MULTIDIMENSIONALITAS PADA TES POTENSI AKADEMIK[1]

Ali Ridho[2]

Abstract

The aim of this research study was to find out characteristics of items and subtests of Tes Potensi akademik (TP) College Admissions (ujian masuk, UM) UGM 2006 approached by unidimensional and multidimensional item response theory with 3 parameter logistic model. Meanwhile, dimensionality investigated by conditional covariance-based approached.

The data for the research consist of UM UGM 2006 applicants' responses. The subjects were 15670. The items were calibrated by unidimensional item response theory (UIRT) and multidimensional item response theory (MIRT) with 3 parameter logistic model using BILOG-MG and BMIRT program. Dimensionality assess by semi confirmatory factor analysis using HCA/CCPROX, DETECT, DIMTEST procedures.

Results of the study show that items of Verbal, Quantitative, and Reasoning subtests grouped by HCA/CCPROX and DETECT procedures into 3 clustered as well as the blueprint. Nevertheless, DIMTEST procedure shows that items of Quantitative and Reasoning subtest were locally independent.

Keyword: TP, item analysis, dimensionality assessment, multidimensional item response theory

A. PENGANTAR

Ujian masuk perguruan tinggi (PT) adalah peristiwa yang krusial. Implikasinya begitu besar, berpengaruh terhadap masa depan para peserta ujian. Sebagai *high stakes testing*, alat ukur yang digunakan dalam ujian masuk PT seharusnya memiliki karakteristik psikometrik yang baik, dalam arti valid. Validitas, menurut Messick (1995), adalah konsepsi tunggal yang harus dilihat dari 6 aspek: isi, substansi, struktural, generalisasi, eksternal, dan konsekuensi. Penelitian ini akan melihat validitas tes dari aspek struktural, yaitu dimensi atribut laten peserta dalam menjawab benar aitem-aitem dalam tes.

Dimensi dalam tes penting untuk diteliti karena hal tersebut mempengaruhi skoring, analisis data dan laporan hasilnya (Abedi, 1997; Kahraman & Thompson, 2011). Selain itu, bila tes terdiri dari beberapa subtes, perlu diperhatikan kombinasi skornya, baru kemudian menginterpretasikan skor komposit tersebut (Ackerman, 1994; Reckase & McKinley, 1991). Yao (2011) memberikan saran:

---

[1] Paper dipresentasikan pada The Second International Conference of Indigenous and Cultural Psychology di Denpasar, Bali – Indonesia (December 21-23, 2011)

[2] Mahasiswa Program Doktor Psikologi Universitas Gadjah Mada Yogyakarta, Dosen di Fakultas Psikologi UIN Maliki Malang. Email: ali.ridho@yahoo.com

pertama, perlu diketahui atribut laten komposit apa yang hendak diukur. Kedua, perlu dipastikan seluruh peserta tes diukur pada skala komposit yang sama melalui suatu model kombinasi linear atribut laten.

Salah satu komponen utama yang biasanya menjadi bagian dari ujian masuk PT adalah tes yang didesain mengukur atribut laten potensi akademik. Tes ini dirancang untuk mengungkap kemampuan individu dalam menghadapi problem kognitif yang perlu diselesaikan dengan strategis dan cepat. Umumnya Tes Potensi (TP) terdiri dari 3 subtes, masing-masing mengukur kemampuan verbal, kuantitatif, dan penalaran (Azwar, 2008). Aitem-aitem dalam TP berbentuk pilihan ganda sehingga dapat diskor secara dikotomi (0 untuk salah dan 1 untuk benar).

Pada tes yang diskor secara dikotomi, diperlukan suatu model yang mampu menjelaskan interaksi antara respons yang diberikan oleh peserta tes dengan aitem-aitem dalam tes. Model yang menjadi tren adalah *unidimensional item response theory* (UIRT). Salah satu asumsi utama yang mendasari UIRT adalah unidimensionalitas. Unidimensionalitas berarti hanya terdapat satu atribut laten yang mendasari para peserta tes dalam menjawab aitem (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Sekumpulan aitem-aitem dalam tes dapat disebut unidimensional bila kinerja pada peserta tes dapat dijelaskan oleh sebuah atribut laten (Hambleton & Rovinelli, 1986). Lebih jauh, 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). Bila sebuah atribut laten belum cukup mampu menjelaskan, dengan sendirinya independensi lokal tidak terpenuhi (Stout, 1984, 1989, 2002). Akhirnya asumsi unidimensional tidak dapat dipertahankan. Implikasinya, sekumpulan aitem disebut sebagai multidimensional. Asumsi unidimensional juga bersifat problematik, yaitu meskipun aitem-aitem tes didesain untuk mengukur satu atribut laten tertentu, sering kali para peserta memerlukan lebih dari satu atribut laten dalam menjawab benar sebuah aitem.

Selain IRT, model lain yang belum begitu berkembang adalah *multidimensional item response therory* (MIRT) (Reckase, 1985; Reckase & Ackerman, 1986). Pada model ini dimungkinkan aitem-aitem direspons benar oleh para peserta tes berdasarkan pada atribut laten lebih dari satu. Menurut pandangan model ini, data respons yang bersifat multidimensi kemudian diperlakukan sebagai data unidimensi, berarti telah menyimpang dari asumsi unidimensionalitas dalam UIRT, sekaligus aspek struktural dari konstrak yang diukur (Messick, 1995). Alat ukur yang terbukti multidimensi kemudian diasumsikan unidimensi bisa 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; E. Stone, Cook, Laitusis, & Frederick, 2010; Zumbo, 1999).

Mendasarkan hal-hal yang diungkap di atas, penelitian ini hendak melakukan penyelidikan terhadap dimensi atribut laten yang mendasari peserta merespons jawaban benar pada aitem-aitem tes TP. Berdasarkan sudut pandang tiap subtes, apa betul bahwa kinerja peserta tes pada Subtes Penalaran, misalnya,

betul-betul ditentukan oleh atribut kemampuan penalaran yang ia miliki? Bisa jadi ada atribut laten lain (misalnya kemampuan kuantitatif) ikut menjadi pendukung penting dalam memahami materi aitem-aitem dalam subtes. Hadirnya atribut laten kedua (kemampuan kuantitatif) yang ikut menentukan kinerja peserta pada aitem selain tujuan utamanya (penalaran) akan menjadikan aitem menjadi multidimensi, tidak lagi unidimensi. Bila dipaksakan unidimensi, skor yang dihasilkan menjadi invalid (Ackerman, 1994).

Penelitian tentang ini penting untuk dilakukan dengan beberapa argumen berikut. Di Indonesia, penelitian tentang tes masih jarang dilakukan. Bilapun ada, sebagian besar masih terbatas menggunakan paradigma teori tes klasik atau *classical test theory* (CTT), yang menurut banyak ahli pengukuran memiliki berbagai kelemahan (misalnya Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, dkk., 1991). 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. Tulisan ini mencoba mengaplikasikan perkembangan-perkembangan ini.

B. METODE

1. Data dan Instrumen

Data respons yang digunakan dalam penelitian ini adalah 15670 data respons peserta tes tulis Tes Potensi (TP) Ujian Masuk (UM) UGM pada tahun 2006 yang terdiri dari 60 aitem, terbagi merata dalam 3 Subtes (Verbal, Kuantitatif, dan Penalaran).

2. Dimensionalitas

Dalam konteks sebuah tes terdiri dari beberapa subtes sebagaimana dalam penelitian ini, hasil evaluasi dimensionalitas tes menentukan perlu tidaknya melaporkan subskor. Bila aitem-aitem dalam dua subtes bersifat unidimensional, maka sebaiknya skor subtes-subtes tersebut dijadikan satu, tidak perlu dilaporkan dua-duanya (Haberman & Sinharay, 2010; C. A. Stone, Ye, Zhu, & Lane, 2010; Yao, 2010, 2011).

a. Hierarchical Cluster Analysis (HCA)

Tes dengan aitem sebanyak N, untuk sembarang pasang aitem i dan j, dapat dibuat $N - 1$ tabel kontingensi, sebuah tabel 2×2 untuk tiap kemungkinan jumlah jawaban benar, k. Ukuran kedekatan untuk mengidentifikasi multidimensionalitas adalah:

$$p_{\mathrm{CCOR}} = \sqrt{2\left(1 - \frac{1}{\Sigma n_k}\sum_{k=0}^{N-2} n_k COR_k\right)} \qquad (1)$$

$$p_{\text{CCOV}} = -\frac{1}{\sum n_k}\sum_{k=0}^{N-2} n_k COV_k + \text{konstan} \qquad (2)$$

dimana sebuah nilai konstan ditambahkan agar memiliki harga ukuran kedekatan berharga ≥ 0, COV_k adalah kovarians antara dua aitem berdasarkan para peserta tes yang memperoleh k benar dari sisa aitem, COR_k adalah korelasi *product-moment* yang bersesuaian dengan COV_k.

Untuk tes yang bersifat multidimensi, bila dua aitem mengukur dimensi yang sama, berarti telah terjadi dependensi lokal positif. Bila dua aitem mengukur dimensi yang berbeda, berarti telah terjadi dependensi lokal negatif.

b. DETECT

Konsep dasar indeks DETECT merupakan komposit terbobot dari koordinat θ pada dimensi ruang. Komposit yang dibobot didefinisikan sebagai

$$\theta_w = \mathbf{w}\boldsymbol{\theta}' = \sum_{k=1}^{m} w_k \theta_k, \qquad (3)$$

dimana $\mathbf{w}$ adalah vektor bobot pada m koordinat pada ruang dan θ_w adalah nilai komposit koordinat. Bobot diskalakan sehingga varians θ_w berharga 1.

Prosedur DETECT menentukan nilai-nilai elemen-elemen vektor $\mathbf{w}$ sehingga memaksimumkan nilai harapan pada ruang $\boldsymbol{\theta}$. $\mathbf{w}^*$ merupakan vektor bobot yang memaksimumkan nilai harapan. Estimasi elemen $\mathbf{w}^*$ menggunakan

$$w_\ell^* = c\sum_{i=1}^{k} E\left\{\frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_\ell}\left[\sum_{I=1}^{k} P_I(\boldsymbol{\theta})Q_I(\boldsymbol{\theta})\right]^{-\frac{1}{2}}\right\} \text{untuk } \ell = 1, 2, ..., m, \qquad (4)$$

dimana w_ℓ^* adalah elemen $\mathbf{w}^*$, c adalah konstanta penskalaan varians agar θ_w bernilai 1, k adalah jumlah aitem dalam tes, $P_i(\boldsymbol{\theta})$ adalah fungsi respons aitem i, dan m adalah jumlah dimensi yang diperlukan untuk memodelkan hubungan dalam matriks data.

Saat mengimplementasikan DETECT, aitem-aitem di partisi menjadi sehimpunan disjoint, $\mathbf{P} = \{A_1, A_2, ..., A_q\}$ yang menunjukkan konstrak yang diukur. Hal ini dapat dilakukan melalui justifikasi ahli, atau menggunakan fasilitas DETECT untuk menemukan partisi yang memaksimumkan indeks DETECT, $D_w(\mathbf{P})$, yaitu

$$D_w(\mathbf{P}) = \frac{2}{k(k-1)}\sum_{1\leq i\leq j\leq k} \delta_{ij}(\mathbf{P}) E[Cov(U_i, U_j \mid \theta_w)] \qquad (5)$$

dimana $\mathbf{P}$ adalah sembarang partisi, k adalah jumlah aitem, i dan j adalah sembarang dua aitem dalam test yang berkorespondensi dengan U_i dan U_j, θ_w adalah komposit yang dimaksud, sedangkan

$$\delta_{ij}(\mathbf{P}) = \begin{cases} 1 & i, j \in A_\ell \\ -1 & \text{selainnya.} \end{cases} \qquad (6)$$

Simbol δ adalah bertanda 1 jika dua aitem dalam partisi yang sama dan -1 bila berbeda partisi. Indeks DETECT yang dihasilkan akan memiliki kemungkinan

nilai 0 sampai dengan 5. Indeks dengan nilai ≥ 1 menunjukkan multidimensionalitas yang besar, nilai 0,4 sampai dengan 1 menunjukkan multidimensionalitas besar menengah, < 0,4 berarti sedang, dan < 0,2 berarti unidimensional (Zhang & Stout, 1999).

c. DIMTEST

Prosedur dalam DIMTEST (Stout, 2002; Zhang & Stout, 1999) didasarkan pada asumsi bahwa interaksi antara para peserta dan aitem-aitem dapat dideskripsikan dalam bentuk umum model MIRT yaitu probabilitas menjawab benar sebagai fungsi $\mathbf{a\theta}' + d$. Model MIRT mengasumsikan bahwa probabilitas menjawab benar akan meningkat secara monoton dengan meningkatnya elemen vektor $\boldsymbol{\theta}$.

Prosedur dalam DIMTEST akan menentukan arah daya beda terbesar tes secara keseluruhan sebagai satu kesatuan menggunakan konsep komposit sebagai referensi. Nilai harapan kovarians antar aitem bersifat kondisional pada komposit elemen $\boldsymbol{\theta}$ sehingga menghasilkan daya beda terbesar, $\theta_Y = \mathbf{a_Y\theta'}$, dimana $\mathbf{a_Y}$ adalah vektor parameter $\mathbf{a}$ yang menentukan arah pengukuran terbaik:

$$E[\mathrm{cov}(U_i, U_j) \mid \theta_Y], \quad i \neq j, \tag{7}$$

dimana *i* dan *j* adalah indeks untuk pasangan-pasangan aitem dalam tes.

Karena θ_Y tidak dapat diamati secara langsung, DIMTEST menggunakan jumlah jawaban benar, *Y*, sebagai pendekatan θ_Y. Uji statistik yang digunakan dalam DIMTEST adalah kovarians antar aitem kondisional pada *Y*

$$E[\mathrm{cov}(U_i, U_j) \mid \theta_Y] = \sum_{k=0}^{n} P(Y = k)\,\mathrm{cov}(U_i, U_j \mid Y = k), \tag{8}$$

dimana *n* adalah jumlah aitem dalam tes.

Bila matriks skor aitem dapat dimodelkan secara akurat oleh sehimpunan fungsi respons aitem yang bersesuaian, maka θ_Y adalah nilai θ yang bersifat unidimensional yang dapat digunakan untuk memodelkan data. Dalam kondisi semacam itu, kovarians harapan kondisional pada θ_Y akan bernilai nol. Dengan demikian DIMTEST menguji hipotesis nol kovarians kondisional bernilai nol vs kovarians kondisional bernilai tidak nol yang mengindikasikan bahwa data tidak dapat direpresentasikan secara akurat dalam model UIRT.

Dalam praktiknya, DIMTEST mensyaratkan sehimpunan aitem dibagi paling sedikit dalam dua kelompok partisi. Kelompok partisi pertama disebut sebagai *partitioned subtest* (PT). Jumlah jawaban benar pada subtes ini digunakan sebagai variabel kondisional pada saat menghitung kovarians inter aitem. Kelompok partisi ke dua terdiri dari aitem-aitem yang mengukur atribut yang paling berbeda dari PT, disebut sebagai *assessment test* (AT). Uji hipotesis didasarkan pada kovarians inter-aitem pada aitem-aitem pada partisi AT,

$$T = \sum_{i \neq j \in \mathrm{AT}} \sum_{k=0}^{n} \mathrm{cov}(U_i, U_j \mid Y_{\mathrm{PT}} = k), \tag{9}$$

dimana Y_{PT} adalah jumlah benar dalam partisi PT dan *n* adalah jumlah aitem dalam partisi PT. Nilai *T* dikonversikan pada skor *z* dengan cara membaginya dengan deviasi standar distribusi sampel dan kemudian dibandingkan dengan nilai kritik berdasarkan distribusi normal standar.

$$T = \frac{\sum_{i \neq j \in \mathrm{AT}} \sum_{k=0}^{n} \mathrm{cov}(U_i, U_j \mid Y_{\mathrm{PT}} = k)}{\sqrt{\sum_{k=0}^{n} s_k^2}}, \tag{10}$$

dimana

$$s_k^2 = \frac{\hat{\mu}_{4k} - \hat{\sigma}_k^4 + \hat{\delta}_{4k}}{J_k}, \tag{11}$$

$$\hat{\mu}_{4k} = \frac{1}{J_k} \sum_{j=1}^{J_k} \left(Y_j^{(k)} - \bar{Y}^{(k)} \right)^4, \tag{12}$$

$$\sigma_k = \sqrt{\frac{1}{J_k} \sum_{j=1}^{J_k} \left(Y_j^{(k)} - \bar{Y}^{(k)} \right)^2}, \tag{13}$$

$$\hat{\delta}_{4k} = \sum \hat{p}_i^{(k)} \left(1 - \hat{p}_i^{(k)} \right) \left(1 - 2\hat{p}_i^{(k)} \right)^2, \tag{14}$$

$$\hat{p}_i^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} U_{ij}^{(k)}, \tag{15}$$

dimana J_k adalah jumlah peserta pada kelompok skor k dalam PT, $Y_j^{(k)}$ adalah skor dalam AT untuk peserta j dengan skor k dalam PT, dan $U_{ij}^{(k)}$ adalah skor peserta j pada aitem i yang memiliki skor k dalam PT.

Stout, dkk. (2003) mengoreksi statistik T pada persamaan (9) dengan memasukkan unsur T_G yang diperoleh dari rata-rata N data simulasi. Uji statistiknya menjadi:

$$T = \frac{T_L - \bar{T}_G}{\sqrt{1 + \frac{1}{N}}}. \tag{16}$$

Partisi AT dapat ditentukan berdasarkan justifikasi ahli, melalui analisis faktor, atau berdasarkan *item clustering*. Tiap teknik sangat mungkin akan menghasilkan kesimpulan yang berbeda. Untuk itu Reckase (2009) menyarankan agar membagi data respons yang diuji menjadi dua bagian (bila terdapat 5000 peserta, dibagi menjadi tiap bagian 2500 peserta). Bagian pertama digunakan untuk melakukan *item clustering*, bagian kedua digunakan untuk menguji hasil *item clustering* bagian pertama. Bagian pertama digunakan untuk menentukan aitem mana saja yang masuk partisi PT dan mana saja yang masuk partisi AT. Bagian ke dua digunakan untuk menguji hipotesis sebagaimana pada persamaan (16).

3. Multidimensional Item Response Theory (MIRT)

Satu di antara fakta dalam tes dengan bentuk pilihan ganda adalah bahwa peserta akan menjawab aitem dengan benar melalui tebakan sehingga probabilitas menjawab benar melibatkan komponen tambahan, yaitu parameter tebakan. Hal

ini sangat dimungkinkan pada aitem yang berbentuk pilihan ganda. Untuk itu Birnbaum (1968, dalam Lord, 1980) memodifikasi model 2PL:

$$P(u_{ij}=1|\,a_i,b_i,\theta_j)=\frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \tag{17}$$

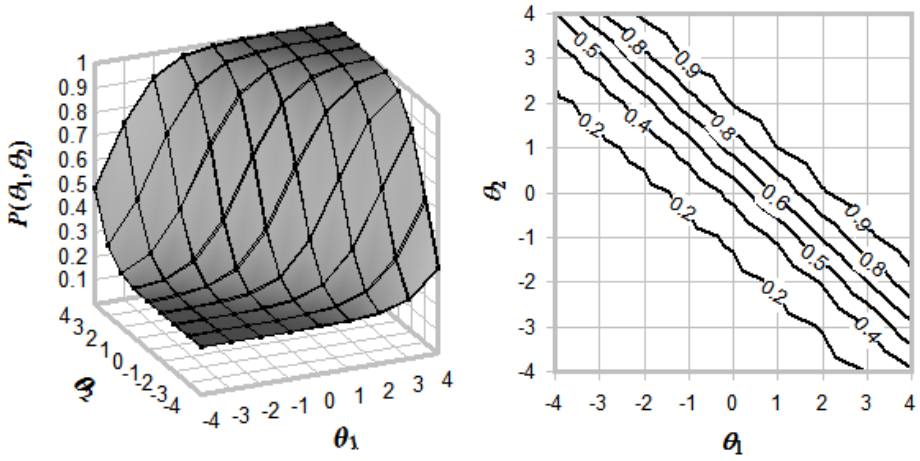
dimana $P(u_{ij}=1|\,a_i,b_i,\theta_j)$ adalah probabilitas menjawab benar aitem *i*, saat level kemampuan peserta tes θ_j, dan a_i, b_i, merupakan parameter tingkat kesukaran dan daya beda aitem *i*; menjadi model logistik tiga–parameter (UIRT 3PL):

$$P(u_{ij}=1|\,a_i,b_i,c_i,\theta_j)=c_i+(1-c_i)\frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \tag{18}$$

dengan memasukkan sebuah parameter tambahan yang merepresentasikan kontribusi tebakan pada probabilitas menjawab benar (c_i).

Model UIRT 3PL dapat ditingkatkan menjadi model MIRT 3PL dengan cara mengembangkan persamaan (18) sehingga menjadi:

$$P(U_{is}=1|\,\boldsymbol{\theta}_s,\mathbf{a_i},c_i,d_i)=c_i+(1-c_i)\frac{e^{\mathbf{a_i}\boldsymbol{\theta}_s'+d_i}}{1+e^{\mathbf{a_i}\boldsymbol{\theta}_s'+d_i}}. \tag{19}$$



Gambar 1. Plot Permukaan dan Kontur Aitem, a_1 = 1,3; a_2 = 1,4; c = 0,2; dan d = -0,7

Sebagai contoh dalam model ini, katakanlah sebuah aitem dengan parameter a_1 = 1,3, a_2 = 1,4, c = 0,2, dan d = -0,7. Berdasarkan parameter ini dapat digambarkan *item characteristic surface* (ICS) sebagaimana pada Gambar 1. Penggambaran MIRT dua dimensi dalam bentuk ICS ini diyakini oleh Ackerman (1996) sangat membantu rasionalisasi pemahaman karakteristik aitem.

Reckase (1985) mengembangkan formula yang menunjukkan daya beda aitem secara keseluruhan dalam MIRT, disebut *maximum discrimination index* (MDISC). Persamaannya yaitu

$$MDISC_i=(\sum_{k=1}^{m}a_{ik}^2)^{1/2}, \tag{20}$$

dimana $MDISC_i$ menunjukkan daya beda multidimensi aitem *i* dan a_{ik} adalah parameter daya beda aitem *i* pada dimensi ke-*k*. Sementara itu, tingkat kesukaran multidimensi aitem *i* ($MDIFF_i$), ditunjukkan oleh

$$MDIFF_i = -\frac{d_i}{\mathbf{a_i}'\mathbf{a_i}} = -\frac{d_i}{MDISC_i}. \quad (21)$$

Untuk menunjukkan arah daya beda aitem terbaik pada dimensi ruang *m*, digunakan persamaan

$$\cos\alpha_{ik} = \frac{a_{ik}}{MDISC_i}, \quad (22)$$

dimana α_{ik} adalah sudut antara vektor aitem *i* dan koordinat *k* dalam dimensi ruang *m*.

4. Prosedur

Peneliti menggunakan prosedur penelitian sebagai berikut.

1) Analisis aitem dengan metode CTT dengan kriteria korelasi *point biserial* (r_{pbis}) ≥ 0,1 pada tiap Subtes;
2) Untuk mengetahui karakteristik aitem tiap Subtes berdasarkan *unidimensional item response theory* (UIRT) dilakukan kalibrasi parameter aitem pada tiap subtes dengan metode *marginal maximum likelihood* (MML) pada aitem-aitem tiap subtes yang hingga diperoleh kecocokan data baik pada level aitem maupun tes. Prosedur ini dilakukan dengan bantuan BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003).
3) Analisis dimensi dalam penelitian ini mengacu pada saran Jang dan Roussos (2007) yaitu dengan menerapkan teknik ekploratori dan konfirmatori pada struktur dimensi TP.
    a. Data respons dipecah menjadi dua secara random menjadi sampel utama dan sampel verifikasi.
    b. Analisis eksploratori dilakukan dengan prosedur HCA/CCPROX pada seluruh respons aitem TP untuk menemukan aitem-aitem yang secara homogen membentuk dimensi yang paling dominan pada sampel utama.
    c. Analisis eksploratori ke dua dilakukan dengan DETECT pada tiap subtes. Sampel utama digunakan terlebih dahulu untuk memaksimumkan partisi, sampel verifikasi digunakan untuk menghitung indeks DETECT.
    d. Temuan berdasarkan HCA/CCPROX dan DETECT digunakan untuk mengembangkan hipotesis menggunakan analisis konfirmatori dengan bantuan DIMTEST pada sampel verifikasi.

    Pilihan prosedur ini juga didasarkan hasil penelitian yang menunjukkan bahwa hasil deteksi DIMTEST menghasilkan power yang tinggi saat mendeteksi simpangan dari udimensional (Finch & Habing, 2007; Nandakumar, 1994).

4) Untuk mengetahui karakteristik aitem berdasarkan *multidimensional item response theory* (MIRT) dilakukan kalibrasi dengan bantuan BMIRT (Yao & Boughton, 2007)

C. HASIL

Hasil yang dilaporkan dalam penelitian ini meliputi: (1) karakteristik aitem tiap subtest dengan pendekatan CTT, (2) karakteristik aitem tiap subtest dengan pendekatan UIRT, (3) uji dimensionalitas, dan (4) karakteristik aitem dengan pendekatan MIRT.

1. Karakteristik Aitem Berdasarkan CTT

Analisis aitem secara klasik dilakukan secara terpisah pada tiap Subtes TP. Hasil analisis pada Subtes Verbal pada putaran ke dua menunjukkan 11 aitem yang memiliki $r_{bis} \geq 0{,}1$. Berdasarkan 11 aitem tersebut diperoleh $M(r_{bis}) = 0{,}23$, $SD(r_{bis}) = 0.08$, $MIN(r_{bis}) = 0{,}12$, dan $MAX(r_{bis}) = 0{,}34$. Pada Subtes Kuantitatif hanya terdapat aitem nomor 22 yang memiliki $r_{bis} \leq 0{,}1$. Deskripsi yang dihasilkan dari 19 aitem Subtes Kuantitatif adalah $M(r_{bis}) = 0{,}39$, $SD(r_{bis}) = 0{,}09$, $MIN(r_{bis}) = 0{,}21$, dan $MAX(r_{bis}) = 0{,}50$. Untuk Subtes Penalaran, aitem nomor 51 dan 56 memiliki $r_{bis} < 0{,}1$. Diantara 18 aitem yang diterima diperoleh $M(r_{bis}) = 0{,}39$, $SD(r_{bis}) = 0.11$, $MIN(r_{bis}) = 0{,}17$, dan $MAX(r_{bis}) = 0{,}52$.

2. Dimensionalitas

Analisis HCA/CCPROX pada level 45 menunjukkan pembagian 3 klaster, yaitu klaster 1 (Kuantitatif), klaster 2 (Penalaran), dan klaster 3 (Verbal). Ringkasan hasil pada Tabel 1 memperlihatkan bahwa secara umum aitem-aitem tiap Subtes TP dikelompokkan sesuai dengan posisinya. Secara khusus, perlu diperhatikan aitem nomor 41, 42, dan 52 yang menjadi anggota klaster Kuantitatif dimana seharusnya tiga aitem ini menjadi bagian klaster Penalaran. Keadaan ini mengindikasikan adanya kovarians kondisional tiga aitem tersebut dengan aitem-aitem dalam Subtes Kuantitatif.

Tabel 1. Rangkuman Hasil Output Analisis HCA/CCPROX

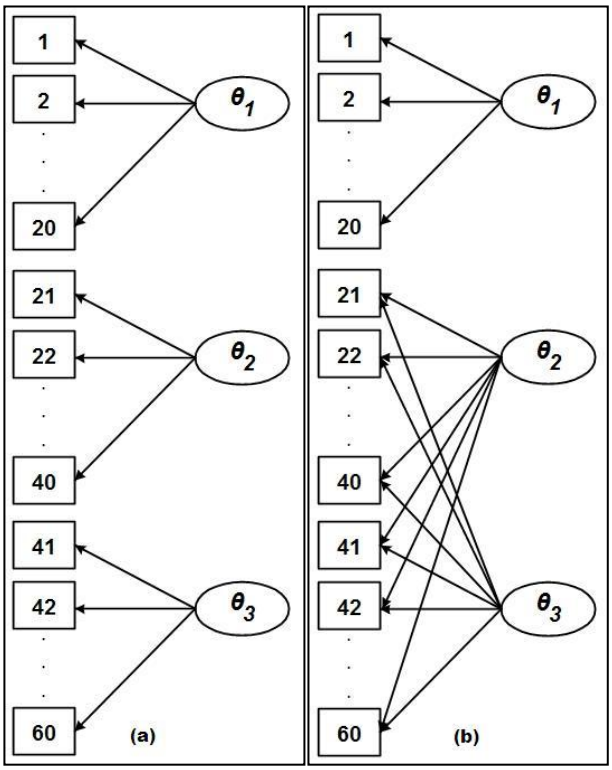
	Klaster	Nomor Aitem	Jumlah
1)	Klaster 1 (Kuantitatif)	23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, ***41, 42, 52***	21
2)	Klaster 2 (Penalaran)	43, 44, 45, 46, 47, 48, 49, 50, 53, 54, 55, 57, 58, 59, 60	15
3)	Klaster 3 (Verbal)	2, 3, 4, 5, 6, 9, 12, 13, 14, 17, 18, ***21***	12
		Total	48

Seluruh aitem yang berjumlah 11 mengelompok pada klaster 3 (Verbal). Namun terdapat aitem nomor 21 yang seharusnya menjadi bagian klaster 1 (Kuantitatif). Hal ini menunjukkan adanya kovarians kondisional aitem nomor 21 dengan aitem-aitem dalam Subtes Verbal.

Analisis DETECT menunjukkan hasil yang konsisten dengan prosedur HCA/CCPROX sebagaimana dalam Tabel 1. Statistik δ = 0,29 menunjukkan aitem-aitem dalam TP bersifat multidimensional dalam taraf sedang. Sementara itu indeks r = 0,78 yang mendekati 1 mengindikasikan struktur konstrak bersifat sederhana.

Berdasarkan hasil prosedur HCA/CCPROX dan DETECT, aitem nomor 41, 42, dan 52 menjadi bagian dari Subtes Kuantitatif, sedangkan aitem nomor 21 menjadi bagian dari Subtes Verbal. Peneliti memutuskan untuk memasukkan kedalam subtes sebagaimana desain awalnya pada analisis selanjutnya. Hal ini dikarenakan dari sudut pandang isi, sebenarnya aitem-aitem tersebut didesain untuk mengukur kemampuan penalaran (aitem nomor 41, 42, dan 52) dan kemampuan kuantitatif (aitem nomor 21).

Secara teoritik, struktur dimensi yang membentuk konstrak TP UM UGM 2006 dikembangkan sehingga bersifat multidimensional antar-aitem. Representasi grafisnya dituangkan dalam Gambar 2a. Hasil analisis klaster dengan HCA/CCPROX dan DETECT menunjukkan hasil yang konsisten dengan struktur dimensi sebagaimana yang diinginkan oleh pengembang, dengan sedikit penyimpangan, yaitu pada aitem-aitem nomor 21, 41, 42, dan 52. Berdasarkan hasil ini, peneliti memutuskan untuk memasukkan aitem-aitem tersebut dalam domain masing-masing pada analisis selanjutnya.



Gambar 2. Dimensionalitas Aitem-aitem TP

Hasil uji independensi lokal aitem-aitem antar subtes melalui DIMTEST dituangkan Tabel 2. Statistik T signifikan pada uji unidimensionalitas antara

aitem-aitem dalam Subtes Verbal dengan Subtes Kuantitatif (V – K) dan Subtes Verbal dengan Subtes Penalaran (V – P). Namun tidak demikian halnya dengan aitem-aitem dalam Subtes Kuantitatif dengan Subtes Penalaran. Dengan demikian dapat dikatakan bahwa aitem-aitem Kuantitatif bersifat unidimensional dengan aitem-aitem Penalaran.

Tabel 2. Statistik DIMTEST

Uji	TL	TGbar	T	*p*
V – K 1	9,4071	6,8569	2,5375	0,0056
V – K 2	11,6446	7,1106	4,5116	0,0000
K – P 1	***14,4102***	***14,5306***	***-0,1198***	***0,5477***
K – P 2	***13,1683***	***14,6231***	***-1,4476***	***0,9261***
V – P 1	10,2238	6,7746	3,4321	0,0003
V – P 2	13,4492	7,1918	6,2264	0,0000

Keterangan: signifikansi pasangan subtes yang perlu menjadi perhatian tercetak tebal dan miring.

Hasil uji simpangan dari independensi lokal ini membawa implikasi bahwa struktur dimensi TP UM UGM tahun 2006 menjadi model multidimensional campuran antara multidimensi antar-aitem dan dalam-aitem. Representasi grafis model ini dapat dilihat pada Gambar 2b. Bila dibandingkan dengan Gambar 2a, struktur Gambar 2b memiliki perbedaan pada kinerja peserta tes yang dipengaruhi oleh kemampuan mereka dalam menjawab aitem. Dimensi Kuantitatif dan Penalaran secara bersama-sama memberikan kontribusi pada probabilitas menjawab benar aitem-aitem nomor 21 sampai dengan 60.

3. Karakteristik Aitem Berdasarkan MIRT

Mendasarkan pada hasil uji dimensionalitas, dapat digambarkan struktur dimensi atribut laten Potensi Akademik (θ_0) terdiri dari Kemampuan Verbal (θ_1), Kemampuan Kuantitatif (θ_2), dan Kemampuan Penalaran (θ_3) sebagaimana pada Gambar 2b. Hasil uji kecocokan dengan data pada Tabel 3 menunjukkan bahwa model θ_2 dan θ_3 yang aitem-aitemnya dikalibrasi bersama sebagai bagian dari dimensi Kuantitatif dan Penalaran menghasilkan model yang relatif lebih cocok. Hal ini dapat dilihat dari harga *Akaike Information Criterion* (AIC) dan *Bayesian Information Criterion* (BIC). Bila dua model dibandingkan, lebih kecilnya harga AIC dan BIC menunjukkan kecocokan data yang lebih baik (Lee, 2007: 128).

Tabel 3. Kecocokan Data

Model	AIC	BIC
θ_2 dan θ_3 terpisah (Gambar 2a)	751219	1102056
θ_2 dan θ_3 bersama (Gambar 2b)	745857	1096694

Selanjutnya adalah mengetahui karakteristik tiap aitem berdasarkan tiga dimensi yang telah teridentifikasi, dengan struktur sebagaimana pada Gambar 2b.

Hasil estimasi parameter daya beda tiap dimensi (a_1, a_2, dan a_3), parameter tingkat kemudahan (d), tebakan semu (c), dapat dilihat secara lengkap pada Tabel 4.

Tabel 4. Parameter a_1, a_2, a_3, d, dan c Model MIRT

Nomor Aitem	a_1	a_2	a_3	d	c
1	-	-	-	-	-
2	1,5251	-	-	-0,0957	0,1572
3	0,7466	-	-	-0,7272	0,1753
4	0,7840	-	-	1,5867	0,1610
5	0,8301	-	-	0,5537	0,1620
6	0,7417	-	-	1,6683	0,1635
7	-	-	-	-	-
8	-	-	-	-	-
9	1,1012	-	-	-0,1733	0,1670
10	-	-	-	-	-
11	-	-	-	-	-
12	0,6461	-	-	1,1159	0,1729
13	1,3482	-	-	-1,4945	0,1532
14	0,9367	-	-	0,7475	0,1128
15	-	-	-	-	-
16	-	-	-	-	-
17	1,3566	-	-	-0,4858	0,1343
18	1,1865	-	-	-2,1592	0,1696
19	-	-	-	-	-
20	-	-	-	-	-
21	-	0,7465	0,6260	-0,8371	0,1698
22	-	-	-	-	-
23	-	0,6710	0,4993	-1,5892	0,1635
24	-	0,4274	0,8169	-1,2314	0,1567
25	-	0,6544	0,7573	-2,1644	0,1550
26	-	0,8270	1,5185	-1,5959	0,1640
27	-	0,8156	1,3649	-1,4284	0,1439
28	-	1,3664	1,1015	0,0775	0,1900
29	-	0,5967	0,6192	-0,0775	0,1816
30	-	1,8293	-	0,4610	0,1605
31	-	1,2752	0,5158	-1,0946	0,1624
32	-	1,0914	1,1928	-0,8916	0,1855
33	-	1,2940	1,1086	-0,5394	0,1433
34	-	1,3611	0,8161	0,7501	0,1473
35	-	0,9001	0,6199	1,2407	0,1061
36	-	0,9004	0,5028	0,0317	0,1917
37	-	1,6957	0,8754	0,8151	0,1840
38	-	0,5242	0,7247	0,7150	0,1565
39	-	1,2397	0,9585	0,8904	0,1532

Nomor Aitem	a_1	a_2	a_3	d	c
40	-	1,1648	0,7794	-0,7412	0,1676
41	-	1,3035	1,1124	1,1130	0,1316
42	-	1,7230	1,3381	0,9184	0,1383
43	-	0,8236	0,7451	-1,7946	0,1534
44	-	0,9945	1,2163	0,5089	0,1442
45	-	0,9440	1,1297	-1,1338	0,1405
46	-	0,5088	0,7712	-1,4364	0,1574
47	-	0,4642	0,4488	-0,7468	0,1840
48	-	1,2315	1,5948	0,5478	0,1778
49	-	0,5542	0,5661	-0,6828	0,1767
50	-	0,9795	0,9315	-1,1619	0,1799
51	-	-	-	-	-
52	-	1,4111	1,2879	0,6518	0,1243
53	-	0,5514	0,7513	-2,0384	0,1606
54	-	0,5310	0,5076	-1,4471	0,2065
55	-	0,9597	1,0798	-0,3050	0,1595
56	-	-	-	-	-
57	-	0,3696	0,4052	2,2955	0,1844
58	-	0,8657	1,0449	-0,5023	0,1838
59	-	0,9264	0,9680	0,2763	0,1633
60	-	0,9440	0,8922	1,2339	0,1222

D. DISKUSI

Selama ini sistem penskoran yang dilakukan terhadap TP dilakukan berdasarkan total jawaban benar. Kemampuan verbal, kuantitatif, dan penalaran dianggap sama-sama mengukur kemampuan potensi akademik. Kondisi ini tidak sejalan dengan realitas yang menunjukkan bahwa TP bersifat multidimensi. Model multidimensionalitas TP menunjukkan eksistensi adanya dimensi-dimensi kemampuan yang mengikuti model Gambar 2b. Dengan demikian, penskoran berdasarkan jumlah benar perlu diperbaiki dengan model MIRT 3PL. Manakala multidimensionalitas ini terbukti ada, MIRT lebih mampu menjelaskan interaksi antara aitem-aitem dalam tes dan jawaban para peserta. Hasil estimasi parameter aitem dan parameter kemampuan para peserta pun terbukti akurat dan efisien (Reckase, 2009).

Hasil penelitian ini menunjukkan bahwa pengambil kebijakan dalam tes masuk perguruan tinggi, khususnya terkait tes potensi, perlu mempertimbangkan kenyataan multidimensionalitas yang terkandung dalam aitem-aitem tes. Tidak seharusnya skor pengukuran tes potensi dilaporkan sebagaimana selama ini, yaitu skor keseluruhan dan skor tiap subtes. Hasil uji dimensionalitas mengarahkan pada saran bahwa sebaiknya subtes penalaran dan subtes kuantitatif dilaporkan sebagai satu subskor, sebut saja skor non verbal. Sementara skor verbal tetap

diberikan sebagaimana selama ini. Dengan demikian skor yang perlu dilaporkan adalah skor keseluruhan, skor verbal, dan skor non-verbal.

Penelitian ini menggunakan model sederhana sebagaimana Gambar 2b. Sementara masih memungkinkan terdapat model-model dengan variasi yang berbeda. Penelitian lanjutan perlu dilakukan dengan mengajukan model-model MIRT lain kemudian mengujinya dengan data riil hasil ujian TP.

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

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.

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

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

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.

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

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

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

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

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

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.

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.

Lee, S.-Y. (2007). *Structural Equation Modeling A Bayesian Approuch*. Chichester, West Sussex: John Wiley & Sons Ltd.

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

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.

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

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.

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. (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 The Annual Meeting of the American Educational Research Association, San Francisco, CA.

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.

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, 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.

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.

Yao, L. (2010). 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., & Boughton, K. A. (2007). A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification. *Applied Psychological Measurement, 31*(2), 83–105.

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.