 1999). Fungsi aitem yang berbeda ini lebih dikenal dengan istilah DIF. Dengan demikian dapat dikatakan bahwa DIF pada tes unidimensi terjadi disebabkan oleh θ tambahan diluar θ utama ikut memberikan kontribusi pada kinerja peserta tes. Implikasi lebih luas, data respons yang bersifat multidimensi, bila diperlakukan sebagai data unidimensi, berarti telah menyimpang dari asumsi unidimensionalitas, juga aspek struktur dari konstrak yang diukur (Messick, 1995).

Mengiringi teridentifikasinya problematika unidimensionalitas pada UIRT, muncullah model *multidimensional item response theory* (MIRT) (Reckase, 1985; Reckase & Ackerman, 1986). MIRT adalah pengembangan unidimensional IRT (UIRT). Pada model MIRT dimungkinkan aitem-aitem direspons benar oleh para peserta tes mendasarkan pada atribut laten lebih dari satu. UIRT dan MIRT memiliki sudut pandang yang berbeda dalam memaknai DIF pada sebuah aitem.

Fenomena DIF terkait dengan konsep invariansi parameter aitem pada dua kelompok. Dua kelompok yang dibandingkan ini biasa disebut sebagai kelompok fokal (F) dan kelompok referensi (R). Apabila dua kelompok berada pada metrik yang sama, sebuah aitem semestinya memiliki parameter yang sama pula pada dua kelompok tersebut. Probabilitas menjawab benar pada aitem tersebut juga akan sama pada dua kelompok peserta dengan tingkat atribut laten yang sama. DIF terjadi manakala parameter pada kedua kelompok tidak sama sehingga

berimplikasi pada probabilitas yang tidak sama dalam merespons sebuah aitem pada keseluruhan atau sebagian kontinum atribut laten.

Pendapat lain menyebutkan bahwa DIF disebabkan oleh adanya kontaminasi atribut tambahan diluar atribut utama yang diukur oleh aitem. Pada akhirnya atribut tambahan ini ikut berpengaruh terhadap probabilitas peserta dalam merespons (Ackerman, 1992b; Roussos & Stout, 1996a). Atribut tambahan ini diistilahkan sebagai *nuisance dimensions* atau dimensi pencemar atau pengganggu (Bolt, 2000; Bolt & Stout, 1996; Gierl & Khaliq, 2001). Dengan kata lain, kehadiran DIF boleh jadi merupakan sinyal hadirnya eksistensi faktor atau dimensi lain yang belum terungkap (Penfield & Camilli, 2007).

Beberapa peneliti memiliki premis bahwa DIF pada tes yang diasumsikan unidimensi disebabkan oleh multidimensionalitas. DIF diyakini dipicu oleh hadirnya paling tidak satu dimensi tambahan diluar dimensi utama yang diukur oleh aitem (misalnya Ackerman, 1994; Camilli, 1992; Gierl, 2005; Gierl, Bisanz, Bisanz, & Boughton, 2003; McDonald, 2000; Oshima & Miller, 1992; Roussos & Stout, 1996a). Jadi, teridentifikasinya DIF pada sebuah aitem pada tes unidimensi menunjukkan hadirnya atribut laten tambahan yang diukur oleh aitem dengan distribusi yang berbeda pada kedua kelompok yang dibandingkan.

Dimensionalitas dalam pengukuran dapat dimaknai sebagai banyaknya atribut laten yang mendasari peserta dalam merespons aitem-aitem tes (Chou & Wang, 2010). Dalam konteks tes kemampuan, dimensionalitas dapat disebut juga sebagai banyaknya kemampuan yang diukur oleh tes atau kumpulan aitem.

Dikaitkan dengan proses konstruksinya, dimensionalitas dapat dipandang sebagai faktor-faktor pengukuran yang didesain untuk diukur oleh tes (Mislevy, Almond, & Lukas, 2003). Namun, bila penekanannya pada data respons hasil

pengukuran, dimensionalitas dipandang sebagai analisis terhadap data respons peserta tes pada sekumpulan aitem (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, Yu, & Zhang, 2011; Reckase, 2009; Zhang, 2008). Penelitian ini mengacu pada kedua sudut pandang ini.

Tes yang menjadi objek dalam penelitian ini adalah tes Potensi Akademik (PA) dalam Seleksi Penerimaan Mahasiswa Baru – Perguruan Tinggi Agama Islam Negeri (SPMB – PTAIN) tahun 2012. Di satu sisi, dimensionalitas dalam penelitian ini ditujukan untuk mengetahui struktur empirik dimensi PA. Data respons dieksplorasi dimensionalitasnya. Di sisi lain, konfirmasi dilakukan dengan mempertimbangkan domain dan komponen pengembangannya. PA didesain dikembangkan oleh tim perancang dengan tujuan ukur unidimensi pada tiap domain Verbal dan Kuantitatif. Domain verbal terdiri dari tiga komponen; yaitu (1) analogis, (2) logis, dan (3) analitis, sementara itu domain kuantitatif terdiri komponen: (1) aritmetika, (2) komparasi, dan (3) geometri. Berdasarkan informasi ini, PA secara keseluruhan telah didesain bersifat multidimensi (dimensi verbal dan kuantitatif).

Meskipun pengertian dimensionalitas dapat dilihat dari sudut pandang yang berbeda, eksplorasi ataupun konfirmasi struktur dimensi merupakan bagian dari proses validasi yang bersifat komprehensif (Jang & Roussos, 2007). Lebih jauh lagi, struktur dimensi antar kelompok juga perlu ditegakkan, bersifat invarians atau tidak, sehingga sejalan dengan prinsip pengukuran yang idealnya bersifat invarians antar kelompok.

Struktur dimensi sebuah tes mengacu pada jumlah dimensi dan karakteristik aitem terkait dengan atribut laten peserta tes (Zhang, 2012b). Dapat dikatakan pula bahwa struktur dimensi menunjukkan hubungan antara aitem-aitem dalam

sebuah tes dan atribut laten yang diyakini diukur oleh tes tersebut. Melalui serangkaian prosedur eksploratori dan konfirmatori, dapat diidentifikasi bagaimanakah hubungan antara aitem-aitem dalam tes dengan dimensi-dimensi yang diukur oleh tes. Dengan kata lain, diantara dimensi-dimensi yang diukur oleh tes, tiap-tiap aitem mengukur dimensi mana saja, hal ini perlu diselidiki lebih jauh.

Multidimensionalitas pada level tes akan terkait dengan struktur konstrak, bagaimanakah struktur dimensi-dimensi dalam mendukung konstrak yang dimaksud. Pada level yang lebih detil, bagaimanakah aitem-aitem dalam mengungkap dimensi. Apakah tiap aitem hanya mengungkap satu dimensi, atau lebih dari satu. Bila masing-masing aitem mengungkap hanya satu dimensi saja, maka struktur semacam ini disebut sebagai multidimensional sederhana. Apabila sebuah aitem mengungkap lebih dari satu dimensi, kondisi semacam ini disebut multidimensional kompleks.

Pada sebuah tes yang bersifat multidimensional, struktur sederhana ataupun kompleks menunjukkan struktur internal tes. Apabila struktur internal tes dibandingkan diantara dua kelompok atau lebih, akan dapat diuji tingkat invariansi strukturnya. Saat kesetaraan struktur tercapai diantara dua kelompok, terjadilah invariansi struktur. Invariansi semacam ini, pada level aitem menunjukkan bebasnya aitem dari DIF.

Penyelidikan DIF diperlukan guna memastikan aitem-aitem dalam tes bersifat tidak bias dan mampu menggambarkan kesetaraan konstrak untuk sembarang peserta tes (Walker, 2011). Penyelidikan ini umumnya dilakukan saat melakukan analisis aitem pada pengembangan alat ukur, adaptasi tes pada kultur yang berbeda, dan secara umum pada proses validasi dalam menyimpulkan skor

tes (Zumbo, 2007). Dengan demikian dapat dikatakan bahwa penyelidikan DIF adalah kegiatan rutin dalam proses validasi skor hasil pengukuran.

Penyelidikan DIF seyogianya terkait dengan kinerja aitem-aitem dalam tes pada kelompok-kelompok yang penting. Kegiatan ini merupakan tahap yang esensial dalam validasi tes (Hambleton, 2006; Walker & Beretvas, 2001), apalagi dalam konteks tes yang berisiko tinggi (Borsboom, Mellenbergh, & Van Heerden, 2002). Tantangannya adalah bagaimana tes menghasilkan skor pengukuran yang presisi sepanjang skala pengukuran, komparabel, dan adil (Liu, Harris, & Schmidt, 2007). Bila pada level aitem atau tes bersifat bias, penarikan kesimpulan berupa estimasi kemampuan para peserta tes menjadi bias pula. Oleh sebab itulah, di Amerika Serikat, analisis DIF sudah menjadi standar pada hampir semua pengembangan tes yang dilaksanakan secara luas (Furlow, Ross, & Gagné, 2009).

Lain halnya dengan di Amerika Serikat, analisis DIF sebagai bagian dari tahapan dalam pengembangan tes belum menjadi suatu kewajiban. Hal ini disebabkan oleh belum adanya regulasi yang mengikat dari pihak pemerintah ataupun himpunan profesional bahwa aitem-aitem dalam tes harus terbebas dari DIF. Selain itu, pengembangan tes khususnya PA SPMB dilakukan oleh sebuah kepanitiaan yang berdiri sendiri serta tidak terintegrasi dari tahun ke tahun.

Lebih jauh, deteksi *differential item functioning* (DIF) terkait dengan isu keadilan dalam tes, keadilan fungsi aitem antar kelompok (Angoff, 1993; Stone, Cook, dkk., 2010; Zumbo, 1999). Bahkan, secara khusus, Liu, dkk. (2007) menegaskan bahwa analisis DIF merupakan prosedur yang harus dilalui pada saat mengembangkan tes masuk perguruan tinggi. Pentingnya aitem agar terbebas dari DIF dibuktikan oleh adanya upaya *Educational Testing Service* untuk merevisi

aitem-aitem SAT-Verbal (Curley & Schmitt, 1993). Pada tes yang diasumsikan unidimensi, masuknya aitem-aitem yang bersifat multidimensi dapat menjadikannya berfungsi secara berbeda (DIF) antar kelompok (Ackerman, 1991; Furlow dkk., 2009). Apabila aitem-aitem tes terbukti multidimensional, kemudian distribusi atribut laten pada dimensi tambahan ini berbeda antar kelompok, diperlukan suatu penyelidikan tentang efek multidimensionalitas yang mendasari peserta tes dalam menjawab benar sebuah aitem terkait mengandung DIF tidaknya aitem tersebut.

Uraian-uraian sebelumnya mengarahkan pada pemahaman bahwa adanya DIF pada sebuah aitem dipandang oleh model UIRT disebabkan oleh adanya atribut laten tambahan, diluar atribut yang menjadi tujuan ukur, yang ikut mempengaruhi kinerja peserta dalam menjawab benar. Berangkat dari pengertian ini, model MIRT memberikan pengertian yang sejalan. Bedanya ada pada banyaknya atribut laten yang dimodelkan.

Kenyataan DIF pada sebuah aitem pada model UIRT memiliki makna bahwa sebuah dimensi saja belum mampu menjelaskan dengan baik kinerja para peserta tes dalam menjawab benar. Pada model MIRT, katakanlah MIRT dua dimensi, berarti dua dimensi saja belum mampu menjelaskan dengan baik kinerja para peserta tes dalam menjawab benar. Dua pengertian ini mengarahkan pada pemahaman bahwa DIF terjadi karena banyaknya dimensi dalam model belum mampu menjelaskan dengan baik kinerja para peserta tes dalam menjawab benar. Dalam bahasa yang lebih ringkas, DIF terjadi disebabkan oleh adanya dimensi tambahan yang tidak bisa dijelaskan oleh model.

Perhatian tentang pentingnya aitem dan tes terbebas dari bias ini sudah lama digulirkan (misalnya Rudner, Getson, & Knight, 1980). Ide tersebut sulit

diimplementasikan karena persoalan teknis teknologi komputasi. Dengan berkembangnya sistem komputasi melalui komputer, teknik implementasi deteksi bias menjadi lebih mudah. Prosedur-prosedur penyelidikan tentang DIF yang berkembang selama ini, didesain untuk mendeteksi *differential item validity* (Camilli & Shepard, 1994), yaitu validitas aitem yang berbeda antar kelompok yang menjadi perhatian.

Berdasarkan uraian sebelumnya, peneliti berpendapat bahwa penting kiranya untuk menyelidiki secara lebih jauh efek dari dimensi tambahan yang terkait dengan keanggotaan peserta tes pada kelompok tertentu (dalam penelitian ini adalah jenis kelamin dan sekolah asal peserta) dapat menjelaskan mengapa terjadi DIF pada aitem. Penelitian ini diarahkan untuk menyelidiki apakah DIF dalam kerangka UIRT adalah manifestasi dari multidimensionalitas dalam sebuah aitem.

3. DIF Berdasarkan Gender dan Sekolah Asal

Pada seleksi masuk perguruan tinggi khususnya di universitas-universitas yang bernaung dibawah Kementerian Agama Republik Indonesia, peserta tes masuk dapat dikategorikan berdasarkan sekolah asal, yaitu peserta lulusan Madrasah Aliyah Negeri (MAN) dan peserta lulusan Sekolah Menengah Atas Negeri (SMAN). Selain itu, para peserta bisa pula dikategorikan berdasarkan gender, yaitu peserta perempuan (P) dan peserta lelaki (L). Dengan demikian, deteksi DIF pada penelitian ini didasarkan pengelompokan peserta berdasarkan kategori gender (PL) dan kategori asal sekolah (MS). Sebagaimana diungkapkan oleh Azwar (2008a), PA didesain memiliki keterkaitan seminimal mungkin dengan silabus/kurikulum di sekolah sehingga harapannya adalah latar belakang asal sekolah peserta tes tidak berpengaruh terhadap kinerja aitem-aitem tes.

Pemilihan kategori PL dan MS dalam penelitian ini bukanlah tujuan utama penelitian ini. Permasalahan utama yang hendak dijawab oleh penelitian ini adalah efek dari multidimensionalitas pada DIF aitem-aitem yang diasumsikan unidimensi. Sebagaimana dalam hasil penelitian awal pada saat mengembangkan proposal disertasi ini (Ridho, 2011), ditemukan bahwa tes Potensi Akademik (PA) terbukti bersifat multidimensi. Multidimensionalitas PA terbukti pula berdasarkan penelitian Azwar dan Ridho (2012).

Penelitian DIF akan selalu dikaitkan dengan konteks variabel kelompok yang dibandingkan. Aplikasi eksplorasi DIF dalam penelitian ini diterapkan pada pengelompokan berdasarkan gender dan sekolah asal. Berdasarkan gender, peserta tes dikelompokkan menjadi perempuan – lelaki (PL), sedangkan berdasarkan asal sekolah, peserta tes dikelompokkan menjadi peserta yang berasal dari lulusan MAN – SMAN (MS). Untuk tema DIF berdasarkan PL, sudah banyak penelitian yang terkait. Namun, pada deteksi DIF berdasarkan MS, pengelompokan ini sejauh peneliti ketahui belum pernah ditemukan.

Sejumlah penelitian mengenai DIF berdasarkan kelompok gender telah dilakukan pada berbagai konteks testing melalui berbagai metode. Penelitian-penelitian tersebut pada umumnya melaporkan bahwa lelaki cenderung memiliki kinerja yang lebih baik pada aitem-aitem yang berisi materi ilmu-ilmu alam dan teknik, sementara perempuan menunjukkan kinerja yang lebih baik dalam bidang ilmu-ilmu sosial, seni, dan humaniora (Curley & Schmitt, 1993; Harris & Carlton, 1993; Pae, 2012; Scheuneman & Gerritz, 1990).

Pada *Graduate Management Admission Test* (GMAT) Scheuneman dan Gerritz (1990) menemukan bahwa aitem-aitem yang berisikan materi ilmu-ilmu alam menguntungkan lelaki, sementara aitem-aitem yang berisikan materi ilmu-

ilmu sosial dan humaniora cenderung menguntungkan perempuan. Pada aitem-aitem verbal dalam SAT mengalami hal yang serupa. Dilaporkan bahwa aitem-aitem *reading comprehension* dalam konteks ilmu-ilmu alam bersifat lebih sulit bagi perempuan (Curley & Schmitt, 1993; Harris & Carlton, 1993).

Young (1991) menemukan bahwa rata-rata skor *Scholastic Aptitude Test* (SAT) –yang sekarang ini berubah nama menjadi *Scholastic Assessment Test* I (SAT I)– berbeda pada kelompok peserta lelaki dan perempuan. Kenyataan ini diperkuat oleh kajian Nankervis (2011) yang mengatakan bahwa berdasarkan hasil tes masuk perguruan tinggi, lelaki secara signifikan memiliki skor *Scholastic Assessment Test* I (SAT I) yang lebih tinggi dibanding perempuan. Dengan demikian terdapat konsistensi perbedaan rata-rata skor antara lelaki dan perempuan dalam SAT. Perbedaan semacam ini memicu para peminat psikometri untuk menyelidiki apakah perbedaan yang terjadi tersebut lantaran memang mencerminkan perbedaan yang sesungguhnya, atau hal tersebut karena adanya aitem-aitem yang mengandung DIF.

Gallagher, Levin, dan Cahalan (2002) meneliti perbedaan gender dalam hal *cognitive processing* pada aitem-aitem kuantitatif *Graduate Record Examination* (GRE). Hasil penelitian mereka mengindikasikan bahkan faktor kognitif yang mempengaruhi perbedaan kinerja berdasarkan gender adalah konteks aitem, berbagai pola menemukan jawaban benar, dan cara cepat menyelesaikan soal. Sejalan dengan Gallagher dkk., Kalaycioğlu dan Berberoğlu (2011) menemukan bahwa aitem-aitem pada bagian kuantitatif dari *university entrance examination* (UEE) Turki, terdeteksi mengandung DIF. Materi yang mengandung *higher order cognitive skills* dan *figural* atau representasi yang bersifat grafis merupakan sumber utama pemicu DIF yang menguntungkan lelaki.

Sebagai bagian dari tes masuk perguruan tinggi di Korea, bagian *English listening* dari *Korea College Scholastic Ability Test* (KCSAT) terbukti mengandung DIF (Park, 2008). Aitem-aitem yang mengandung materi belanja dan hiburan cenderung menguntungkan perempuan, sementara aitem-aitem yang mengandung materi olahraga dan *traveling* menguntungkan lelaki.

Penelitian tentang DIF pada dasarnya dipicu oleh isu tentang gap atau ketimpangan antara lelaki dan perempuan dalam hal kemampuan matematika yang masih terus mengemuka. Perbandingan lelaki dan perempuan di berbagai negara menunjukkan bahwa dalam hal kemampuan matematika, misalnya, pada sebagian negara terbukti terdapat celah perbedaan yang nyata. Hasil meta analisis yang dilakukan oleh Else-Quest, Hyde, dan Linn (2010) berdasarkan data *Trends in International Mathematics and Science Study* (TIMSS) dan *Program for International Student Assessment* (PISA) menunjukkan bahwa terdapat perbedaan kemampuan matematika antara lelaki dan perempuan dengan ukuran efek d = 0,15 yang menguntungkan lelaki.

Berdasarkan pada uraian-uraian sebelumnya, dapat dikatakan bahwa penyelidikan tentang DIF pada alat ukur yang diadministrasikan secara luas telah dilakukan di berbagai negara, berdasarkan kelompok lelaki dan perempuan. Kegiatan penelitian DIF tersebut ditujukan guna menjamin kesetaraan konstrak pengukuran pada kedua kelompok. Pengupayaan kesetaraan konstrak melalui perbaikan aitem-aitem yang terdeteksi DIF adalah hal yang krusial demi menjamin validitas konstrak pengukuran. Pentingnya deteksi DIF dibuktikan dengan dimasukkannya deteksi DIF sebagai bagian dari standar pengembangan tes oleh organisasi-organisasi testing semisal *American College Testing* (ACT), *Educational Testing Service* (ETS), *College Board* (CB), *Illinois Alternate*

Assessment (IAA), dan *Philippine Aptitude Classification Test* (PACT) (ACT, 2011; Augemberg & Morgan, 2008; ETS, 2002; Fraillon, Schulz, & Ainley, 2010; IAA, 2011; Lantano & Gatchalian, 2010; Yang, 2004).

Bila di negara maju, Amerika misalnya, perhatian tentang testing begitu besar, di Indonesia belum sampai pada taraf standar sebagaimana di negara-negara maju. Di Amerika kajian DIF adalah kegiatan rutin dalam pengembangan tes, dilakukan sebelum tes tersebut diadministrasikan secara luas. Dengan demikian, tes dalam skala luas yang diadministrasikan sudah relatif bebas DIF. Di Indonesia, tradisi semacam itu belum membudaya. Kajian terhadap DIF dilakukan pasca administrasi dilaksanakan secara luas. Hal ini disebabkan oleh pengembangan tes yang secara umum dibentuk dalam kepanitiaan yang tidak berkesinambungan dari waktu ke waktu.

Pada tes dalam skala besar, khususnya tes masuk perguruan tinggi, misalnya, peneliti belum temukan kajian yang khusus tentang deteksi DIF pada aitem-aitem tes. Oleh sebab itu, kajian dan penelitian tentang DIF pada tes masuk perguruan tinggi merupakan kegiatan yang krusial untuk dilaksanakan. Apalagi mengingat tes masuk perguruan tinggi merupakan tes yang berisiko tinggi. Penelitian ini akan fokus pada aitem-aitem yang digunakan dalam tes masuk perguruan tinggi, khususnya pada mata uji Potensi Akademik (PA). Kelompok gender (lelaki dan perempuan) dipilih pada penelitian ini dengan harapan dapat memberikan kontribusi pada upaya-upaya kesetaraan gender dalam testing pada ujian masuk perguruan tinggi.

Berbeda dengan deteksi DIF berdasarkan gender yang telah banyak diteliti, tidak demikian dengan deteksi DIF berdasarkan asal sekolah, dalam hal ini adalah

MAN – SMAN. Pilihan terhadap pengelompokan ini dilandaskan pada argumentasi berikut.

Pada umumnya tes PA dirancang untuk mengungkap kemampuan kognitif potensial, disusun berdasar konsep kemampuan dasar, memiliki kaitan minimal dengan kurikulum (Azwar, 2008a). Pada SPMB-PTAIN 2012, PA dirancang terdiri dari enam komponen; yaitu (1) analogis, (2) logis, (3) analitis, (4) aritmetika, (5) komparasi, dan (6) geometri. Oleh sebab itu, idealnya aitem-aitem tidak terpengaruh oleh latar belakang pendidikan peserta. Pertanyaannya kemudian adalah, sejauh mana tim pengembang dan penulis aitem PA mampu menghasilkan aitem-aitem yang memiliki kinerja setara antara dua kelompok yang memiliki latar belakang pendidikan yang berbeda, hal ini perlu diidentifikasi.

Terdapat kesenjangan yang terjadi antara SMA dan MA. Berdasarkan informasi dari Badan Akreditasi Nasional Sekolah / Madrasah (BAN – S/M) yang diakses melalui http://www.ban-sm.or.id/statistik pada 12 Desember 2012, diperoleh data bahwa terdapat 13678 SMA dan 5014 MA. Hasil akreditasi menunjukkan bahwa secara keseluruhan di Indonesia, perbedaan mencolok terjadi pada peringkat akreditasi A, yaitu 36,52% pada SMA dan 18,71% pada MA. Hasil akreditasi ini adalah cerminan dari kenyataan bahwa memang terdapat perbedaan antara SMA dan MA. Perbedaan ini tentu akan berimplikasi pada kualitas lulusannya. Lebih jauh, perbedaan antara SMA dan MA boleh jadi berimplikasi pada perbedaan kinerja lulusannya pada saat mengikut seleksi masuk perguruan tinggi, termasuk SPMB-PTAIN.

Selain itu, di MA terdapat pelajaran Fiqh, Aqidah Akhlak, Al Quran Hadits, Bahasa Arab, sementara di SMA tidak terdapat mata pelajaran tersebut. Dengan alokasi yang waktu yang sama antara MA dan SMA, sedangkan terdapat beban

mata pelajaran yang lebih banyak, hal ini akan berdampak pada pengalaman belajar yang tidak sama antara siswa-siswa MA dan SMA.

Kesenjangan antara SMA dan MA tampak pula pada kasus terakhir dimana hanya hanya lima lulusan MA se-Indonesia yang diterima menjadi mahasiswa Universitas Gadjah Mada (UGM) melalui jalur Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) tahun 2013 ini. Bahkan lulusan MA di DIY tak satupun yang diterima melalui jalur ini (sumber: http://www.republika.co.id/berita/pendidikan/eduaction/13/05/29/mnjthu-tak-satu-pun-lulusannya-diterima-ugm-madrasah-aliyah-seyogyakarta-kecewa). Mengacu pada keterangan Kepala Dinas Pendidikan Pemuda dan Olahraga (Dikpora) Daerah Istimewa Yogyakarta Baskara Aji, dibandingkan dengan SMA, MA memang menjadi pilihan kedua dan beban pelajarannya lebih banyak (sumber: http://daerah.sindonews.com/read/2013/05/31/22/744815/kadisdikpora-diy-kualitas-madrasah-aliyah-dibawah-sma).

Bila perbedaan kinerja pada aitem-aitem tes disebabkan oleh atribut laten yang memang berbeda, hal tersebut bukanlah suatu masalah. Namun, bila perbedaan tersebut terjadi karena kelompok asal sekolah yang ikut berkontribusi saat menyelesaikan aitem-aitem soal pada tes seleksi, berarti telah terjadi DIF pada aitem-aitem tes seleksi.

B. Rumusan Permasalahan

Terdapat beberapa hal yang dipertimbangkan dalam arah penelitian ini. Pertimbangan I, terkait sudut pandang paradigma teori pengukuran. Teori pengukuran yang berkembang selama ini paling tidak dapat dibagi dua yaitu *classical test theory* (CTT) dan *item response theory* (IRT). Meskipun CTT adalah teori yang muncul pada tahun 1950-an (Embretson & Reise, 2000; Hambleton

dkk., 1991), keberadaannya tetap eksis sampai sekarang. Hal ini karena CTT relatif lebih mudah dipelajari dibanding IRT, baik dari sisi teoritik ataupun praktis. Namun demikian, mengingat kelebihan IRT dibanding CTT sebagaimana disinggung di awal bagian Latar Belakang, penelitian ini akan menggunakan kerangka *item response theory* (IRT) dan *multidimensional item response theory* (MIRT).

Pertimbangan II terkait dengan dimensi tes. Tes yang pada kenyataannya bersifat multidimensi, bila diperlakukan penskoran secara unidimensi, hasil estimasi parameternya akan bias. Akibatnya skor yang dijadikan dasar memaknai atribut laten yang diungkap menjadi tidak tepat. Oleh sebab itu semestinya penskoran dilakukan setelah dimensionalitas respons hasil pengukuran berdasarkan tes tersebut diselidiki, unidimensi atau multidimensi. Bukti unidimensionalitas ini ditentukan setelah dilakukan uji dimensionalitas. Dalam penelitian ini, untuk menyelidiki dimensi yang melatarbelakangi peserta dalam menjawab benar aitem-aitem, peneliti menggunakan kerangka unidimensionalitas berarti terpenuhinya asumsi independensi lokal (Stout, 1984, 1989, 2002; Swaminathan, Hambleton, & Rogers, 2007). Uji independensi lokal aitem-aitem dilakukan dengan bantuan software DETECT (Zhang & Stout, 1999) dan DIMTEST (Stout & Nandakumar, 2006).

Pertimbangan III terkait dengan model multidimensi dalam IRT. Paling tidak terdapat dua model: model kompensasi (*compensatory*) dan model nonkompensasi (*noncompensatory*) (Reckase, 2009). Model kompensasi didasarkan pada hubungan kombinasi linier koordinat vektor atribut laten, $\boldsymbol{\theta}$. Sementara itu model nonkompensasi memisahkan atribut-atribut laten dalam merespons aitem serta menggunakan model UIRT untuk tiap atribut laten. Model

MIRT yang akan digunakan dalam penelitian ini adalah model kompensasi (*compensatory*) (Reckase & Ackerman, 1986; Reckase, Ackerman, & Carlson, 1988). Untuk menyelidiki DIF tidaknya aitem-aitem dalam kerangka MIRT, peneliti menggunakan metode *Lord's chi-square* (LCS) dalam kerangka multidimensi. Guna mendeteksi DIF multidimensional, peneliti menggunakan bantuan software BMIRT yang dikembangkan oleh Yao (2010a).

Isu tentang multidimensionalitas atribut laten yang mendasari peserta tes menjawab benar aitem-aitem tes telah lama digulirkan (misalnya Ackerman, 1988; Flowers & Oshima, 1994; McDonald, 1967; Reckase, 1985; Reckase & Ackerman, 1986). Berbagai metode telah dikembangkan oleh para ahli psikometri untuk mengidentifikasi dimensionalitas aitem-aitem dalam alat ukur. Software yang digunakan juga berkembang. Sebagai contoh adalah TESTFACT (Bock dkk., 2003), NOHARM (Fraser & McDonald, 1999, 2003), DIMTEST (Stout & Nandakumar, 2006), dan BMIRT (Yao, 2010a; Yao & Boughton, 2007). Diantara software yang ada, pemilihan software sebagai alat bantu merupakan permasalahan tersendiri sebagai pertimbangan IV.

Pertimbangan terakhir didasarkan pada pemilihan kelompok referensi (R) dan fokal (F) yang dipilih. Penelitian ini akan terbatas pada pengelompokan peserta berdasarkan gender dan sekolah asal. Hal ini didasarkan pada pertimbangan: (1) penekanan utama penelitian ini adalah hendak mengetahui efek multidimensionalitas aitem pada DIF UIRT; (2) sumber data yang melekat secara langsung pada lembar jawab SPMB-PTAIN tahun 2012 adalah jenis kelamin dan sekolah asal.

Memperhatikan uraian yang telah dikemukakan sebelumnya, peneliti merumuskan masalah utama yang diangkat dalam penelitian ini adalah:

1) Aitem-aitem PA SPMB-PTAIN tahun 2012 mana saja yang mengandung DIF berdasarkan kelompok gender dan jenis sekolah asal?
2) Sejauh manakah multidimensionalitas memberikan kontribusi terhadap DIF pada aitem-aitem yang diasumsikan unidimensi?
3) Karakteristik psikometrik seperti apakah yang memicu munculnya DIF MIRT?

C. Tujuan dan Manfaat

Tujuan penelitian ini adalah untuk:

1) Mengevaluasi aitem-aitem PA SPMB-PTAIN tahun 2012 yang mengandung DIF berdasarkan kelompok gender (perempuan – lelaki) dan sekolah (MAN – SMAN);
2) Mengevaluasi sejauh mana efek multidimensionalitas pada DIF UIRT.
3) Mengevaluasi karakteristik psikometrik secara umum pemicu DIF MIRT pada tes PA SPMB-PTAIN 2012.

Penelitian ini memberikan manfaat:

1) Secara teoritik:
    a. menambah kajian tentang dimensionalitas dalam tes;
    b. memahami struktur konstrak PA secara empirik;
    c. penguatan eksistensi model *multidimensional item response theory* (MIRT); dan
    d. mengetahui efek multidimensionalitas pada DIF yang dimodelkan unidimensi.

2) Secara praktis:

   a. memberikan informasi psikometrik tentang PA SPMB-PTAIN, berdasarkan UIRT dan MIRT;

   b. identifikasi aitem-aitem PA SPMB-PTAIN mana saja yang mengandung DIF berdasarkan kelompok gender dari sudut pandang kerangka UIRT dan MIRT. Informasi ini dapat digunakan untuk memperbaiki jenis-jenis aitem sehingga bisa terbebas dari bias;

   c. memberikan usulan metode skoring pada tes multidimensi.

D. Keaslian

Di Indonesia, masih sedikit penelitian tentang tes yang menggunakan sudut pandang teori respons aitem atau *item response theory* (IRT). Sementara teori tes berkembang begitu pesat. IRT yang pada mulanya didasarkan pada asumsi unidimensi, mengalami kendala dalam melakukan penskoran pada tes-tes yang bersifat multidimensi (misalnya Ackerman, 1989; Cheng, Wang, & Ho, 2009; DeMars, 2006; Dirir & Sinclair, 1996; Oshima & Miller, 1990; Reise, Moore, & Haviland, 2010; Yao, 2011). Dengan demikian, tes multidimensi yang diskor berdasarkan paradigma unidimensi akan mengalami ketidaktepatan. Menurut (Snow & Oshima, 2009), penelitian tentang DIF sejauh ini lebih banyak disandarkan pada kerangka UIRT. Masih sedikit yang mengeksplorasi DIF dalam kerangka MIRT sebagaimana dalam penelitian ini. Dalam rentang waktu tahun 2000-2012 penelitian DIF dalam kerangka MIRT yang peneliti temukan adalah:

1) Yao dan Li (2010) mengembangkan prosedur deteksi DIF dalam kerangka MIRT dengan metode bayesian. Metode tersebut diaplikasikan pada data riil dan simulasi.
2) Snow dan Oshima (2009) mendeteksi DIF berbasis model satu dan tiga dimensi pada data simulasi dua dimensi. Data dua dimensi terdeteksi DIF tidak seragam berdasarkan model satu dimensi.
3) Walker, Zhang, dan Surber (2008) mendeteksi DIF berdasarkan etnis pada aitem-aitem matematika dalam instrumen *Terra Nova* menggunakan model dua dimensi MIRT dengan bantuan NOHARM. Hasil penelitiannya menunjukkan bahwa kinerja siswa pada aitem-aitem tertentu ikut ditentukan oleh level kemampuan membacanya.
4) Wagiran dan Retnawati (2008, 2009) membandingkan beberapa metode DIF dalam kerangka MIRT. Hasil penelitiannya menyimpulkan bahwa metode *likelihood ratio test* (LRT) paling sensitif dibanding metode lainnya. Secara aplikatif, deteksi DIF dilakukan berdasarkan data respons Ujian Nasional (UN).
5) Mendes-Barnett dan Ercikan (2006) mengidentifikasi sumber-sumber DIF berdasarkan kelompok gender (5069 lelaki dan 4335 perempuan) pada *British Columbia Principles of Mathematics Exam*.
6) Stoneberg Jr. (2004) mendeteksi DIF *Idaho Standards Achievement Tests* berdasarkan kelompok gender dan etnis. Teknik yang ia gunakan adalah *Simultaneous Item Bias Test* (SIBTEST).
7) Stout dkk. (2003) mengembangkan metodologi deteksi DIF yang terjadi pada kelompok gender dan warna kulit dengan kemampuan sama dan

tidak sama. Data yang digunakan adalah data respons Subtes Kuantitatif *Graduate Record Examinations* (GRE).

8) Walker dan Beretvas (2001) meneliti DIF berdasarkan kelompok siswa sekolah dasar kelas 4 dan 7 yang pandai dan tidak. Aitem-aitem yang diteliti adalah aitem-aitem esai matematika (*Washington Assessment of Student Learning*, WASL). Melalui metode Poly-SIBTEST, terbukti bahwa aitem-aitem terdeteksi DIF disebabkan oleh perbedaan distribusi pada kemampuan komunikasi matematik.

Khusus terkait disertasi pada Program Doktor Psikologi UGM, disertasi yang erat dengan penelitian ini adalah disertasi yang berjudul “Perbandingan Berbagai Metode untuk Mendeteksi Bias Butir” (Wagiran, 2005) dan “Validitas Prediktif Diferensial pada Ujian Tulis UM UGM” (Azwar, 2008b). Dibandingkan dengan dua disertasi ini, terdapat beberapa persamaan dan perbedaan. Persamaan dengan disertasi ini terletak pada isi yang sama-sama mengkaji masalah DIF.

Wagiran (2005) fokus pada perbandingan berbagai metode deteksi DIF dalam kerangka UIRT, sementara dalam disertasi ini kerangka yang digunakan adalah UIRT dan MIRT. Meski tidak sedalam pada disertasi Wagiran (2005), Azwar (2008b) memasukkan kajian tentang DIF sebagai bagian kecil dari disertasinya dalam kerangka UIRT. Selain menggunakan kerangka UIRT, DIF dalam disertasi ini dianalisis pula dalam kerangka MIRT. Dari sisi permasalahan yang diangkat, penelitian ini berupaya membuktikan pengaruh multidimensionalitas pada DIF.

Dengan melihat hasil penelitian terdahulu tersebut, dapat dikatakan bahwa pengembangan ataupun penelitian DIF dalam kerangka MIRT belum mendapatkan perhatian sebanyak kerangka UIRT. Selain itu, aplikasi kerangka

DIF MIRT pada pengukuran PA, sejauh peneliti ketahui, belum pernah dilakukan. Dengan demikian dapat disebutkan bahwa belum ada penelitian yang sama dengan penelitian ini.

BAB VI

KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan uraian dalam hasil dan pembahasan, dapat diambil kesimpulan sebagai berikut.

1. Aitem-aitem PA SPMB-PTAIN tahun 2012 terbukti mengungkap 4 dimensi: kosakata (θ_1), verbal (θ_2), kuantitatif (θ_3), dan simbol (θ_4). Struktur konstrak PA SPMB-PTAIN tahun 2012 bersifat multidimensi semi kompleks: aitem-aitem kosakata – verbal secara bersama-sama mengungkap dimensi kosakata dan verbal; sementara aitem-aitem kuantitatif – simbol sama-sama mengungkap dimensi kuantitatif dan dimensi simbol.
2. Aitem-aitem PA SPMB-PTAIN tahun 2012 yang mengandung DIF UIRT:
   a. Berdasarkan kelompok perempuan – lelaki,
      i. ada 19 aitem menguntungkan lelaki yang tersebar pada komponen analogis, analitis, dan geometri (simbol);
      ii. ada 14 menguntungkan perempuan yang tersebar pada komponen aritmetika dan komparasi;
      iii. berdasarkan ES DIF, 11 aitem memiliki ES sedang dan besar. Ada 6 aitem menguntungkan lelaki pada komponen analogis dan analitis; dan 5 aitem menguntungkan perempuan pada komponen aritmetika dan komparasi.
   b. Berdasarkan kelompok MAN – SMAN,
      i. ada 12 aitem menguntungkan siswa lulusan SMAN pada komponen analogi, logis, aritmetika, dan komparasi;

ii. ada 3 aitem menguntungkan siswa lulusan MAN pada komponen analogi dan aritmetika;

iii. berdasarkan ES DIF, 8 aitem memiliki ES sedang dan besar. Ada 6 aitem menguntungkan SMAN pada komponen analogis, logis, dan aritmetika, dan komparasi; 2 aitem menguntungkan MAN pada komponen analogi dan aritmetika.

3. Aitem-aitem PA SPMB-PTAIN tahun 2012 yang mengandung DIF MIRT:
   a. Berdasarkan kelompok perempuan – lelaki, terdapat 16 aitem DIF dengan rincian:
      i. ada 9 aitem pada komponen analogi, analitik, aritmetika dan geometri yang menguntungkan lelaki; dan
      ii. ada 7 aitem pada komponen aritmetika dan komparasi yang menguntungkan perempuan
   b. Berdasarkan kelompok MAN – SMAN, terdapat 5 aitem DIF pada komponen analogi, logis, dan aritmetika yang semuanya menguntungkan siswa lulusan SMAN.
4. Multidimensionalitas aitem yang berbeda pada kelompok fokal dan referensi tidak berefek pada aitem sehingga berperilaku DIF. Perbedaan multidimensionalitas antara kelompok fokal dan referensi tidak memicu munculnya DIF pada aitem.
5. Aitem-aitem yang terdeteksi DIF MIRT memiliki karakteristik sensitivitas daya beda yang berbeda pada kelompok fokal dan referensi.

B. Saran

Mendasarkan pada kesimpulan yang dihasilkan dalam penelitian ini, dapat diberikan saran sebagai berikut.

1. Untuk tim skoring, metode penskoran PA SPMB-PTAIN hendaknya diubah dari skor total menjadi komposit berdasarkan fungsi informasi maksimum model MIRT sehingga menghasilkan skor-skor: (1) komposit PA, (2) kosakata, (3) verbal, (4) kuantitatif, dan (5) simbol. Selain itu, penskoran hendaknya dilakukan dengan mengeliminir aitem-aitem yang terdeteksi DIF.
2. Untuk tim pengembang dan penulis tes PA SPMB-PTAIN:
   a. perlu melakukan analisis isi secara lebih mendalam terhadap aitem-aitem yang terdeteksi DIF terutama pada aitem-aitem yang memiliki ES DIF sedang dan besar sehingga dapat dilakukan perbaikan-perbaikan pada pengembangan tes PA SPMB-PTAIN di masa mendatang agar meningkatkan validitas konstrak tes PA SPMB-PTAIN, juga menjamin kesetaraan dan keadilan bagi peserta lelaki – perempuan dan MAN – SMAN;
   b. analisis DIF hendaknya dijadikan sebagai salah satu poin prosedur proses pengembangan PA SPMB. Dengan demikian pada saat tes tersebut digunakan secara luas, aitem-aitem sudah relatif bebas DIF;
   c. pengembangan aitem-aitem diupayakan betul-betul mengungkap potensi akademik saja, tidak dipengaruhi oleh latar belakang pendidikan formil.
3. Untuk panitia SPMB, pengembangan tes PA hendaknya diserahkan kepada lembaga / badan profesional yang khusus bekerja dibidang testing sehingga penanganan tes menjadi lebih profesional.

4. Untuk para peneliti psikometri, karena temuan dalam penelitian ini menunjukkan bahwa klaim yang menyebutkan bahwa multidimensionalitas merupakan penyebab aitem mengandung DIF tidak terbukti, perlu diselidiki secara lebih jauh dalam penelitian tersendiri tentang sumber-sumber DIF pada aitem. Selain itu, tema tentang sensitivitas aitem dalam konteks MIRT perlu dieksplorasi lebih mendalam.

DAFTAR PUSTAKA

Abedi, J. (1997). Dimensionality of NAEP Subscale Scores in Mathematics *CSE Technical Report 428*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California.

Ackerman, T. A. (1988). *An Explanation of Differential Item Functioning from A Multidimensional Perpective*. Paper dipresentasikan pada Annual Meeting of the American Educational Research Association, New Orleans, LA.

Ackerman, T. A. (1989). Unidimensional IRT Calibration of Compensatory and Noncompensatory Multidimensional Items. *Applied Psychological Measurement, 13*(2), 113-127.

Ackerman, T. A. (1991). *A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective*. Paper dipresentasikan pada Annual Meeting of the American Educational Research Association, Chicago.

Ackerman, T. A. (1992a). *Assessing Construct Validity Using Multidimensional Item Response Theory*. Paper dipresentasikan pada Annual Meeting of the American Educational Research Association, San Francisco, CA.

Ackerman, T. A. (1992b). A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T. A. (1996). Graphical Representation of Multidimensional Item Response Theory Analyses. *Applied Psychological Measurement, 20*(4), 311-329.

Ackerman, T. A., & Evans, J. A. (1994). The Influence of Conditioning Scores In Performing DIF Analyses. *Applied Psychological Measurement, 18*(4), 329-342. doi: 10.1177/014662169401800404

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational and Psychological Measurement, 22*(3), 37–51.

ACT. (2011). *2009-2010 Fairness Report for the EXPLORE® Tests*. Iowa: American College Testing.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement, 36*(3), 185-198.

Andrich, D., & Hagquist, C. (2012). Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387-416. doi: 10.3102/1076998611411913

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. Dalam W. H. Angoff (Ed.), *Differential Item Functioning* (Edisi ke-1, hh. 3-24). New Jersey: Lawrence Erlbaum Associates.

APA. (2010). *Publication manual of the American Psychological Association* (Edisi ke-6). Washington, DC: American Psychological Association.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36*(5), 258-267.

Augemberg, K. E., & Morgan, D. L. (2008). *Differential Performance of Test Items by Geographical Regions*. Paper dipresentasikan pada annual meeting of the National Council on Measurement in Education, New York.

Azwar, S. (2008a). Kualitas Tes Potensi Akademik Versi 07A. *Jurnal Penelitian dan Evaluasi Pendidikan, 12*(2), 232-250.

Azwar, S. (2008b). *Validitas Prediktif Diferensial pada Ujian Tulis UM UGM*. Yogyakarta: Pascasarjana UGM. Disertasi. Tidak diterbitkan.

Azwar, S., & Ridho, A. (2012). *Abilitas Komposit dalam Tes Potensi*. Fakultas Psikologi Universitas Gadjah Mada.

Baker, F. B. (2001). *The Basics of Item Response Theory*. New York: ERIC Clearinghouse on Assessment and Evaluation.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing Impact, DIF, and DFF in Accommodated Item Scores. *Educational and Psychological Measurement, 72*(5), 754-773. doi: 10.1177/0013164412440998

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4*: Scientific Software International, Inc.

Bolt, D., Wollack, J., & Suh, Y. (2012). Application of a Multidimensional Nested Logit Model to Multiple-Choice Test Items. *Psychometrika, 77*(2), 339-357. doi: 10.1007/s11336-012-9257-5

Bolt, D. M. (2000). A SIBTEST Approach to Testing DIF Hypotheses Using Experimentally Designed Test Items. *Journal of Educational Measurement, 37*(4), 307-327.

Bolt, D. M., & Stout, W. F. (1996). Differential Item Functioning: Its Multidimensional Model and Resulting SIBTEST Detection Procedure. *Behaviormetrika, 23*(1), 67-95.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias. *Applied Psychological Measurement, 26*(4), 433-450.

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the Bifactor Model to Assess the Dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement, 71*(1), 170-185.

Camilli, G. (1992). A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model. *Applied Psychological Measurement, 16*(2), 129-147.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Bias Test Item*. Thousan Oaks, CA: Sage Publication.

Camilli, G., Wang, M.-m., & Fesq, J. (1995). The Effects of Dimensionality on Equating the Law School Admission Test. *Journal of Educational Measurement, 32*(1), 79-96.

Casey, M. B., Nuttall, R. M., & Pezaris, E. (1997). Mediators of Gender Differences in Mathematics College Entrance Test Scores: A Comparison of Spatial Skills with Internalized Beliefs and Anxieties. *Developmental Psychology, 33*(4), 669-680.

Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch Analysis of a Psychological Test With Multiple Subtests: A Statistical Solution for the Bandwidth--Fidelity Dilemma. *Educational and Psychological Measurement, 69*(3), 369-388.

Chou, Y.-T., & Wang, W.-C. (2010). Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. *Educational and Psychological Measurement, 70*(5), 717-731.

Chuderski, A., & Ngcka, E. (2012). The Contribution of Working Memory to Fluid Reasoning: Capacity, Control, or Both? *Journal of Experimental Psychology. Learning, Memory & Cognition, 38*(6), 1689-1710. doi: 10.l037/a0028465

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Edisi ke-2). Hillsdale, New Jersey: Lawrence Erlbaum.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An Assessment of the Dimensionality of Three SAT-Verbal Test Editions. *Journal of Educational Statistics, 13*(1), 19-43.

Curley, W. E., & Schmitt, A. P. (1993). Revising SAT-Verbal Items to Eliminate Differential Item Functioning *College Board Report No. 93-2, ETS RR No. 93-61*. New York: The College Board.

de la Torre, J., & Patz, R. J. (2005). Making the Most of What We Have: A Practical Application of Multidimensional Item Response Theory in Test Scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311.

DeMars, C. E. (2006). *Scoring Subscales Using Multidimensional Item Response Theory Models*. Paper dipresentasikan pada Annual Meeting of the American Psychological Association, Washington, DC.

DeMars, C. E. (2011). An Analytic Comparison of Effect Sizes for Differential Item Functioning. *Applied Measurement in Education, 24*(3), 189-209.

Deng, N., Wells, C. S., & Hambleton, R. K. (2008). *A Confirmatory Factor Analytic Study Examining the Dimensionality of Educational Achievement Tests*. Paper dipresentasikan pada Northeastern Educational Research Association (NERA) Annual Conference, Rocky Hill, Connecticut.

Diones, R., Bejar, I. I., & Chaffin, R. (1996). The dimensionality of responses to SAT analogy items. *ETS Research Report RR-96-1*. Princeton, NJ: Educational Testing Service.

Dirir, M. A., & Sinclair, N. (1996). *On Reporting IRT Ability Scores When the Test Is Not Unidimensional*. Paper dipresentasikan pada Annual Meeting of the National Council on Measurement in Education, New York.

Dorans, N. J., Holland, P. W., & Educational Testing Service, P. N. J. (1992). DIF Detection and Description: Mantel-Haenszel and Standardization.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis. *Psychological Bulletin, 136*(1), 103–127.

Embretson, S. E. (1997). The Factorial Validity of Scores from a Cognitively Designed Test: The Spatial Learning Ability Test *Educational and Psychological Measurement, 67*(1), 99-107.

Embretson, S. E. (2000). Dynamic Cognitive Testing: What Kind of Information is Gained by Measuring Response Time and Modifiability? *Educational and Psychological Measurement, 60*(6), 837-863.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologist*. NJ: Lawrence Erlbaum Associates Inc.

Enright, M. K., Tucker, C. B., & Katz, I. R. (1995). A Cognitive Analysis of Solutions for Verbal, Informal, and Formal-Deductive Reasoning Problems. *RR-95-06; GREB-90-04P*. Princeton, NJ: Educational Testing Service.

ETS. (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

ETS. (2007). Factors That Can Influence Performance on the GRE General Test 2006-2007. *Test Fairness and Score Use.* Retrieved 30 Oktober, 2011, from http://www.ets.org/Media/Tests/GRE/pdf/gre_0809_factors_2006-07.pdf

ETS. (2011). *Practice Book for the Paper-Based GRE® revised General Test*. Princeton, NJ: Educational Testing Service.

Finch, W. H., & French, B. F. (2007). Detection of Crossing Differential Item Functioning. *Educational and Psychological Measurement, 67*(4), 565-582. doi: 10.1177/0013164406296975

Finch, W. H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-Based Statistics for Testing Unidimensionality *Applied Psychological Measurement, 31*(4), 292-307.

Fiske, D. W. (2002). Validity for what? Dalam H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (Edisi ke-1, hh. 169-178). Mahwah, NJ: Lawrence Erlbaum.

Flowers, C. P., & Oshima, T. C. (1994). *The Consistency of DIF/DTF Across Different Tes Administrations: A Multidimensional Perspective*. Paper dipresentasikan pada Annual Meeting of the American Educational Research Association, New Orleans, LA.

Fraillon, J., Schulz, W., & Ainley, J. (2010). *Regional Differential Item Function in the International Civics and Citizenship Education Study.* Paper dipresentasikan pada 4th IEA International Research Conference, Gothenburg, Sweden.

Fraser, B. J., Walberg, H. J., Welch, W. W., & Hattie, J. A. (1987). Syntheses of educational productivity research. *International Journal of Educational Research, 11*(2), 147-252.

Fraser, C., & McDonald, R. P. (1999). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Fraser, C., & McDonald, R. P. (2003). NOHARM: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.

Froelich, A. G., & Habing, B. (2008). Conditional Covariance-Based Subtest Selection for DIMTEST. *Applied Psychological Measurement, 32*(2), 138-155.

Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The Impact of Multidimensionality on the Detection of Differential Bundle Functioning Using Simultaneous Item Bias Test. *Applied Psychological Measurement, 33*(6), 441-464.

Gallagher, A. M., Levin, J., & Cahalan, C. (2002). Cognitive Patterns of Gender Differences on Mathematics Admissions Tests. *GRE Board Report No. 96-17P*. Princeton, NJ: Educational Testing Services.

Gessaroli, M. E., & Champlain, A. F. D. (1996). Using an Approximate Chi-Square Statistic to Test the Number of Dimensions Underlying the Responses to a Set of Items. *Journal of Educational Measurement, 33*(2), 157-179.

Gierl, M. J. (2005). Using Dimensionality-Based DIF Analyses to Identify and Interpret Constructs That Elicit Group Differences. *Educational Measurement: Issues and Practice, 24*(1), 3–14.

Gierl, 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.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006). Evaluating DETECT Classification Accuracy and Consistency When Data Display Complex Structure. *Journal of Educational Measurement, 43*(3), 265-289.

Gierl, M. J., Tan, X., & Wang, C. (2005). Identifying Content and Cognitive Dimensions on the SAT®. *College Board Research Report No. 2005-11*. New York: The College Board.

Glanville, J. L., & Wildhagen, T. (2007). The Measurement of School Engagement: Assessing Dimensionality and Measurement Invariance across Race and Ethnicity. *Educational and Psychological Measurement, 67*(6), 1019-1041.

González-Garrido, A. A., Gómez-Velázquez, F. R., Sequeira, H., Ramos-Loyo, J., & López-Franco, A. L. (2013). Gender Differences in Visuospatial Working Memory -- Does Emotion Matter? *International Journal of Psychological Studies, 5*(1), 11-21. doi: 10.5539/ijps.v5n1p11

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. Dalam H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (Edisi ke-1, hh. 621-642). Washington, DC, US: American Psychological Association.

Haberman, S. J., & Sinharay, S. (2010). Reporting of Subscores Using Multidimensional Item Response Theory. *Psychometrika, 74*(2), 209-227.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (Edisi ke-3). Mahwah, NJ: Lawrence Erlbaum.

Halpern, D. F. (2004). A Cognitive-Process Taxonomy for Sex Differences in Cognitive Abilities. *Current Directions in Psychological Science, 13*(4), 135-139.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest*(1), 1. doi: 10.2307/40062381

Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. *Medical Care, 44*(11), S182-S188.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement, 10*(3), 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Application*. Boston, MA: Kluwer Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publication Inc.

Hamilton, L. S., & Snow, R. E. (1998). Exploring Differential Item Functioning on Science Achievement Tests. Los Angeles: Center for the Study of Evaluation.

Harris, A. M., & Carlton, S. T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151. doi: 10.1207/s15324818ame0602_3

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57–63.

Harvey, R. J., & Hammer, A. L. (1999). Item Response Theory. *The Counseling Psychologist, 27*(3), 353-383.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An Assessment of Stout's Index of Essential Unidimensionality. *Applied Psychological Measurement, 20*(1), 1-14.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential Item Functioning Detection and Effect Size: A Comparison between Logistic Regression and Mantel-Haenszel Procedures. *Educational and Psychological Measurement, 64*(6), 903-915. doi: 10.1177/0013164403261769

Hirnstein, M., Freund, N., & Hausmann, M. (2012). Gender Stereotyping Enhances Verbal Fluency Performance in Men (and Women). *Zeitschrift für Psychologie, 220*(2), 70-77.

IAA. (2011). *Illinois Alternate Assessment 2011 Technical Manual*. Illinois: Illinois State Board of Education.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: an illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement, 68*(4), 695-709.

Jang, E. E., & Roussos, L. (2009). Integrative Analytic Approach to Detecting and Interpreting L2 Vocabulary DIF. *International Journal of Testing, 9*(3), 238-259. doi: 10.1080/15305050903107022

Jang, E. E., & Roussos, L. A. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement, 44*(1), 1–21.

Kahraman, N., & Thompson, T. (2011). Relating Unidimensional IRT Parameters to a Multidimensional Response Space: A Review of Two Alternative Projection IRT Models for Scoring Subscales. *Journal of Educational Measurement, 48*(2), 146-164.

Kalaycioğlu, D. B., & Berberoğlu, G. (2011). Differential Item Functioning Analysis of the Science and Mathematics Items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478. doi: 10.1177/0734282910391623

Kim, S.-H., & Cohen, A. S. (1995). A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education, 8*(4), 291-312.

Lantano, A., & Gatchalian, C. (2010). *Revisiting the Philippine Aptitude Classification Test: Analysis of Potentially Biased Test Items*. Paper dipresentasikan pada 36th International Association for Educational Assessment Annual Conference, Bangkok.

Lawrence, I. M., & Dorans, N. J. (1987). *An Assessment of the Dimensionality of SAT-Mathematical*. Paper dipresentasikan pada Annual Meeting of the National Council on Measurement in Education, Wahington, DC.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior Predictive Model Checking for Multidimensionality in Item Response Theory. *Applied Psychological Measurement, 33*(7), 519-537.

Liu, J., Harris, D. J., & Schmidt, A. (2007). Statistical Procedures Used in College Admissions Testing. Dalam C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (Edisi ke-1, hh. 1057-1091). Amsterdam: Elseveir.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Martín, E. S., Pino, G. d., & Boeck, P. D. (2006). IRT Models for Ability-Based Guessing. *Applied Psychological Measurement, 30*(3), 183–203.

McDonald, R. P. (1967). Nonlinear Factor Analysis. *Psychometric Monograph No. 15.* Retrieved 1 Mei, 2011, from http://www.psychometrika.org/journal/online/MN15.pdf

McDonald, R. P. (2000). A Basis for Multidimensional Item Response Theory. *Applied Psychological Measurement, 24*(2), 99–114.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728-743.

Meade, A. W., & Wright, N. A. (2012). Solving the Measurement Invariance Anchor Item Problem in Item Response Theory. *Journal of Applied Psychology*. doi: 10.1037/a0027934

Meara, K., Robin, F., & Sireci, S. G. (2000). Using Multidimensional Scaling to Assess the Dimensionality of Dichotomous Item Data. *Multivariate Behavioral Research, 35*(2), 229-259.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology*. Belmont, CA: Wadsworth/ThomsonLearning.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining Sources of Gender DIF in Mathematics Assessments Using a Confirmatory Multidimensional Model Approach. *Applied Measurement in Education, 19*(4), 289-304.

Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. J. (1996). Validity and Washback in Language Testing. *Research Report No. 96-17*. Princeton, NJ: Educational Testing Service.

Messick, S. J. (1998). Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological Assessment. *Research Report No. 98-48*. Princeton, NJ: Educational Testing Service.

Miller, T. R., & Spray, J. A. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30*(2), 107-122.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A Brief Introduction to Evidence Centered Design. *Research Report RR-03-16*. Princeton: Educational Testing Services.

Monahan, P. O., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of Exploratory and Cross-Validated DETECT Index Under Unidimensionality. *Applied Psychological Measurement, 31*(6), 483-503. doi: 10.1177/0146621606292216

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological Testing: Principles and Applications*. Upper Saddle River, NJ: Pearson Prentice Hall.

Nandakumar, R. (1993). Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

Nandakumar, R. (1994). Assessing Dimensionality of a Set of Item Responses: Comparison of Different Approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Nandakumar, R., Yu, F., & Zhang, Y. (2011). A Comparison of Bias Correction Adjustments for the DETECT Procedure. *Applied Psychological Measurement, 35*(2), 127–144.

Nankervis, B. (2011). Gender Inequities in University Admission Due to the Differential Validity of the SAT. *Journal of College Admission, 213*, 24-30.

Narayanon, P., & Swaminathan, H. (1996). Identification of Items that Show Nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274.

Nunnally, J. C. (1981). *Psychometric Theory*. New Delhi: McGraw-Hill Company Limited.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-Based Item Invariance Indexes: The Effect of Between-Group Variation in Trait Correlation. *Journal of Educational Measurement, 27*(3), 273-283.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and Item Bias in Item Response Theory. *Applied Psychological Measurement, 16*(3), 237-248.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Journal of Educational Measurement, 34*(3), 253-272.

Osterlind, S. J. (1983). *Test Item Bias*. Beverly Hills, CA: Sage Publication.

Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*. doi: 10.1177/0265532211434027

Park, G.-P. (2008). Differential Item Functioning on an English Listening Test across Gender. *TESOL Quarterly, 42*(1), 115-123. doi: 10.2307/40264430

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of Model Selection Strategies for Cross-Level Two-Way Differential Item Functioning Analysis. *Educational and Psychological Measurement, 72*(1), 44-51.

Penfield, R. D., & Camilli, G. (2007). Differential Item Functioning and Item Bias. Dalam C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (Edisi ke-1, hh. 125-167). Amsterdam: Elseveir.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing Differential Item Functioning in Performance Assessment: Review and Recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15.

Raju, N. S. (1990). Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational & Psychological Measurement, 53*, 301-314. doi: 10.1177/0013164493053002001

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M. D. (1997). The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reckase, M. D., & Ackerman, T. A. (1986). *Building a Test Using Items That Require More than One Skill to Determine a Correct Answer*. Paper dipresentasikan pada The Annual Meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a Unidimensional Test Using Multidimensional Items. *Journal of Educational Measurement, 25*(3), 193-203.

Reckase, M. D., & McKinley, R. L. (1991). The Discriminating Power of Items That Measure More Than One Dimension. *Applied Psychological Measurement, 15*(4), 361-373.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment, 92*(6), 544-559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Ridho, A. (2011). *Multidimensionalitas Tes Potensi Akademik*. Paper dipresentasikan pada Second International Conference of Indigenous and Cultural Psychology, Denpasar, Bali.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT Population Parameter and Evaluation of DETECT Estimator Bias. *Journal of Educational Measurement, 43*(3), 215–243.

Roussos, L. A., & Stout, W. F. (1996a). A Multidimensionality-Based DIF Analysis Paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement, 33*(2), 215-230.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using New Proximity Measures with Hierarchical Cluster Analysis to Detect Multidimensionality. *Journal of Educational Measurement, 35*(1), 1-30.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased Item Detection Techniques. *Journal of Educational Statistics, 5*(3), 213-233.

Santelices, M. V., & Wilson, M. (2012). On the Relationship Between Differential Item Functioning and Item Difficulty: An Issue of Methods? Item Response Theory Approach to Differential Item Functioning. *Educational and Psychological Measurement, 72*(1), 5-36.

Scarpati, S. E., Wells, C. S., Lewis, C., & Jirka, S. (2011). Accommodations and Item-Level Analyses Using Mixture Differential Item Functioning Models. *The Journal of Special Education, 45*(1), 54–62.

Scheuneman, J. D., & Gerritz, K. (1990). Using Differential Item Functioning Procedures to Explore Sources of Item Difficulty and Group Performance Characteristics. *Journal of Educational Measurement, 27*(2), 109-131.

Schilling, S. G. (2007). The Role of Psychometric Modeling in Test Validation: An Application of Multidimensional Item Response Theory. *Measurement: Interdisciplinary Research and Perspectives, 5*, 93-106.

Seo, M., & Roussos, L. A. (2010). Formulation of a DIMTEST Effect Size Measure (DESM) and Evaluation of the DESM Estimator Bias. *Journal of Educational Measurement, 47*(4), 413–431.

Seraphine, A. E. (2000). The Performance of Dimtest When Latent Trait and Item Difficulty Distributions Differ. *Applied Psychological Measurement, 24*(1), 82-94. doi: 10.1177/01466216000241005

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194. doi: 10.1007/bf02294572

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Social Indicators Research, 102*(3), 443-461.

Snow, T. K., & Oshima, T. C. (2009). A Comparison of Unidimensional and Three-Dimensional Differential Item Functioning Analysis Using Two-Dimensional Data. *Educational and Psychological Measurement, 69*(5), 732-747.

Socha, A., & DeMars, C. E. (2013). A Note on Specifying the Guessing Parameter in ATFIND and DIMTEST. *Applied Psychological Measurement, 37*(1), 87-92. doi: 10.1177/0146621612464693

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the Effects of Differential Item (Functioning and Differential) Test Functioning on Selection Decisions: When Are Statistically Significant Effects Practically Important? *Journal of Applied Psychology, 89*(3), 497-508. doi: 10.1037/0021-9010.89.3.497

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415. doi: 10.1037/1082-989x.11.4.402

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing Subscale Scores for Diagnostic Information: A Case Study When the Test is Essentially Unidimensional. *Applied Measurement in Education, 23*(1), 63-86.

Stone, C. A., & Yeh, C.-C. (2006). Assessing the Dimensionality and Factor Structure of Multiple-Choice Exams : An Empirical Comparison of Methods Using the Multistate Bar Examination. *Educational and Psychological Measurement, 66*(2), 193-214.

Stone, E., Cook, L., Laitusis, C. C., & Frederick, C. (2010). Using Differential Item Functioning to Investigate the Impact of Testing Accommodations on an English-Language Arts Assessment for Students who are Blind or Visually Impaired. *Applied Measurement in Education, 23*(2), 132-152.

Stoneberg Jr., B. D. (2004). A Study of Gender-Based and Ethnic-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi Square Test. Retrieved 1 Juni, 2011, from http://www.eric.ed.gov/PDFS/ED489949.pdf

Stout, W. F. (1984). A Statistical Procedure for Assessing Test Dimensionality. *Measurement Series 84-2*. Washington, D.C.: ERIC Clearinghouse.

Stout, W. F. (1989). A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation. *Cognitive Science Program*. Champaign, IL: Department of Statistics - Univ. of Illinois.

Stout, W. F. (2002). Psychometrics: From Practice to Theory and Back (15 Years of Nonparametric Multidimensional IRT, DIF/Test Equity, and Skills Diagnostic Assessment). *Psychometrika, 67*(4), 485-518.

Stout, W. F., Bolt, D. M., Froelich, A. G., Habing, B., Hartz, S., & Roussos, L. A. (2003). Development of a SIBTEST Bundle Methodology for Improving Test Equity, With Applications for GRE Test Development. *GRE Research*. Princeton, NJ: Educational Testing Service.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional Covariance-Based Nonparametric Multidimensionality Assessment. *Applied Psychological Measurement, 20*(4), 331-354.

Stout, W. F., & Nandakumar, R. (2006). DIMTEST 2.1 [Computer Software]. Missoula: Assessment System Corporation.

Suryabrata, S. (2000). *Pengembangan Alat Ukur Psikologi*. Yogyakarta: Andi.

Svetina, D., & Levy, R. (2012). An Overview of Software for Conducting Dimensionality Assessment in Multidimensional Models. *Applied Psychological Measurement, 36*(8), 659-669. doi: 10.1177/0146621612454593

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the Fit of Item Response Theory Models. Dalam C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (hh. 683-718). Amsterdam: Elseveir.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tate, R. (2000). Performance of a Proposed Method for the Linking of Mixed Format Tests with Constructed Response and Multiple Choice Items. *Journal of Educational Measurement, 37*(4), 329-346.

Tate, R. (2002). Test Dimensionality. Dalam G. Tindal & T. M. Haladyna (Eds.), *Large-Scale Assessment Program for All Students: Validity, Technical Adequacy, and Implementation* (Edisi ke-1, hh. 181-211). Mahwah, NJ: Lawrence Erlbaum.

Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Responses to Test Items. *Applied Psychological Measurement, 27*(3), 159–203.

Teresi, J. A., & Fleishman, J. A. (2007). Differential Item Functioning and Health Assessment. *Quality of Life Research, 16*(ArticleType: research-article / Issue Title: Supplement 1 / Full publication date: Aug., 2007 / Copyright © 2007 Springer), 33-42.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. Dalam P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* (Edisi ke-1, hh. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Thomas, R. M. (2005). *High-Stakes Testing: Coping with Collateral Damage*. New Jersey: Lawrence Erlbaum.

Togut, T. D. (2011). High-Stakes Testing: Educational Barometer for Success, or False Prognosticator for Failure. Retrieved 15 Agustus 2011, from http://www.harborhouselaw.com/articles/highstakes.togut.htm#1

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. *Applied Psychological Measurement, 28*(1), 3–24.

Vaughn, B. K., & Wang, Q. (2008). Classification Based on Tree-Structured Allocation Rules. *Journal of Experimental Education, 76*(3), 315-340.

Vaughn, B. K., & Wang, Q. (2010). DIF Trees: Using Classification Trees to Detect Differential Item Functioning. *Educational and Psychological Measurement, 70*(6), 941-952. doi: 10.1177/0013164410379326

Wagiran, B. K. (2005). *Perbandingan Berbagai Metode untuk Mendeteksi Bias Butir*. Yogyakarta: Program Pascasarjana UGM. Disertasi. Tidak diterbitkan.

Wagiran, B. K., & Retnawati, H. (2008). Mengembangkan Metode Pendeteksian Keberfungsian Butir Pembeda (Differential Item Functioning, DIF) Berdasarkan Teori Respons Butir. *Laporan Penelitian Fundamental (Tahun I)*. Yogyakarta: Universitas Negeri Yogyakarta.

Wagiran, B. K., & Retnawati, H. (2009). Mengembangkan Metode Pendeteksian Keberfungsian Butir Pembeda (Differential Item Functioning, DIF) Berdasarkan Teori Respons Butir. *Laporan Penelitian Fundamental (Tahun II)*. Yogyakarta: Universitas Negeri Yogyakarta.

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. Dalam P. W. Holland & H. Wainer (Eds.), *Differential*

Item Functioning (Edisi ke-1, hh. 123-135). Hillsdale, NJ: Lawrence Erlbaum.

Walker, C. M. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment, 29*(4), 364-376.

Walker, C. M., & Beretvas, S. N. (2001). An Empirical Investigation Demonstrating the Multidimensional DIF Paradigm: A Cognitive Explanation for DIF. *Journal of Educational Measurement, 38*(2), 147-163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing Multidimensional and Unidimensional Proficiency Classifications: Multidimensional IRT as a Diagnostic Aid. *Journal of Educational Measurement, 40*(3), 255-275.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a Multidimensional Differential Item Functioning Framework to Determine if Reading Ability Affects Student Performance in Mathematics. *Applied Measurement in Education, 21*(2), 162-181.

Wang, W.-C., & Su, Y.-H. (2004). Effects of Average Signed Area Between Two Item Characteristic Curves and Test Purification Procedures on the DIF Detection via the Mantel-Haenszel Method. *Applied Measurement in Education, 17*(2), 113-144.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of Anchor Item Methods on Differential Item Functioning Detection With Likelihood Ratio Test. *Applied Psychological Measurement, 27*(6), 479-498.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The Comparative Effects of Compensatory and Noncompensatory Two-Dimensional Data on Unidimensional IRT Estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The Performance of IRT Model Selection Methods With Mixed-Format Tests. *Applied Psychological Measurement, 36*(3), 159-180. doi: 10.1177/0146621612440305

Woods, C. M. (2009). Empirical Selection of Anchors for Tests of Differential Item Functioning. *Applied Psychological Measurement, 33*(1), 42–57.

Yang, W.-L. (2004). Sensitivity of Linkings Between AP Multiple-Choice Scores and Composite Scores to Geographical Region: An Illustration of Checking for Population Invariance. *Journal of Educational Measurement, 41*(1), 33-41. doi: 10.1111/j.1745-3984.2004.tb01157.x

Yao, L. (2010a). BMIRT: Bayesian multivariate item response theory. [Computer Software]. Monterey, CA: Defense Manpower Data Center.

Yao, L. (2010b). Reporting Valid and Reliable Overall Scores and Domain Scores. *Journal of Educational Measurement, 47*(3), 339–360.

Yao, L. (2011). Multidimensional Linking for Domain Scores and Overall Scores for Nonequivalent Groups. *Applied Psychological Measurement, 35*(1), 48–66.

Yao, L. (2012). Multidimensional CAT Item Selection Methods for Domain Scores and Composite Scores: Theory and Applications. *Psychometrika, 77*(3), 495-523. doi: 10.1007/s11336-012-9265-5

Yao, L., & Boughton, K. A. (2007). A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification. *Applied Psychological Measurement, 31*(2), 83–105.

Yao, L., & Li, F. (2010). *A DIF Detection Procedure in Multidimensional Item Response Theory Framework and its Applications*. Paper dipresentasikan pada Annual Meeting of the National Council on Measurement in Education, Colorado, Denver.

Yen, S. J., & Walker, L. (2007). *Multidimensional IRT Models for Composite Scores*. Paper dipresentasikan pada Annual Meeting of the National Council of Measurement in Education, Chicago, IL.

Young, J. W. (1991). Gender Bias in Predicting College Academic Performance: A New Approach Using Item Response Theory. *Journal of Educational Measurement, 28*(1), 37-47.

Zhang, B. (2008). Application of Unidimensional Item Response Models to Tests With Items Sensitive to Secondary Dimensions. *Journal of Experimental Education, 77*(2), 147.

Zhang, J. (2012a). Calibration of Response Data Using MIRT Models With Simple and Mixed Structures. *Applied Psychological Measurement, 36*(5), 375-398. doi: 10.1177/0146621612445904

Zhang, J. (2012b). A Procedure for Dimensionality Analyses of Response Data from Various Test Designs. *Psychometrika*, 1-22. doi: 10.1007/s11336-012-9287-z

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 231–249.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3). Lincolnwood, IL: Scientific Software International.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.