PENGARUH JUMLAH KATEGORI, RENTANG THRESHOLD DAN METODE ESTIMASI TERHADAP AKURASI SKOR TES PADA BEBERAPA MODEL POLITOMI

Adiyo R[1]
Fakultas Psikologi UIN Jakarta

Abstrak

Tskor tes yang akurat merupakan salah satu tujuan utama yang ingin dicapai dari sebuah pengetesan. Namun umumnya masyarakat masih menggunakan cara klasik untuk memperoleh skor tes. Misalnya menjumlahkan hasil jawaban benar dari sebuah tes. Sebagaimana diketahui, terdapat beberapa asumsi pada tes klasik yang amat sulit dijumpai pada data skor tes. Maka itu pendekatan tes tidak lagi dilakukan dengan pendekatan klasik, tetapi dilakukan dengan pendekatan teori tes modern. Salah satu aspek yang dibahas dari teori tes modern ialah cara atau prosedur penskoran terhadap tes. Terdapat beberapa hal yang memengaruhi keakurasian skor tes pada pendekatan tes modern, yaitu dalam penelitian ini model politomi, jumlah kategori, rentang threshold dan metode estimasi. Model politomi (GRM dan GPCM), jumlah kategori (3 dan 4) dan rentang threshold (equal dan unequal) dipilih karena variabel tersebut berkaitan erat dengan data khususnya data kategorik. Data dalam penelitian ini berupa kategorik seperti skala Likert (0, 1, 2, 3...dst). Pada teori tes modern, penskoran dihitung berdasarkan susunan respon peserta tes terhadap butir item pada tes. Maka itu, diperlukan metode estimasi (MLR, WLSMV, EAP dan MLE). Penelitian ini menggunakan simulasi studi dengan kondisi 25 item, 500 responden dan 25 replikasi untuk setiap satu data. Interaksi variasi dari seluruh variabel independen sejumlah 2x2x2x4 menghasilkan 32 data. Variabel dependen dalam penelitian ini yaitu RMSE dan standard error. Analisis data menggunakan uji-F. Hasil penelitian menunjukkan bahwa pada kriteria RMSE, model data jumlah kategori 4 dan rentang threshold equal, serta dikalibrasi dengan model politomi GRM dan metode estimasi MLR dan WLSMV menghasilkan nilai RMSE terkecil dibandingkan dengan interaksi model data dan kalibrasi yang lain. sedangkan pada kriteria standard error, model data jumlah kategori 4 dan rentang equal, serta dikalibrasi dengan model GPCM dan metode estimasi MLE menghasilkan nilai standard error terkecil. Dari kedua kriteria tersebut, perbedaan nilai yang signifikan hanya pada kriteria RMSE. Untuk penelitian simulasi, kriteria RMSE lebih sensitif dalam menghasilkan keakurasian skor tes. Untuk aplikasi pada data empiris, model politomi GRM dan metode estimasi WLSMV atau MLR diduga lebih menghasilkan skor tes yang presisi.

Kata Kunci : Jumlah kategori, rentang *threshold*, metode estimasi, model politomi, akurasi skor tes

---

[1] Penulis adalah alumni Magister Sains Fakultas Psikologi UIN Jakarta Korespondensi tentang artikel ini dapat menghubungi : redaksi_jp3i@yahoo.co.id

Pendahuluan

Tes terdiri dari kumpulan butir soal atau item yang digunakan untuk memperoleh gambaran kepribadian, sikap, atau kognitif manusia . Berdasarkan respon orang terhadap butir soal atau item yang diujikan, maka diperoleh respon data. Adapun respon data tersebut berupa $F_{v\ x\ i}$, dimana F adalah variabel laten yang diukur melalui soal pada tes, v adalah banyaknya orang yang menempuh tes, dan i adalah banyaknya soal pada tes. Jika tes tersebut merupakan tes kemampuan atau mungkin tes prestasi atau potensi belajar maka data yang diperoleh berupa data dikotomi (0, 1), sedangkan jika tes tersebut merupakan tes kepribadian atau sikap, maka data yang dimiliki berupa data politomi (1, 2, 3,...,i_k, dimana k adalah banyaknya respon jawaban dari sebuah soal). Melalui respon data tersebut, terdapat banyak informasi yang dapat diperoleh baik mengenai soal maupun orang yang menempuh tes. Salah satu informasi utama yang dapat diperoleh yaitu informasi tentang skor tes orang. Dimana skor tes tersebut diasumsikan sebagai informasi mengenai letak atau posisi orang terhadap variabel yang diukur baik itu kepribadian, sikap maupun kognitif orang tersebut.

Terdapat dua pendekatan yang digunakan untuk menskor orang pada suatu tes, yaitu melalui pendekatan teori tes klasik atau *classical test theory* (CTT) dan pendekatan *item response theory* (IRT). Pada pendekatan CTT, umumnya skor tes seseorang diperoleh dengan cara menjumlahkan pilihan jawaban orang tersebut. Andaikan seseorang telah menempuh tes pilihan ganda dengan format jawaban benar salah (jika benar diskor 1, dan salah diskor 0), maka untuk menghitung skor orang tersebut, biasanya dihitung jumlah jawaban benar orang tersebut pada tes yang diujikan. Jika jawaban benar orang tersebut sebanyak 7 dari 10 soal, maka ia diberi skor 7. Atau 7/10 merupakan proporsi jawaban yang benar. Dengan kemungkinan skor yang ada yaitu 11 macam, mulai dari salah semua (0/10) sampai kepada benar semua (10/10). Cara tradisional seperti ini yang paling lazim digunakan untuk menskor orang, baik itu tes prestasi, tes kemampuan dsb. Mulai dari tes pada tingkatan kelas, hingga tes pada skala nasional sekalipun masih banyak yang menggunakan cara tradisional seperti itu. Bahkan beberapa ilmuwan psikologi atau psikolog yang menggunakan alat ukur berupa format skala Likert umumnya juga menggunakan cara tradisional untuk menskor orang, seperti menghitung atau menjumlahkan pilihan jawaban. Jika terdapat 20 item pernyataan dengan tiap pernyataan ada pilihan respon 'sangat tidak setuju=1' sampai kepada 'sangat setuju=4', maka skor orang adalah hasil penjumlahan pilihan respon dari item ke-1 s/d ke-20, sehingga diperoleh skor hasil penjumlahan orang tersebut. Hal ini menunjukkan bahwa betapa populernya cara penskoran dengan cara menjumlah-jumlahkan pilihan jawaban tiap item atau soal.

Cara tradisional untuk menskor orang dengan cara menjumlahkan tiap pilihan jawaban pada item lebih banyak berdampak merugikan terhadap hasil pengetesan, baik itu bagi orang yang dites maupun bagi tes itu sendiri. Bagi orang yang dites, tidak menutup kemungkinan bahwa apabila terdapat banyak orang menempuh tes yang sama akan menghasilkan skor yang sama persis, terlebih item hanya ada sedikit, misal 20. Andaikan seorang Guru akan meranking siswanya berdasarkan hasil tes prestasi, ternyata dari tes prestasi tersebut terdapat 15 siswa yang mendapatkan skor yang sama yaitu 20 dari 20 soal yang diujikan. Maka siapakah diantara 15 siswa tersebut yang menjadi ranking 1?. Atau sebaliknya, apabila seorang manajer perusahaan sedang menyeleksi karyawan dengan kuota untuk 10 orang, kriteria penerimaannnya adalah jawaban benar dari hasil tes kemampuan yaitu minimal 7. Ternyata ada 13 orang yang memiliki skor 7. Maka siapakah diantara 13 orang tersebut yang harus digagalkan sebanyak 3 orang oleh manajer perusahaan tersebut?.

Kemudian kerugian dari segi tes itu sendiri misalnya pada suatu tes kemampuan terdapat dua soal yang menanyakan konten yang agak berbeda, misalnya soal pertama yaitu (52 + 37=...), sedangkan soal kedua yaitu (52 + (0,40/20)=...). Tentu dari kedua soal tersebut, tingkat kesukaran soal yang kedua nampaknya lebih sulit daripada tingkat kesukaran soal yang pertama. Sedangkan jika menjawab benar pada kedua soal tersebut, sama-sama diberi skor 1. Padahal skor 1 dari soal yang kedua, memiliki makna skor yang berbeda daripada skor 1 dari soal yang pertama. Artinya tes tersebut tidak mampu membedakan kemampuan orang yang diperlukan untuk menjawab soal yang berbeda pula.

Maka itu cara tradisional untuk menskor orang dengan cara menjumlah-jumlahkan respon jawaban orang sudah seharusnya ditinggalkan. Sebab cara penskoran pada CTT, menskor orang tanpa memperhitungkan parameter tentang item, menyebabkan skor orang yang dihasilkan menjadi bias. Bahkan jika dilihat dari keakurasian skor, cara tradisional tersebut sama sekali tidak akurat untuk menskor orang. Padahal salah satu tujuan utama tes ialah menskor orang dengan hasil yang diperoleh seakurat mungkin. Dengan demikian, menskor orang dengan pendekatan CTT memang sudah seharusnya ditinggalkan dan diganti dengan cara penskoran yang lebih baik dan modern.

Dalam pendekatan *item response theory* (IRT), penskoran bukanlah difokuskan kepada skor tes, melainkan berdasarkan respon peserta terhadap soal atau respon item. Sebab IRT merupakan teori matematika tentang performa peserta tes terhadap item dan bagaimana hubungan antara kemampuan yang diukur oleh item

pada sebuah tes dengan probabilitas seseorang pada item tersebut (Hambleton dkk, 1991). Pada IRT unidimensional diasumsikan bahwa secara dominan hanya ada satu *trait* l a t e n y a n g m e n y e b a b k a n bervariasinya pola respon jawaban peserta tes. Kemudian hubungan performa orang terhadap sebuah item d i g a m b a r k a n m e l a l u i k u r v a k a r a k t e r i s t i k i t e m (*i t e m characteristics curve* atau ICC), dimana fungsi kurva tersebut berupa *monotone* atau membentuk huruf "S". Kurva ICC menunjukkan probabilitas bagi tiap orang terhadap masing – masing item. Bagi peserta tes yang memiliki kemampuan yang tinggi, maka memiliki peluang lebih tinggi untuk menjawab benar, sedangkan b a g i p e s e r t a y a n g m e m i l i k i kemampuan rendah, maka memiliki peluang lebih kecil untuk menjawab benar. Tentu *feature* yang demikian tidak terdapat pada CTT. Maka itu, IRT lebih banyak memiliki keunggulan dibandingkan CTT. Salah satu keunggulan IRT atas CTT ialah m e n g e n a i s k o r o r a n g y a n g *independent* terhadap tes yang ditempuhnya atau dengan kata lain *test independent* (Embretson & Reise, 2000). Pada IRT, meskipun orang menempuh paket tes yang berbeda, maka skor tes tersebut tetap dapat dibandingkan atau dapat ditentukan titik nol-nya (Hambleton dkk, 1991; Embretson & Reise, 2000). Dalam CTT, skor tes dapat dibandingkan hanya apabila asumsi paralelitas antar tes tercapai (Gulliksen, 1950, dalam Embretson & Reise, 2000). Namun kenyataannya, kecil kemungkinan untuk memperoleh setidaknya dua buah tes yang *strictly* paralel. Dikarenakan skor orang pada IRT bersifat *test independent*, maka terdapat beberapa hal yang dapat diteliti, salah satunya mengenai keakurasian skor orang IRT. Namun demikian terdapat beberapa hal yang mempengaruhi keakurasian skor orang pada IRT, yang penulis bahas yaitu model penskoran, format respon, rentang *threshold* dan metode e s t i m a s i . K e e m p a t h a l y a n g mempengaruhi keakurasian skor akan penulis bahas secara ringkas berikut ini.

D a l a m p e n d e k a t a n I RT setidaknya terdapat 3 model parameter yang lazim dikenal, yaitu 1 *parameter logistic* (1-pl) dimana hanya ada satu parameter yaitu tingkat kesukaran soal (disimbolkan dengan "b"), kemudian 2 *parameter logistic* (2-pl) yaitu model parameter tentang daya pembeda soal (disimbolkan dengan huruf "a") dan tingkat kesukaran soal, dan terakhir model 3 *parameter logistic* (3-pl) yaitu model parameter yang terdiri dari parameter tentang menebak atau *guessing* (disimbolkan dengan huruf "c"), daya pembeda soal, dan tingkat kesukaran soal. Seluruh parameter tersebut digambarkan melalui ICC. ICC ini yang mengekspresikan probabilitas seseorang terhadap suatu item berdasarkan karakteristik soal dan karakteristik orang. Sebagai

contoh, jika seseorang memiliki kemampuan yang sama dengan tingkat kesukaran soal, maka probabilitas menjawab benar untuk orang tersebut yaitu 0,5. Kemudian misalnya soal tersebut berupa pilihan ganda, apabila menjawab benar maka diberi skor 1 dan apabila salah diberi skor 0. Atau jika seorang peserta memiliki probabilitas menjawab soal secara benar diatas 0.5, maka diberi skor 1, dan sebaliknya.

Awalnya IRT muncul dengan tipe data 1 dan 0 atau dikotomi. Tipe data tersebut banyak digunakan pada konteks pendidikan dan intelegensi. Sedangkan pada bidang pengukuran kepribadian dan sikap, pola respon tidak hanya menggunakan kategori benar atau salah. Sebab dalam pengukuran kepribadian dan sikap tidak ada jawaban yang bersifat normatif, dalam hal ini jawaban benar atau salah, jawaban baik atau buruk. Oleh sebab itu kategori respon pengukurannya tidak lagi menjadi 1 dan 0, tetapi menjadi lebih banyak pilihannya. Atau istilah psikometris untuk kategori respon yang lebih dari dua yaitu *polytomous* atau *multiple response categories*. Embretson dan Reise (2000) menerangkan bahwa alasan digunakannya respon politomi yaitu kategori respon tersebut lebih informatif dan *reliable* dibandingkan hanya sekedar jawaban benar atau salah. Dikarenakan ada data yang berupa politomi, maka muncullah model penskoran IRT untuk data politomi. Model penskoran IRT politomi pertama kali muncul pada tahun 1969, dimana Fumiko Samejima menjelaskan tentang *the graded response model* (GRM) (dalam Embretson & Reise, 2000). GRM merupakan perluasan dari model 2-pl. Kemudian Geofferey Masters (1982) mengajukan model alternatif politomi IRT yang disebut dengan *the partial credit model* (PCM). Model tersebut merupakan perluasan dari Rasch model atau model 1-pl. Kemudian pada tahun 1992, Eiji Muraki (1992) membuat model yang lebih general dari model PCM sebelumnya, yaitu *the generalized partial credit model* (GPCM) . Dan masih terdapat beberapa model IRT politomi yang lain.

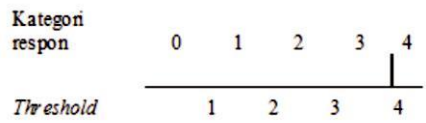
Model IRT Politomi

Pada penjelasan model IRT sebelumnya penulis telah menuliskan tentang model parameter IRT untuk data 1 dan 0 (dikotomi). Namun data yang ada pada tes umumnya tidak hanya berupa 1 dan 0, tetapi juga bisa berupa data politomi (1, 2, 3, ..., k). Tes yang memiliki tipe data seperti itu misalnya saja tes sikap, tes kepribadian dan soal essay. Maka itu, muncullah teori IRT untuk data politomi. Dan data yang penulis gunakan dalam penelitian ini ialah tipe data politomi tersebut. Maka itu penulis perlu membahas teori IRT pada tipe data politomi tersebut. Untuk selanjutnya, teori IRT untuk data politomi disebut sebagai model

penskoran IRT politomi. Pembahasan masing – masing model penskoran tersebut sebagai berikut.

Graded-Response Model (GRM)

Model penskoran *graded-response* umumnya digunakan untuk tipe data yang diperoleh berdasarkan hasil skor *graded* pada tiap itemnya (Samejima, 1969; Susan & Embretson, 2000). Misalnya penskoran terhadap hasil pengukuran sikap dapat diskor melalui model penskoran GRM. Pada model penskoran GRM, setiap soal hanya memiliki satu parameter a_i, sedangkan *threshold* (b_{ij}) (j = 1, ..., j_i) sebanyak $j_i - 1 = m_i$. *Threshold* merupakan ambang batas antar kategori . Menurut Susan dan Embretson (2000) terdapat dua tingkat untuk menghitung probabilitas kategori respon pada GRM. Andaikan ada sebuah soal dengan 5 kategori respon yaitu mulai dari 0 s/d 4 sehingga dapat ditulis $m = 5$, sedangkan $j = 1 \ldots 4$. Jika ditulis maka menjadi berikut ini:



Langkah pertama pada GRM untuk mengestimasi probabilitas orang maka perlu dihitung terlebih dahulu kurva untuk setiap soal. Berikut rumusnya (Susan & Embretson, 2000),

$$P_{ix}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]} \qquad 2.9$$

Keterangan:

$X = j = 1, \ldots, m_i$

α_i = *item discrimination*

β_i = *item difficulties*

Persamaan 2.9 menunjukkan probabilitas seseorang terhadap rentang *threshold* antar kategori respon ($j = 1 \ldots m_i$) yang mana probabilitas tersebut bergantung pula kepada letak atau posisi trait orang yang diukur (θ). Susan dan Embretson (2000) menjelaskan bahwa kurva P_{ix} disebut sebagai *operating characteristic curves*. Pada GRM setiap *threshold* harus memiliki masing – masing *operating characteristic curves*. Jika terdapat 5 respon kategori, maka terdapat empat parameter $\beta_{i\ j}$ yang diestimasi. Kemudian parameter β_{ij} tersebut diartikan sebagai nilai level trait yang diperlukan oleh responden untuk mendapatkan probabilitas diatas 0.50. Dan yang perlu diingat bahwa *constraint* pada GRM yaitu nilai parameter a atau *item discrimination* bernilai sama untuk tiap kategori pada sebuah item.

Setelah menghitung *operating characteristic curves*, kemudian menghitung probabilitas tiap kategori respon mulai dari 0 s/d 4. Maka itu akan ada 5 probabilitas kategori, sesuai dengan banyaknya kategori yang ada. Rumusnya yaitu sebagai berikut,

$$P_{ix}(\theta) = P_{ix}(\theta) - P_{i(x+1)}(\theta) \quad 3.0$$

Jika ditulis secara satu persatu kategorimakamenjadi,

$$P_{t0}(\theta) - 1.0 - P_{t1}(\theta)$$
$$P_{t1}(\theta) - P_{t1} - P_{t2}(\theta)$$
$$P_{t2}(\theta) - P_{t2} - P_{t3}(\theta)$$
$$P_{t3}(\theta) - P_{t3} - P_{t4}(\theta)$$
$$P_{t4}(\theta) - P_{t4} - 0$$

Secara definisi, probabilitas tertinggi untuk memilih kategori paling rendah ialah P_{i0} (θ) = 1.0, sedangkan probabilitas terendah untuk memilih kategori paling tinggi ialah P_{i5} (θ) = 0.0. Jika seluruh probabilitas untuk setiap respon kategori telah dihitung, maka dapat digambar melalui kurva tiap angka probabilitas kategori tersebut, selanjutnya kurva tersebut dinamakan dengan *category response curves* (Susan & Embretson, 2000). Parameter item pada GRM menunjukkan tentang bentuk dan lokasi kurva kategori respon dan kurva *operating characteristic curves*. Umumnya, semakin tinggi parameter slope item (α_i) semakin tegak pula kurva *operating characteristic*. Dan semakin menyempit pula kurva kategori respon . Hal tersebut menunjukkan bahwa kategori respon mampu membedakan level trait orang dengan cukup baik. Kemudian, *threshold* antar kategori menentukan lokasi kurva *operating characteristic*

.

***The Partial Credit Model* (PCM)**

Umumnya penskoran untuk *performance* seseorang berupa data 1 dan 0. Namun, menurut Masters (1982) ada model penskoran yang berurutan misal mulai dari 0, 1, 2, ..., *m*, yang mana m adalah banyaknya kategori. Masters (1982) menuliskan bahwa salah satu tipe data pada soal atau item untuk penilaian *performance* dapat berupa *ordered level* atau tingkatan yang berurutan. Maka itu penskorannya dilakukan dengan cara memberikan *partial credit* jika benar sebagian atau benar semua pada satu soal atau item tersebut. Model penskoran seperti ini umumnya ditemui pada soal dengan format jawaban essay. Melalui penskoran berupa *partial credit* tersebut, diharapkandidapatestimasi kemampuan secara lebih akurat dibandingkan tipe data yang hanya berupa benar atau salah. Masters (1982) memberikan contoh untuk tiga level penskoran untuk satu soal, yaitu,

$$\sqrt{\frac{7.5}{0.3} - 16 = ?}$$

Level kategori pertama yaitu jika seseorang mampu menyelesaikan pembagian 7.5/0.3 maka diberikan skor 1, kemudian level kategori kedua yaitujikaseseorangmampu menyelesaikan pengurangan 25 – 16, maka diberikan skor 2. Dan terakhir, level kategori ketiga jika seseorang mampu mendapatkan hasil dari √ 9,

maka diberikan skor 3. Kemudian pada setiap level kategori tersebut memiliki ambang batas (umumnya disebut *item step parameter*, disimbolkan dengan τ). *Step parameter* tersebut dapat diartikan sebagai tingkat kesulitan untuk tiap level kategori pada PCM. Jika sebuah soal memiliki 3 level kategori, maka terdapat 2 *step parameter* (τ_1 & τ_2) pada soal tersebut, yaitu *step* antara kategori ke-1 dengan kategori ke-2 disebut τ_1, kemudian *step* antara kategori ke-2 dengan kategori ke-3 disebut τ_2. Untuk setiap level kategori tersebut, dapat dihitung probabilitas tiap orang dengan *theta* tertentu yaitu θ_n, rumusnya adalah sebagai berikut:

$$P\{y_{ni} = 1 \mid \beta_n, \lambda_{ik} = \frac{\exp[\theta n - (\beta_i - \tau_{ik})]}{1 + \exp[\theta n - (\beta_i - \tau_{ik})]} \qquad 3.1$$

Keterangan:

θ_n adalah kemampuan atau theta orang ke-n

β_i adalah tingkat kesukaran soal/item ke-i

τ_{ik} adalah step parameter kategori k pada item i

Tingkat kesulitan antar kategori bersifat *ordered*, yang mana kategori terendah diasumsikan lebih mudah, sedangkan kategori tertinggi lebih sulit. Maka itu jumlah orang yang menjawab kategori *k* tidak akan lebih banyak daripada kategori *k-1*. Oleh sebab itu asumsi *ordered categories* pada model penskoran *partial credit* sama dengan asumsi pada model penskoran *graded response*. Artinya, untuk mencapai *step* kategori *k*, maka seseorang harus mencapai terlebih dahulu kategori *k – 1*. Sebab kategori tersebut bersifat *ordered* atau *sequence*. Jika semua probabilitas seseorang ditotal dari tiap masing – masing kategori, maka jumlah probabilitas tersebut yaitu 1, dan sebaliknya jika dikurangi dari 1 hingga ke tiap probabilitas kategori maka akan berjumlah yaitu 0. Dengan demikian, model penskoran *partial credit* ialah probabilitas orang ke-*n* untuk menskor kategori *k* pada item *i* berdasarkan lokasi seseorang yaitu β_n pada variabel yang diukur dan tingkat kesukaran pada item *i*.

The Generalized Partial Credit Model

Pada model PCM, diasumsikan bahwa *item slope* memiliki nilai yang sama untuk semua item atau soal pada satu skala. Namun Muraki (1992, dalam Embretson & Reise, 2000) membuat model penskoran hampir sama seperti PCM, tetapi item – item pada skala tersebut memiliki *item slope* yang berbeda-beda. Selanjutnya, model penskoran tersebut dikenal dengan *generalized partial credit model* (GPCM). Probabilitas orang untuk menskor kategori *k* pada item *i* yang ada pada GPCM hampir sama dengan probabilitas orang pada model penskoran PCM. Hanya saja, pada GPCM ditambahkan parameter *a* yaitu *item slope* atau daya pembeda item, sehingga rumusnya menjadi berikut ini

$$P(j|\theta_n, i_{jk}) = \frac{Exp\,\alpha_i[\theta n-(\beta_i-\tau_{ik})]}{1+Exp\,\alpha_i[\theta n-(\beta_i-\tau_{ik})]} \qquad 3.2$$

Keterangan:
Maksud dari β_i dan τ_{ik} sama seperti keterangan pada persamaan 3.1
α_i adalah item slope atau daya pembeda item

Berdasarkan persamaan 3.2 diatas, maka terdapat 1 parameter item yang bertambah pada model penskoran GPCM yaitu indeks α_i atau daya pembeda item. Interpretasi dari daya pembeda item tersebut tidak sama dengan interpretasi daya pembeda item pada model 2-pl. Hal ini disebabkan karena daya pembeda item pada model polytomus bergantung pada kombinasi parameter slope dan *category intersections* (Embretson & Reise, 2000). Kemudian intepretasi kategori τ_{ik} sama seperti interpretasi τ_{ik} pada model penskoran PCM, yaitu sebagai *category intersections* atau titik potong antar kurva tiap kategori.

Rating Scale Model (RSM)

Rating scale model atau RSM digunakan apabila seluruh item memiliki format respon yang sama dengan item – item lainnya atau bersifat *rating scale*, sehingga setiap item tersebut hanya memiliki satu parameter *item location*. *Item location* ini mencerminkan tentang tingkat kemudahan atau kesukaran item tersebut. Kemudian pada RSM setiap *threshold* (i_j) yaitu $j = k - 1$ dideskripsikan melalui kategori *intersection parameter* (δ_j) (Embretson & Reise, 2000). Disamping itu, menurut Embretson & Reise (2000) bahwa item yang diskor dengan model RSM memiliki asumsi yang sama seperti model PCM yaitu tiap itemnya memiliki daya pembeda yang sama dengan item lainnya dan skor mentah (*raw score*) merupakan data yang dapat digunakan untuk mengestimasi level *trait* atau kemampuan peserta tes.

Embretson & Reise (2000) menambahkan bahwa disamping memiliki kesamaan dengan PCM, RSM juga memiliki perbedaan dengan PCM yaitu pada RSM diasumsikan bahwa tingkat kesukaran kategori respon antar item relatif homogen. Bahkan sebetulnya secara *strict* Andrich (1978) mengatakan bahwa RSME merupakan *special case* dari model PCM, yaitu nila *step parameter* atau *location parameter* sama untuk tiap kategori pada seluruh item. Lebih lanjut lagi, Embretson & Reise (2000) mengatakan bahwa jika pada item matematika kemungkinan terdapat variasi mengenai tingkat kesukaran antar item dan antar kategori respon jawaban yang ada. Tetapi jika menggunakan format respon seperti 1= tidak setuju, 2 = agak setuju, dan 3 = sangat setuju, maka tingkat kesukaran kategori respon tersebut tidak begitu bervariasi. Model RSM memiliki rumus penskoran yang sama seperti PCM dan GPCM. Adapun rumus penskoran pada model RSM sebagai berikut:

$$P\{y_{ni} = 1 \mid \beta_n, \lambda_{ik} = \frac{Exp\ \alpha_i[\theta n - (\beta_i - \tau_{ik})]}{1 + Exp\ \alpha_i[\theta n - (\beta_i - \tau_{ik})]} \quad 3.3$$

Keterangan:
Keterangan seluruh elemen sama seperti persamaan 3.2

Dengan demikian RSM, PCM dan GPCM merupakan *nested models*. Yang mana model RSM memiliki asumsi yang lebih *strict* dibandingkan dengan model PCM dan GPCM. Ketiga model penskoran politomi tersebut merupakan *family rasch model*.

Rentang *Threshold* dan Jumlah Kategori

Dodd dan Koch (1985) pada penelitiannya menggunakan model penskoran PCM untuk tipe data politomi, dan mengevaluasi hasil *item information function* pada penskoran tersebut. Hasil penelitian mereka yaitu *item information function* pada model PCM dapat berbeda antar item dikarenakan *step estimates* atau rentang *threshold* pada item . Perbedaan bentuk *information functions* pada item merupakan efek dari rentangan *threshold* yang pertama dan terakhir pada suatu item. Item dengan rentangan yang kecil antara *threshold* pertama dan terakhir, akan menghasilkan informasi yang paling tinggi, tetapi hanya untuk rentang theta yang agak lebih sempit. Sedangkan item yang memiliki rentangan lebih besar antara *threshold* yang pertama dan terakhir, memiliki bentuk *information functions* yang lebih kecil atau rendah, tetapi untuk rentang theta yang lebih luas. Sebenarnya hampir pada semua kondisi bahwa *information functions* akan maksimum apabila mendekati dengan rentangan *step estimates* atau *threshold* pada suatu item. Disamping itu, Dodd & Koch (1987) mendapati bahwa item yang memiliki *threshold* sebanyak empat lebih banyak menghasilkan informasi, daripada item yang memiliki *threshold* sebanyak tiga. Artinya respon kategori sebanyak 5 lebih informatif dibandingkan dengan respon kategori sejumlah 4. Sedangkan dalam penelitian ini, penulis menggunakan 2 kategori respon yaitu sebanyak 3 dan 4 kategori. Untuk hal ini penulis berasumsi bahwa respon dengan 4 kategori akan memberikan informasi yang lebih tinggi dibandingkan dengan respon 3 kategori. Dengan sampel 500, 25 item dan 4 respon kategori, kemungkinan akan mencukupi untuk menghasilkan estimasi trait yang presisi. Kemudian temuan lainnya ialah bentuk *information functions* yang paling tinggi apabila *order estimates* mulai dari yang tersulit hingga ke yang termudah (Dodd & Koch, 1987).

Fung (2002) melalui disertasinya melakukan penelitian tentang *threshold distance* terhadap estimasi kemampuan orang. Variasi yang ia lakukan pada *threshold distance* yaitu *unequal-close at the low end*, *equal threshold* dan *unequal at the high end*. Urutan tersebut juga

menunjukkan hasil dari *recovery rates* yang tertinggi hingga ke yang terendah dan RMSE yang terendah hingga ke yang tertinggi. Namun perbedaan *recovery rates* dan RMSE yang dihasilkan kecil dan tidak signifikan. Fung menyatakan bahwa jarak antar *threshold* mempengaruhi keakurasian estimasi kemampuan orang. Selanjutnya, dengan empat kombinasi distribusi kemampuan (*ability distributions*), ketiga variasi *threshold* tidak menunjukkan perbedaan *recovery rate* yang signifikan, tetapi pada kriteria RMSE justru menunjukkan perbedaan yang signifikan. Ketika distribusi kemampuan bersifat normal (*normal distribution*) maka tidak terdapat perbedaan mean antar ketiga variasi *threshold*. Namun ketika distribusi kemampuan bersifat *skewed to the left*, maka terdapat perbedaan mean yang signifikan antar ketiga variasi *threshold* tersebut, urutannya sesuai dengan yang telah penulis sebutkan. Tetapi ketika distribusi kemampuan bersifat *skewed to the right* dan bimodal, maka urutan terendah hingga ke terkecil pada variasi *threshold* menjadi *unequal at the high end, equal threshold* dan *unequal-close at the low end*. Sebagai ikhtisar, pengaruh *threshold configuration* menandakan bahwa estimasi kemampuan akan menjadi kurang akurat apabila kategori antar *threshold* saling berdekatan pada rentang kemampuan yang bagian bawah (*lower end of the ability continum*).

Format respon kategori umumnya dibahas sebagai bagian dari *psychometric properties* yaitu misalnya estimasi parameter item, model fit, dsb (Rodriguez, 2005; Fabiola, dkk, 2005; Comrey & Montag, 1982). Misalnya saja Comrey dan Montag (1982) meneliti tentang analisis faktor skala "Comrey Personality", yang mana mereka menggunakan 7 dan 2 respon kategori. Berdasarkan hasil penelitian mereka ditemukan bahwa hasil *loading factor* baik item maupun 8 dimensi *Comrey's personality* lebih tinggi dengan menggunakan format respon 7 kategori dibandingkan dengan format respon 2 kategori. Secara rata-rata, *loading factor* untuk format 7 kategori yaitu 0.52, sedangkan untuk format 2 kategori rata-ratanya yaitu 0.44. Hasil yang sama juga ditemukan oleh Symonds (1924, dalam Comrey & Montag, 1982), ia menemukan bahwa reliabilitas *rating scales* meningkat sesuai dengan banyaknya kategori hingga 7 format. Menurut Comrey dan Montag, dikarenakan skala kepribadian berupa *self-rating* maka responden akan semakin mudah untuk menilai dirinya apabila diberikan pilihan respon yang cukup beragam, dalam hal ini misalnya 7 kategori. Dan juga, reliabilitas *self-rating* akan meningkat sesuai dengan banyaknya respon kategori. Pada pendekatan teori klasik, jika estimasi reliabilitas tinggi, maka *standard error of measurement* akan semakin rendah. Dengan kata lain, tes yang memiliki format respon cukup beragam, akan meningkatkan

reliabilitas tes tersebut, sehingga estimasi skor menjadi semakin akurat. Hal ini menjadi sangat penting untuk pengembangan tes yang memiliki data berupa *polytomous score,* khususnya tes kepribadian atau jawaban essay. Dengan menggunakan pendekatan IRT, maka seharusnya estimasi skor menjadi semakin akurat dengan adanya kombinasi format respon kategori, rentang *threshold* dan *standard error of measurement* yang berbeda untuk *trait level* yang berbeda pula.

Terdapat beberapa penelitian IRT yang membahas format respon kategorikaitannyadengan *psychometric properties* misalnya Hernandez, dkk (2000, dalam Fabiola 2012) mendapati bahwa apabila semakin banyak respon kategori jawaban pada model GRM maka diperlukan algoritma yang sangat besar agar mendapatkan estimasi model yang konvergen. Melalui penelitiannya, Hernandez dkk menyarankan bahwa sebaiknya digunakan 6 respon kategori yang paling maksimal agar mendapatkan kriteria konvergen yang ideal. Fabiola dkk (2012) menelititentang banyaknya format respon kategori kaitannya dengan validitas konstruk alatukurtersebut. Mereka menggunakan kriteria validitas *internal structure* dan menguji hubungan skor variabel hasil instrumen dengan variabel lainnya. Berdasarkan hasil penelitiannya ditemukan bahwa format 5 kategori menghasilkan korelasi yang paling tinggi yaitu 0.77 dengan skala penilaian untuk orang tua dan 0.68 dengan skala hasil penilaian guru. Sedangkan format 7 kategori berada diposisi kedua untuk hasil korelasi dengan variabel lain (0.71, skor hasil orang tua; 0.62, skor hasil guru) dan format 3 kategori menghasilkan nilai korelasi terkecil yaitu 0.60 dengan skor hasil orang tua dan 0.53 dengan skor hasil guru. Kemudian untuk kriteria *informations functions* format 7 kategori menghasilkan *information functions* yang paling tinggi, kemudian selanjutnya format respon kategori 5 dan terakhir format respon kategori 3 yang menghasilkan *information functions* yang paling kecil. Kriteria terakhir, format kategori yang menunjukkan *best fit measurement model* yaitu 7 kategori dengannilaichi-square = 486 (p=0.89;df=526), kemudian kategori 5 dengannilaichi-square = 374 (p=0.87;df=406). Dan terakhir, kategori 3 menunjukkan model fit dengannilaichi-square = 276 (p=0.77;df=295). Fabiola dkk (2007) menyimpulkan bahwa pengukuran yang menghasilkan aspek psikometris ideal adalah kategori respon lebih dari 3. Namun banyaknya kategori tersebut tidak lebih dari 7 respon.

Dalam penelitian ini penulis menggunakan pola respon 3 dan 4 kategori, kemudian rentang *threshold* yang penulis gunakan ialah *equal* dan *unequal*. Berdasarkan hasil penelitian terdahulu mengenai rentang *threshold*

dan respon kategori (misalnya Dood & Koch, 1987; Fung, 2002) maka hipotesis yang penulis miliki untuk variabel rentang *threshold* dan format respon yaitu:

1a : kombinasi rentang *threshold* dan format respon yang menghasilkan estimasi trait paling akurat ialah kombinasi rentang *threshold* yang *equal* dengan 4 respon kategori.

Model Politomi dan Metode Estimasi

Sepanjang pengetahuan penulis, belum begitu banyak studi yang membahas mengenai model penskoran pada IRT politomi. Khususnya dengan model penskoran ordinal seperti GRM, PCM, GPCM dan RSM. Namun begitu, penulis mendapati beberapa literatur yang telah membahas mengenai model penskorankaitannyadengan parameter orang khususnya estimasi skor tes. Misalnya Reise dan Yu (1990) meneliti tentang estimasi kemampuan orang dan item yang diskor dengan model GRM dan diestimasi dengan pendekatan *marginal maximum likelihood.* Reise dan Yu membuat penelitian dengan menggunakan data simulasi yaitu 25 item dengan lima respon jawaban seperti skala Likert. Mereka menyimpulkan bahwa untuk mendapatkan estimasi parameter 25 item yang stabil pada model GRM maka setidaknya diperlukan sampel orang sebanyak 500 responden.

Kemudian Dodd (1984) (dalam Maydeu, dkk, 1994) meneliti tentang parameter item dan orang. Dodd menggunakan model politomi GRM, modifikasi GRM (M-GRM) yang mana parameter *a* dibuat konstan untuk seluruh item, dan terakhir model PCM. Estimator yang digunakan untuk item dan orang ialah *joint maximum likelihood*, serta data yang digunakan ialah perbandingan data empiris dan data simulasi. Kriteria evaluasi yang digunakan ialah mengkorelasikan estimasi parameter item dan orang untuk setiap design atau model. Dan juga mengevaluasi *test information functions* (TIF) yang dihasilkandaritiapmodel. Berdasarkan hasil korelasi seluruh parameteritemdanorang menunjukkan bahwa ketiga model penskoran memiliki korelasi yang cukup tinggi (tidak disebutkan besarannya). Namunjika menggunakan kriteria evaluasi *information functions*, diperoleh bahwamodelpertamaGRM menghasilkan informasi yang lebih sedikittentangorangjikadibandingkandengan model modifikasi GRM dan PCM.

Samejima (1996) melakukan penelitian mengenai *test reporting* berdasarkan penskoran *summed score* dan politomi. Dari penelitiannya ditemukan bahwa model penskoran politomi memberikan *test information* yang lebih tinggi daripada *summed score*. Dengan semakin tingginya *test information* tersebut maka semakin

akurat pula estimasi trait atau kemampuan yang dihasilkan. Lebih lanjut lagi, Samejima mengatakan bahwa penskoran politomi akan semakin akurat apabila model penskoran atau format respon yang digunakan tepat. Misalnya, format respon sebanyak 7- dan 3-respon. Tentu untuk format 7 respon lebih sulit setiap kategori terisi oleh responden dibandingkan dengan format 3 respon. Apabila setiap kategori respon terisi, maka estimasi trait menjadi semakin akurat. Hal yang sama juga ditemukan oleh Donoghue (1993) (dalam Alagoz, 2000). Akkerman (1998, dalam Fung 2002) meneliti tentang model penskoran yang menghasilkan keakurasian estimasi kemampuan. Ia mengatakan bahwa fokus utama pada IRT ialah mengenai penskoran orang. Data yang ia gunakan ialah data empiris dan model penskoran yang digunakan yaitu *graded, parallel* dan *sequential scoring*. Hasil yang ia temukan yaitu tiga tipe penskoran tersebut harus digunakan sesuai dengan tipe data politomi yang ada. Sebab ketiga model penskoran tersebut ternyata berfungsi bagus ketika memang data yang digunakan sesuai untuk masing-masing model penskoran yang ada. Secara lebih spesifik yaitu model *continuation-ratio* digunakan untuk tipe respon data yang berupa skoring *sequential*, kemudian *cumulative probability* cocok untuk penskoran yang berupa *graded*, dan terakhir *adjacent category* cocok untuk penskoran yang *parallel*. Terakhir, Akkerman menyimpulkan bahwa perbedaan model penskoran dapat mempengaruhi estimasi penskoran trait atau kemampuan.

Maydeu-Olivares dkk (1994) menguji bagaimanakah model fit IRT politomi dengan data 20 item dan 5 respon jawaban, serta sampel orang sebesar 1.053. Model IRT politomi yang digunakan ialah GRM, PCM dan Nominal Model. Estimator yang digunakan ialah *marginal maximum likelihood* (MML) pada Multilog 6 (Thissen, 1991), *generalized least squares* pada Liscomp (Muthen, 1987). Kriteria fit yang digunakan ialah nilai chi-square (χ^2) untuk setiap model yang diuji. Berdasarkan penelitian tersebut diperoleh hasil bahwa 1) *full information* pada GRM menunjukkan hasil yang terbaik dibandingkan dengan model yang lainnya, 2) pada seluruh estimator, GRM menunjukkan model fit yang terbaik dibandingkan dengan model lainnya yaitu PCM, dan Nominal Model, 3) dari model PCM dan Nominal Model, model yang paling banyak parameter yaitu Nominal Model menunjukkan fit statistik yang lebih baik daripada model yang lebih sedikit parameternya, yaitu PCM. Sedangkan Dodd dan Koch (1985) menemukan bahwa justru model penskoran PCM menghasilkan information function yang lebih tinggi dibandingkan model penskoran yang lainnya. Dengan dihasilkannya *information function* lebih tinggi tersebut, maka dapat disesuaikan

dengan level theta yang hendak diukur, sehingga hasil pengukuran menjadi semakin akurat.

Embretson dan Reise (2000) meneliti tentang perbandingan estimasi *true score* berdasarkan model penskoran GRM, PCM, GPCM, RSM dan *modified graded response model* atau M-GRM. Semua estimasi menggunakan pendekatan *maximum likelihood* kecuali M-GRM menggunakan estimasi EAP dan GRM menggunakan estimasi MAP. Hasil yang diperoleh ialah seluruh skor *latent trait* berkorelasi setidaknya ≥ 0.70, dan bahkan seluruh skor hasil estimasi IRT politomi berkorelasi tinggi dengan *raw scores*. Korelasi terendah dengan *raw scores* ialah model penskoran G-PCM yaitu 0.97. Menurut Embretson & Reise, model penskoran yang ada asumsi bahwa *raw score* dapat digunakan untuk mengestimasi *trait level examinee* yaitu RSM dan PCM merupakan transformasi *non-linear* dari *raw scores* itu sendiri. Maka itu jika terdapat korelasi antara *raw scores* dengan model penskoran yang tidak ada *slope parameter*, maka korelasi tersebut dapat diterima secara konseptual. Tetapi sebaliknya, jika ada korelasi antara *raw scores* dengan model penskoran yang ada *slope parameter*, maka korelasi tersebut tidak dapat diterima secara konseptual. Namun perlu dicatat bahwa, menurut penulis meskipun terdapat korelasi antara model penskoran IRT politomi dengan *raw scores* tetapi hasil skor dari *raw scores* tidak menunjukkan tingkat akurasi yang berbeda-beda. Artinya standar error pada *raw scores* berlaku sama untuk seluruh *trait level*. Hal ini berbeda dengan konsep IRT, dimana tiap *trait level* memiliki standar error yang berbeda-beda.

Alagoz (2000) juga meneliti tentang model penskoran PCM, GRM dan GPCM kaitannya dengan estimasi kemampuan peserta tes. Korelasi antara GPCM dengan GRM yaitu sebesar 0.998, yang mana korelasi tersebut lebih besar daripada korelasi antara GPCM dengan PCM. Dan korelasi PCM dengan GRM paling kecil yaitu sebesar 0.995. Hal ini wajar mengingat model GPCM dam GRM sama-sama mengakomodasi *item slope parameter*, sedangkan pada model PCM tidak ada *item slope*. Namun secara keseluruhan, model GRM menghasilkan korelasi yang paling besar dengan model-model IRT lainnya. Sedangkan berdasarkan kriteria *information functions*, model penskoran GPCM memberikan informasi paling maksimum untuk level *theta* bagian tengah ($0 \leq \theta \leq 1$), sedangkan theta bagian bawah model PCM menunjukkan informasi yang paling maksimum ($-3 \leq \theta \leq -1$), dan untuk theta level atas model GRM memberikan informasi yang paling maksimum ($1.5 \leq \theta \leq 3$). Fung (2002) dalam penelitiannya ia membahas mengenai perbandingan model politomi. Hasil penelitiannya ialah dengan dikombinasikannya konfigurasi *threshold* dan distribusi

ability maka model GPCM - 1 menunjukkan estimasi kemampuan yang paling baik dibandingkan dengan model NCM dan GPCM. Dengan demikian, model penskoran pada data politomi ikut mempengaruhi keakurasian estimasi kemampuan peserta tes. Hasil penskoran tersebut tidak hanya dikorelasikan, tetapi juga perlu untuk diteliti model mana yang menghasilkan estimasi trait paling akurat. Samejima (1969, 1976) telah meneliti juga mengenai GRM yang menghasilkan pengukuran secara cermat. Koch (1983) juga menggunakan GRM sebagai penskoran untuk tipe data Skala Likert. Kemudian Bock (1972) dan Thissen (1976) menunjukkan bahwa *nominal response model* memberikan informasi yang lebih banyak khususnya untuk level theta bagian bawah disuatu rentang kontinum, daripada model penskoran dikotomi.

Embretson & Reise (2000) menuliskan bahwa untuk tipe data politomi yang memiliki banyaknya kategori sama untuk seluruh item, maka lebih tepat menggunakan model penskoran PCM. Dan PCM memiliki asumsi bahwa *discrimination power* berlaku sama untuk semua item. Sedangkan pada model GPCM merupakan pengembangan dari PCM yaitu bahwa tiap item memiliki *discrimination power* yang berbeda-beda, sehingga pada GPCM terdapat parameter a untuk setiap item. Baik PCM maupun GPCM tepat digunakan pada respon kategori yang jumlahnya sama untuk seluruh item pada sebuah skala. Hal tersebut berbeda dengan model GRM, yang mana *threshold* atau *category intersection* antar item dapat berbeda-beda banyaknya. Disamping itu pula, tiap item memiliki *discrimination power* dan *threshold* yang berbeda-beda. Terakhir, model RSM memiliki postulat bahwa setiap item memiliki nilai *threshold* yang sama, tetapi hanya berbeda letak atau lokasi itemnya. Dan tiap item juga tidak memiliki parameter *discrimination power*. Selain itu, item yang diskor melalui RSM harus memiliki kategori respon yang banyaknya sama.

Perbandingan antar model penskoran IRT politomi menjadi perlu untuk diteliti dalam kaitannya dengan estimasi trait. Sebab agar dapat diketahui model penskoran mana yang menghasilkan estimasi *trait level* yang paling akurat . Bahkan model penskoran tersebut dikombinasikan dengan kondisi tes lainnya, seperti format respon, rentang *threshold*. Dengan begitu, akan diperoleh estimasi yang terbaik atas informasi mengenai kemampuan responden atau peserta tes.

Metode estimasi terhadap *trait level* yang akan penulis bahas ialah pendekatan *maximum likelihood* (ML). Penskoran pada pendekatan ML diperoleh dengan cara menemukan nilai θ maksimum bagi seseorang berdasarkan suatu pola respon tertentu (Hambleton, dkk, 1991; Reise, 2000). Melalui pendekatan ML ini skor theta

dihitung mulai dari *negative infinity* hingga *positive infinity* pada rentangan theta *continuum*, dan setiap nilai likelihood tersebut dapat dihitung berdasarkan pola respon tertentu. Apabila nilai likelihood tersebut telah dihitung untuk seluruh rentangan theta, maka seorang peneliti dapat menentukan nilai likelihood yang maksimum, sehingga diketahui nilai theta orang tersebut. Dengan catatan, bahwa pada saat mengestimasi skor theta orang, parameter item telah diketahui. Disamping itu, pada penskoran ML peneliti berasumsi bahwa terjadi *local-item independence*. Artinya probabilitas untuk memilih respon pada suatu item, tidak terpengaruhi oleh probabilitas memilih respon pada item lainnya. Dengan demikian, fungsi probabilitas antar item bersifat *multiplicative*.

Beberapa studi telah membahas mengenai keakurasian estimator ML (Drasgow, 1985; Mislevy & Stocking, 1989; Allen & Yen, 1979).

Drasgow (1989) menuliskan bahwa *marginal maximum likelihood estimator* (MMLE) lebih akurat daripada *joint maximum likelihood estimator* (JMLE) apabila sampel orang sedikit, namun kedua estimator tersebut menghasilkan parameter yang sama apabila sampel berjumlah sangat banyak. Disamping itu, Drasgow (1989) menemukan bahwa keakurasian MMLE meningkat seiring dengan meningkatnya jumlah sampel. Hal tersebut juga ditulis kan oleh Hambleton (1991) bahwa ketika estimator ML digunakan maka sebaiknya diaplikasikan pada sampel jumlah besar. Jika yang diestimasi adalah parameter item, maka sampel yang dimaksud adalah orang, dan sebaliknya. Namun Seong (1990) menambahkan bahwa estimasi theta berdasarkan MMLE atau pendekatan *likelihood* lainnya, akan semakin akurat apabila *quadrature point* semakin diperbanyak. Hasil penelitian Seong (1990) ditemukan bahwa RMSE dan *mean bias* semakin kecil dalam setiap kondisi *quadrature point* yang lebih banyak (20 poin).

Diao dan Reckase (2009) membandingkan metode estimasi *maximum likelihood* dan bayesian pada simulasi *computerized adaptive testing*. Dari hasil penelitian mereka ditemukan bahwa pada panjang tes baik yang berjumlah 20 dan 50 soal, estimasi *maximum likelihood* menghasilkan RMSE dan mean bias yang lebih besar dibandingkan dengan estimas bayesian. Lebih jauh lagi, apabila nilai *theta true* bernilai negatif, maka estimasi theta berdasarkan bayesian menghasilkan bias yang positif, sedangkan estimasi theta berdasarkan *maximum likelihood* menghasilkan bias yang negatif. Bahkan metode estimasi *maximum likelihood* memiliki keterbatasan utama yaitu tidak dapt mengestimasi theta apabila respon item berupa benar atau salah semua (untuk konteks politomi bersifat *all endorsed* atau *all not-endorsed*) (Hambleton, 1991; Embretson & Reise, 2000). Kemudian

untuk mengatasi permasalahan pada estimasi *maximum likelihood*, maka digunakan estimasi bayesian. Pada estimasi bayesian, peneliti menggunakan *prior information* untuk menghitung fungsi *likelihood* pada data respon. Dengan adanya *prior information* tersebut, maka akan memudahkan untuk mengestimasi theta responden. Hal ini telah dibuktikan dalam banyak penelitian, misalnya Linden dan Pashley (2010), Bock dan Mislevy (1982). Dengan demikian, berdasarkan informasi diatas maka penulis menyusun hipotesis mengenai pendekatan metode estimasi trait sebagai berikut:

Berdasarkan asumsi dan keunggulan masing-masing model penskoran, maka hipotesis yang penulis miliki yaitu bahwa:

1b : model GRM akan lebih presisi dibandingkan model GPCM. Kemudian, pada berbagai kondisi tes, metode estimasi bayesian (EAP) akan menghasilkan estimasi trait yang paling akurat dibandingkan dengan metode estimasi lainnya.

1c: kombinasi antara rentang *threshold equal* , jumlah kategori 4, model GRM dan metode estimasi EAP akan menghasilkan skor tes yang paling akurat dibandingkan dengan model kalibrasi lainnya.

Metodologi Desain Eksperimen

Untuk menjawab rumusan masalah pada bab 1 dan hipotesis penelitian pada bab 2, maka diperlukan data. Data tersebut dapat diperoleh melalui dua cara, yaitu pertama data dari lapangan dan yang kedua data dari hasil studi simulasi. Dikarenakan data yang diperoleh dari dilapangan tidak dapat dilakukan variasi (bersifat *ex post facto*), maka penulis memilih menggunakan data simulasi. Artinya setiap kondisi yang diteliti, diciptakan datanya lalu diuji pertanyaan dan hipotesis penelitiannya. Pada studi simulasi orang dapat menetapkan data sesuai dengan kondisi yang diinginkan. Kondisi tersebut ialah jumlah kategori x rentang *threshold* x metode estimasi x model politomi. Dari variasi tersebut diperoleh beberapa kondisi eksperimen. Setiap kondisi eksperimen penulis tetapkan nilai *true parameter* (nilai *true parameter* dapat dilihat pada lampiran 1 s/d 4). Untuk lebih rincinya tentang kondisi eksperimen yang dibuat dapat dilihat pada tabel 3.1 berikut ini :

Tabel 3.1 Desain Eksperimen

	3	4
equal	a	c
unequal	b	d

Keterangan: 3 dan 4 adalah jumlah kategori; eq=*equal threshold;* un=*unequal threshold;*

Pada tabel 3.1 diatas terdapat 4 desain eksperimen, yaitu desain a dengan kondisi *equal threshold* dan jumlah kategori 3, kemudian desain b dengan kondisi *unequal threshold* dan jumlah kategori 3, desain c dengan kondisi *equal threshold* dan jumlah kategori 4 dan terakhir desain d dengan kondisi *unequal threshold* dan jumlah kategori 4.

Pada seluruh kondisi eksperimen tersebut, penulis tetapkan responden sebanyak 500 dengan menempuh item sebanyak 25, serta replikasi sebanyak 25 kali. Kemudian pada setiap kondisi dilakukan estimasi kemampuan orang (dalam hal ini *theta*) sebanyak 25 kali replikasi. Selanjutnya skor *theta* tersebut yang akan menjadi unit analisis dalam penelitian ini.

Seperti yang telah dituliskan sebelumnya bahwa pada setiap desain eksperimen tersebut, dapat diestimasi kemampuan orang (*theta*). Estimasi kemampuan orang dilakukan oleh *statistical software* yaitu Mplus (Muthen & Muthen, 2006). Pada estimasi tersebut dapat divariasikan beberapa hal diantaranya yaitu metode estimasi dan model politomi. Metode estimasi divariasikan menjadi 4 yaitu metode estimasi *maximum likelihood with robust standard error* (MLR), *expected a posterior* (EAP), *weighted least square with full diagonal weight matrix* (WLSMV), *maximum likelihood estimation* (MLE). Kemudian model politomi divariasikan menjadi 2 yaitu *graded response model* (GRM) dan *generalized partial credit model* (GPCM). Dengan demikian terdapat estimasi sebanyak 4 x 2 yaitu 8 kali estimasi kemampuan orang. Namun demikian, dalam penelitian ini tidak terdapat estimasi kemampuan orang untuk variasi model GPCM dengan metode estimasi MLR dan WLSMV. Hal ini disebabkan belum ada *statistical software* yang dapat mengestimasi kemampuan orang pada variasi tersebut.

Prosedur Menciptakan Data

Untuk mendapatkan kondisi eksperimen yang diinginkan, penulis menggunakan *statistical software* yaitu Mplus (Muthen & Muthen, 2006). Didalam *software* Mplus terdapat studi Monte Carlo. Melalui studi Monte Carlo dapat diciptakan data sesuai dengan kondisi yang dinginkan. Disamping itu pula, pada studi Monte Carlo dapat dibuat satu *true score* yang berlaku untuk seluruh kondisi eksperimen. *True score* tersebut sangat penting dikarenakan menjadi skor acuan (kriteria) terhadap estimasi theta pada berbagai kondisi eksperimen.

Kriteria Evaluasi Estimasi

Terdapat dua kriteria yang penulis gunakan untuk menentukan keakurasian skor theta, yaitu pertama *root mean square error* (RMSE). RMSE adalah indeks tentang varians error dari rata-rata estimasi *theta* terhadap nilai *theta true*. Jika nilai RMSE kecil, maka artinya hasil estimasi *theta* akurat dan sebaliknya. Adapun rumus RMSE ini yaitu sebagai berikut:

$$\text{RMSE} = \sum_{j=1}^{n} \sqrt{\sum_{j=1}^{n} \frac{(\theta'_j - \theta_j)^2}{n}} \qquad 3.3$$

Simbol θ' merupakan *theta* hasil estimasi, sedangkan simbol θ merupakan nilai *theta true* bagi orang ke-*j*. Simbol n merupakan jumlah

responden yaitu sebanyak 500 responden. Nilai RMSE tersebut penulis hitung sebanyak replikasi dilakukan yaitu 25. Selanjutnya nilai RMSE tersebut yang dijadikan data untuk menjawab hipotesis penelitian. Pada studi simulasi umumnya menggunakan nilai RMSE sebagai kriteria.

Kriteria kedua yang penulis gunakan ialah standard error hasil estimasi *theta*. Interpretasi terhadap kedua kriteria tersebut tetap sama, jika nilainya kecil, maka artinya skor tes akurat dan sebaliknya. Kemudian untuk menganalisis data dan menguji hipotesis penelitian, penulis menggunakan uji-F terhadap masing-masing kondisi atau interaksi variabel independen dalam penelitian ini. Uji-F dihitung sebanyak 2 kali sesuai dengan banyaknya kriteria evaluasi yang penulis gunakan.

Interaksi Jumlah Kategori vs Rentang *Threshold* (Hipotesis 1a)

Pada tahapan ini penulis menguji interaksi antara rentang *threshold* yaitu *equal* dan *unequal* dengan jumlah kategori 3 dan 4 (2 x 2). Selanjutnya interaksi ini disebut sebagai variabel independen. Kemudian variabel dependen dalam analisis ini yaitu RMSE dan *standard error*. Namun khusus untuk *standard error*, tidak ada analisis metode estimasi wlsmv, sebab memang pada metode estimasi wlsmv tidak terdapat *standard error*. Berikut hasil analisis two way manova 2 x 2.

Tabel 4.5 Tabel Anova Rentang *Threshold* vs Jumlah Kategori

Source	df	Mean Square	F	Sig.
distance_*thresholds*	1	.003	12.3	.001
categories	1	.007	30.7	.000
distance_*thresholds* * categories	1	.002	6.5	.012

RMSE : Rsquared = ,300

Berdasarkan tabel 4.4 diatas pada variabel dependen RMSE, variabel rentang *threshold* memiliki nilai F=12.372 ($p<0.05$) yang artinya terdapat pengaruh yang signifikan rentang *threshold* terhadap nilai RMSE estimasi theta. Nilai selisih mean RMSE yang dihasilkan antara *equal* dengan *unequal* yaitu -0,010 ($\mu_{RMSE\ (equal)} - \mu_{RMSE\ (unequal)}$). Hal ini berarti bahwa secara rata-rata nilai RMSE estimasi theta yang dihasilkan rentang *threshold equal* lebih kecil daripada nilai RMSE estimasi theta yang dihasilkan oleh rentang *threshold unequal*. Kemudian pada kriteria RMSE, variabel jumlah kategori berpengaruh signifikan terhadap nilai RMSE yang dihasilkan dengan nilai F=30.759 ($p<0.05$). Nilai selisih mean yang dihasilkan antara kategori 3 dan 4 yaitu 0.016 ($\mu_{rnse\ (cat3)} - \mu_{RMSE\ (cat4)}$). Artinya yaitu berdasarkan nilai RMSE, jumlah kategori 3 menghasilkan nilai RMSE yang lebih besar dibandingkan dengan nilai RMSE pada jumlah kategori 4.

Terakhir, interaksi antara jumlah kategori dengan rentang *threshold* menghasilkan nilai F=6.55 ($p<0.05$) untuk variabel dependen RMSE. Artinya interaksi antara jumlah

kategori dengan rentang *threshold* berpengaruh secara signifikan. Hasil ini konsisten dengan hasil RMSE pada masing-masing variabel independen. Berdasarkan nilai selisih mean yang penulis peroleh pada variabel RMSE, rentang *threshold equal* menghasilkan nilai RMSE yang lebih kecil bila dibandingkan dengan nilai RMSE yang dihasilkan dari rentang *threshold unequal*. Hal tersebut berlaku baik pada kategori 3 maupun 4. Dan juga sebaliknya, pada rentang *equal* dan *unequal*, nilai RMSE kategori 4 jauh lebih kecil daripada nilai RMSE kategori 3. Untuk lebih jelasnya dapat dilihat pada tabel 4.6 berikut ini.

Tabel 4.6 Nilai RMSE hasil Interaksi Jumlah Kategori vs Rentang *Threshold*

Dependent Variable	categories	distance_*thresholds*	Mean
RMSE	cat 3	equal	,370
		unequal	,387
	cat 4	equal	,361
		unequal	,364

Selanjutnya penulis menguji efek interaksi jumlah kategori dan rentang *threshold*, dengan variabel dependen *standard error*. Hasilnya dapat dilihat pada tabel 4.7 berikut ini.

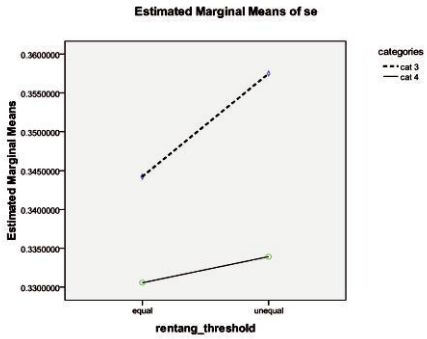
Tabel 4.7 Tabel Anova Jumlah Kategori vs Rentang *Threshold* terhadap *Standard error*

Source	df	Mean Square	F	Sig.
rentang_*threshold*	1	,002	,614	,435
categories	1	,009	3,081	,082
rentang_*threshold* * categories	1	,001	,219	,641

R squared = ,039

Berdasarkan tabel diatas, variabel rentang *threshold* memiliki nilai F = 0.614 (p>0.05), sedangkan variabel jumlah kategori memiliki nilai F = 3.081 (p>0.05). Artinya baik variabel rentang *threshold* maupun jumlah kategori tidak berpengaruh signifikan terhadap nilai *standard error*. Begitupun juga dengan interaksi antara rentang *threshold* dan kategori tidak berpengaruh signifikan terhadap nilai *standard error* yang dihasilkan dari estimasi theta. Nilai F hasil interaksi kedua variabel tersebut yaitu 0.219 (p>0.050). Dengan demikian, jika menggunakan kriteria nilai *standard error*, variabel jumlah kategori dan rentang *threshold* tidak menghasilkan perbedaan signifikan *standard error* estimasi theta. Namun begitu secara nilai mean *standard error* yang penulis peroleh, nilai mean kategori 3 berbeda dengan nilai mean kategori 4, yaitu nilai mean standar error kategori 3 lebih tinggi daripada nilai mean kategori 4. Hal tersebut berlaku baik pada rentang *threshold equal* dan *unequal*. Perbedaan *standard error* dapat dilihat pada figur 6 dibawah. Dengan demikian dapat ditulis bahwa untuk menjawab rumusan masalah diatas, pengaruh jumlah kategori dan rentang *threshold* terhadap keakurasian skor theta bergantung pada kriteria yang digunakan. Apabila menggunakan kriteria nilai *standard error*, maka interaksi kedua variabel independen tersebut tidak signifikan. Namun jika menggunakan kriteria RMSE maka

terdapat pengaruh yang signifikan jumlah kategori dan rentang *threshold* terhadap keakurasian skor theta responden.



Figur 6. Mean *Standard error* berdasarkan Jumlah Kategori pada tiap Rentang *Threshold*

Interaksi Model Politomi vs Metode Estimasi (Hipotesis 1b)

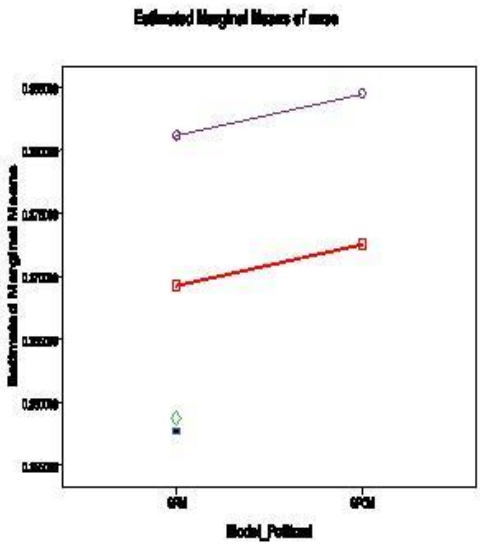
Selanjutnya penulis menguji pengaruh interaksi model politomi dan metode estimasi terhadap keakurasian skor theta. Variabel model politomi memiliki 2 variasi yaitu model GRM dan GPCM, sedangkan variabel model estimasi memiliki 4 variasi yaitu MLR, WLSMV, EAP dan MLE. Selanjutnya kedua variabel dan variasinya disebut sebagai variabel independen. Namun interaksi dari kedua variabel tersebut tidak sepenuhnya berjumlah 8 (2 x 4). Hal ini dikarenakan, pada software yang penulis gunakan baik Mplus maupun Parscale tidak ada kalibrasi untuk model GPCM dengan estimasi MLR dan WLSMV. Maka itu interaksi yang ada hanyalah berjumlah 6. Disamping itu juga khusus pada kriteria *standard error*, model estimasi WLSMV tidak memiliki nilai *standard error*, maka itu hanya ada 5 data interaksi pada kriteria *standard error*. Kriteria yang penulis gunakan masih sama yaitu nilai RMSE dan *standard error*. Hasil nilai RMSE pada interaksi model politomi dan metode estimasi dapat dilihat pada tabel 4.8 berikut ini:

Source	Dependent Variable	df	Mean Square	F	Sig.
Model_Politomi	RMSE	1	,000	,898	,345
Metode_Estimasi	RMSE	3	,003	12,02	,000
Model_Politomi * Metode_Estimasi	RMSE	1	,000	,000	,995

Berdasarkan tabel diatas, pada variabel dependen RMSE nilai F model politomi sebesar 0,898 (p>0,05). Hal ini artinya, berdasarkan nilai RMSE, model politomi baik GRM maupun GPCM tidak menghasilkan nilai RMSE yang berbeda secara signifikan. Atau dengan kata lain, tidak ada pengaruh yang signifikan dari model politomi terhadap nilai mean RMSE baik pada model GRM maupun GPCM. Kemudian, nilai F metode estimasi untuk variabel dependen RMSE sebesar 12,024 (p<0,05). Dari hasil kriteria tersebut dapat ditulis bahwa jika menggunakan kriteria RMSE, metode estimasi berpengaruh secara signifikan terhadap nilai RMSE theta.

Artinya pada RMSE, metode estimasi MLR, WLSMV, EAP dan MLE menghasilkan perbedaan mean RMSE theta yang signifikan. Namun begitu hasil interaksi model politomi dengan metode estimasi tidak berpengaruh signifikan baik terhadap RMSE dengan nilai F sebesar 0,00 (p>0,05).

Untuk model politomi, pada selisih mean RMSE model GRM dan GPCM menghasilkan perbedaan yang signifikan. Nilai selisih mean RMSE antara model GRM dan GPCM sebesar -0,012. Hal ini artinya model GRM menghasilkan nilai RMSE yang lebih kecil daripada nilai RMSE yang dihasilkan oleh model GPCM. Kemudian untuk metode estimasi, secara rata-rata RMSE metode MLR dan WLSMV hanya selisih 0,001. Maka itu selisih mean RMSE kedua metode estimasi tersebut tidak signifikan. Kedua metode estimasi tersebut menghasilkan nilai RMSE terkecil. Yang menarik ialah jika nilai RMSE dibandingkan baik dari metode MLR dan WLSMV dengan nilai RMSE yang dihasilkan oleh EAP dan MLE, maka keduanya menghasilkan perbedaan mean RMSE yang signifikan. Kemudian nilai RMSE terkecil diikuti oleh nilai RMSE yang dihasilkan dari metode EAP, dan yang terakhir nilai RMSE terbesar dihasilkan dari metode estimasi MLE. Untuk lebih mudahnya dapat dilihat pada figur dibawah ini.



Figur 7. Mean RMSE berdasarkan Model Politomi vs Metode Estimasi. Ket: ◯ = metode MLE; □ = metode EAP; ◊ = metode WLSMV; ▬ = metode MLR

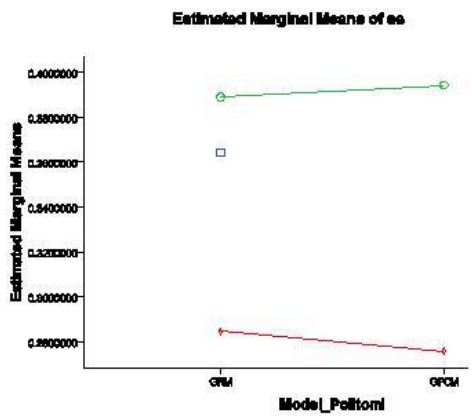
Selanjutnya penulis menguji interaksi antara model politomi dengan metode estimasi, tetapi yang menjadi kriteria adalah nilai rata-rata *standard error*. Hasilnya dapat dilihat pada tabel 4.9 berikut ini:

Tabel 4.9 Tabel Anova Model Politomis vs Metode Estimasi terhadap *Standard error*

Source	df	Mean Square	F	Sig.
Model_Politomi	1	,000	,325	,570
Metode_Estimasi	2	,130	691,470	,000
Model_Politomi * Metode_Estimasi	1	,001	5,375	,023

Rsquared = 0,937.

Berdasarkan tabel 4.9 diatas, pada kriteria *standard error* variabel model politomi memiliki nilai F sebesar 0,325 (p>0.05). Hal ini berarti bahwa variabel model politomi tidak berpengaruh signifikan terhadap nilai *standard error* yang dihasilkan. Kemudian variabel metode estimasi memiliki nilai F sebesar 691,470 (p<0.05), artinya variabel metode estimasi yaitu MLR, EAP dan MLE menghasilkan perbedaan yang signifikan pada nilai mean *standard error* dari theta. Selanjutnya, hasil interaksi antara model politomi dan metode estimasi diperoleh nilai F sebesar 5,375 (p<0.05). Artinya kedua interaksi variabel independen tersebut berpengaruh secara signifikan terhadap nilai *standard error* yang dihasilkan. Tentu hal ini menarik mengingat model politomi tidak berpengaruh signifikan terhadap nilai *standard error* untuk seluruh metode estimasi, namun sebaliknya metode estimasi berpengaruh signifikan terhadap nilai *standard error* pada seluruh model GRM dan GPCM. Artinya perbedaan metode estimasi yang digunakan untuk seluruh model politomi menghasilkan atau menentukan tinggi rendahnya nilai *standard error* pada theta. Hal tersebut juga terlihat dari nilai rsquare yang dihasilkan oleh variabel interaksi yaitu sebesar 0,937 atau 93,7% bervariasinya *standard error* theta disebabkan oleh variabel interaksi model politomi dengan metode estimasi. Untuk secara spesifiknya dapat dilihat pada figur berikut ini.



Figur 8. Mean *Standard error* Berdasarkan Model Politomi dan Metode Estimasi
Ket: ○ = metode estimasi EAP; □ = metode estimasi MLR; ◊ = metode estimasi MLE

Dari figur 4 diatas, terlihat bahwa pada seluruh kondisi model GRM dan GPCM, metode estimasi MLE menghasilkan nilai *standard error* yang kecil dibandingkan dengan metode estimasi MLR dan EAP. Perbedaan antara mean *standard error* dari metode estimasi MLE, MLR dan EAP signifikan yaitu nilai F keseluruhan sebesar 691,470 (p<0,05). Kemudian perbedaan mean *standard error* antara MLE dengan MLR yaitu menghasilkan nilai *mean difference* sebesar -0.084 (p<0,050), selisih *mean difference* MLE dengan EAP yaitu -0,112 (p<0,05) dan selisih *mean difference* EAP dengan MLR yaitu 0,027 (p<0,05). Kendatipun mean *standard error* yang dihasilkan oleh metode estimasi MLE paling kecil dan bahkan signifikan perbedaannya, namun hal ini bukan berarti menjadi jaminan bahwa

metode estimasi MLR dan EAP tidak akurat dalam mengestimasi theta. Sebab jika digunakan ukuran RMSE justru nilai mean RMSE pada metode estimasi MLE paling besar diantara yang lainnya. Tentu menurut penulis, ukuran RMSE sebetulnya lebih presisi sebab ada skor yang dijadikan pembandingnya yaitu *true score*.

Dengandemikianuntuk menjawab rumusan masalah no 2 dapat disimpulkan bahwa metode estimasi berpengaruh signifikan terhadap keakurasian skor theta baik menggunakan kriteria RMSE maupun *standard error*. Secara berurutan, metode estimasi yang secara konsisten menghasilkan keakurasian yang tinggi (berdasarkan RMSE dan *standard error*) yaitu MLR, WLSMV, EAP dan terakhir MLE. Kemudian model politomi tidak berpengaruh secara signifikan terhadap keakurasian skor theta baik menggunakan kriteria RMSE dan *standard error*. Namun berdasarkan kriteria RMSE model politomiGRMmenghasilkan keakurasianyanglebihbaik dibandingkan dengan model GPCM. Namun sebaliknya jika menggunakan kriteria *standard error*, justru model GPCM menghasilkan keakurasian yang lebih baik dibandingkan dengan model GRM.

Interaksi Jumlah Kategori, Rentang *Threshold*, Model Politomi dan Metode Estimasi (Hipotesis 1c)

Pada tahapan ini penulis menguji interaksi seluruh variabel independen dalam penelitian ini. Dari variabel jumlah kategori, rentang *threshold*, model politomi dan metode estimasi dihasilkan variasi berupa 2 x 2 x 2 x 4 sehingga total kondisi seharusnya menjadi 32 kondisi. Namun ada kondisi yang tidak ada yaitu metode estimasi WLSMV dan MLR pada metode politomi GPCM. Maka itu menjadi 30 kondisi. 30 kondisi ini selanjutnya disebut sebagai variabel independen. Kemudian variabel dependen dalam analisis ke-3 ini sama seperti dua analisis sebelumnya yaitu RMSE dan *standard error*. Adapun hasil pengaruh interaksi keempat variabel independen tersebut terhadap RMSE dan *standard error* dapat dilihat pada tabel 4.10 dan 4.11 berikut ini.

Tabel 4.10 Tabel Anova Hasil Interaksi 4 Variabel Independen

Source	DV	df	Mean Square	F	Sig.
rentang_*thresholds* * categories[a]	RMSE	1	,002	19,6	,000
	SE	1	,001	14.9	,000
rentang_*thresholds* * model_politomi[b]	RMSE	1	,001	13,0	,000
	SE	1	,001	8,0	,005
rentang_*thresholds* * metode_estimasi[c]	RMSE	3	,001	10,5	,000
	SE	2	,000	3,1	,043
categories * model_politomi[d]	RMSE	1	,002	19,4	,000
	SE	1	,000	3,86	,051
categories * metode_estimasi[e]	RMSE	3	,001	5,45	,001
	SE	2	,000	2,04	,132
model_politomi * metode_estimasi[f]	RMSE	1	,000	,000	,990
	SE	1	,002	31,5	,000
rentang_*thresholds* * categories * model_politomi[g]	RMSE	1	,001	7,75	,006
	SE	1	,000	2,64	0,10
rentang_*thresholds* * categories * metode_estimasi[h]	RMSE	3	,000	3,525	,016
	SE	2	,000	1,348	,262
rentang_*thresholds* * model_politomi * metode_estimasi[i]	RMSE	1	,001	10,7	,001
	SE	1	,000	7,74	,006
categories * model_politomi * metode_estimasi[j]	RMSE	1	,001	13,9	,000
	SE	1	,000	1,09	,298
rentang_*thresholds* * categories * model_politomi * metode_estimasi[k]	RMSE	1	,001	7,79	,006
	SE	1	,000	2,55	,112

Dari tabel diatas, secara ringkas dapat dituliskan sebagai berikut. Pada kriteria nilai RMSE, seluruh interaksi variabel independen menghasilkan perbedaan nilai RMSE yang signifikan ($p<0.05$), kecuali interaksi antara model politomi dengan metode estimasi (kondisi f). Kemudian pada kriteria *standard error*, interaksi variabel hanya berpengaruh signifikan pada kondisi a, b, c, f dan i. Sedangkan pada kondisi lainnya interaksi antar variabel tidak berpengaruh signifikan terhadap variabel *standard error*. Interaksi antara rentang *threshold* dengan jumlah kategori menghasilkan nilai F sebesar 19,698 ($p<0.05$) pada nilai RMSE. Kemudian interaksi antara rentang *threshold* dengan model politomi menghasilkan nilai F sebesar 13,082 ($p<0 . 05$) . Berdasarkan p e n g a m a t a n p e n u l i s t e r h a d a p *estimated marginal mean* (EMM) RMSE yang dihasilkan dari interaksi antara rentang *threshold* dengan model politomi, diperoleh bahwa pada seluruh rentang *equal* dan *unequal*, model GRM menghasilkan mean RMSE terkecil bila dibandingkan dengan model GPCM. Pada variabel dependen *standard error*, interaksi rentang *threshold* dengan jumlah kategori menghasilkan nilai F sebesar 14,998 ($p<0.05$). Kemudian, interaksi antara rentang *threshold* dan model politomi menghasilkan nilai F sebesar 8,024 ($p<0 . 05$) pada variabel dependen *standard error*. Berdasarkan nilai EMM, baik pada rentang *threshold* yang *equal* dan *unequal*, jumlah kategori 4 menghasilkan nilai *standard error* yang lebih kecil daripada jumlah kategori 3. Kemudian rentang *threshold equal* lebih menghasilkan nilai RMSE yang lebih kecil dibandingkan dengan rentang *threshold unequal*. Hal tersebut berlaku baik pada kategori 3 maupun 4. Selanjutnya interaksi antara rentang *threshold* dengan metode estimasi menghasilkan nilai F sebesar 10,514 ($p<0.05$) pada kriteria RMSE. Dan pada kriteria *standard error*, interaksi antara rentang *threshold* dengan metode estimasi memnghasilkan nilai F sebesar 3,195 ($p<0.05$). Hal ini berarti interaksi antara rentang *threshold* dengan metode estimasi signifikan dampaknya hanya pada nilai RMSE dan *standard error*. Dari grafik EMM nilai RMSE, penulis mendapati bahwa baik pada rentang *equal* dan *unequal* metode estimasi MLR dan WLSMV menghasilkan nilai RMSE terkecil. Kemudian diikuti oleh nilai RMSE dari metode estimasi EAP dan terakhir nilai RMSE terbesar dihasilkan oleh metode estimasi MLE. Pada seluruh metode estimasi tersebut nilai RMSE terkecil dihasilkan pada rentang *threshold* yang *equal*, sedangkan nilai RMSE pada rentang *unequal* agak lebih besar. Meskipun perbedaan tersebut signifikan ($p<0.05$), tetapi selisih mean RMSE antara *equal* dan *unequal* hanya 0.010. Selanjutnya pada nilai EMM *standard error*, rentang *threshold equal* menghasilkan nilai *standard error* yang lebih kecil dibandingkan dengan

nilai rentang *threshold unequal* pada seluruh metode estimasi. Dan estimasi MLE lebih kecil nilai *standard error*nya dibandingkan dengan metode estimasi MLR dan EAP.

Interaksi antara variabel jumlah kategori dengan model politomi menghasilkan nilai F pada kriteria RMSE sebesar 19,422 ($p<0.05$). Dan pada kriteria *standard error*, interaksi jumlah kategori dengan model politomi menghasilkan nilai F sebesar 3,86 ($p>0.05$). Dengan demikian, interaksi antara variabel jumlah kategori dengan model politomi signifikan pengaruhnya terhadap nilai mean RMSE. Dari grafik EMM RMSE, penulis mendapati bahwa baik pada jumlah kategori 3 maupun 4, model GRM memiliki nilai RMSE yang lebih kecil bila dibandingkan dengan model GPCM. Kemudian nilai RMSE pada kedua model tersebut lebih kecil pada jumlah kategori 4 dibandingkan dengan jumlah kategori

3. Pada jumlah kategori 3, selisih nilai RMSE antara GRM dan GPCM hanya sekitar 0,01. Namun pada jumlah kategori 4, selisih nilai RMSE antara GRM dan GPCM semakin besar yaitu sekitar 0,02. Dengan demikian secara sementara dapat disimpulkan semakin banyak jumlah kategori, semakin menurun pula nilai RMSE dan semakin lebar jaraknya antara model GRM dengan model GPCM. Kemudian interaksi antara jumlah kategori dengan metode estimasi menghasilkan nilai F sebesar 5,451 ($p<0.05$) pada kriteria RMSE, serta nilai F sebesar 2,046 ($p>0.05$) pada kriteria *standard error*. Dengan demikian, terdapat pengaruh yang signifikan interaksi jumlah kategori dengan metode estimasi terhadap nilai RMSE. Dari figur EMM RMSE, metode estimasi MLR dan WLSMV menghasilkan nilai RMSE terkecil, kemudian diikuti oleh metode estimasi EAP dan terakhir metode MLE. Hal tersebut berlaku baik pada kategori 3 maupun kategori 4. Selanjutnya interaksi antara model politomi dan metode estimasi menghasilkan nilai F sebesar 0,000 ($p>0.05$) pada kriteria RMSE, dan nilai F sebesar 31,510 ($p<0.05$) pada kriteria *standard error*. Dengan demikian, interaksi antara model politomi dan metode estimasi hanya berpengaruh signifikan terhadap nilai *standard error*. Namun pada kriteria RMSE, interaksi antara model politomi dan metode estimasi tidak berpengaruh signifikan. Hal ini bisa dikarenakan hanya metode estimasi EAP dan MLE yang dihitung nilai RMSE baik pada model GRM maupun GPCM. Sedangkan metode estimasi MLR dan WLSMV tidak ada nilai RMSE pada model GPCM, sehingga tidak dapat dibandingkan nilai RMSE dari metode MLR dan WLSMV pada model GRM dan GPCM. Namun jika melihat perbandingan RMSE antara metode estimasi EAP dan MLE baik pada model GRM dan GPCM, penulis mendapati bahwa nilai RMSE yang dihasilkan dari metode estimasi EAP lebih kecil nilainya daripada metode

estimasi MLE. Kemudian jika yang dibandingkan adalah model politomi, maka baik pada metode estimasi EAP dan MLE, nilai RMSE pada model GRM lebih kecil daripada nilai RMSE pada model GPCM.

Selanjutnya interaksi antara variabel rentang *threshold*, jumlah kategori dan model politomi terhadap RMSE dan *standard error* menghasilkan nilai F masing-masing sebesar 7,755 ($p<0.05$) dan 2,649 ($p>0.05$). Dengan demikian, interaksi ketiga variabel tersebut hanya berpengaruh signifikan terhadap variabel kriteria RMSE, sedangkan pada kriteria *standard error* tidak berpengaruh signifikan. Berdasarkan figur EMM RMSE, rentang *threshold equal* menghasilkan nilai RMSE yang lebih kecil apabila dibandingkan dengan nilai RMSE yang dihasilkan pada rentang *threshold unequal* baik pada jumlah kategori 3 dan 4 maupun pada model politomi GRM dan GPCM. Kemudian jumlah kategori 4 menghasilkan nilai RMSE yang lebih kecil bila dibandingkan dengan jumlah kategori 3. Hal tersebut terjadi pada seluruh kondisi rentang *threshold* dan model politomi. Terakhir, model politomi GRM menghasilkan nilai RMSE yang lebih kecil daripada model GPCM. Meskipun perbedaan mean RMSE yang dihasilkan hanya sekitar 0,01 tetapi perbedaan tersebut signifikan ($p<0.05$). Hal tersebut berlaku baik untuk rentang *threshold equal* dan *unequal* maupun pada jumlah kategori 3 dan 4. Seluruh pola interaksi ketiga variabel tersebut tidak penulis temui pada kriteria *standard error*.

Interaksi antara variabel rentang *threshold*, jumlah kategori dan metode estimasi terhadap RMSE dan *standard error* menghasilkan nilai F masing - masing sebesar 3,525 ($p<0.05$); 1,348 ($p>0.05$). Hal ini menunjukkan bahwa interaksi antara ketiga variabel independen tersebut berpengaruh secara signifikan terhadap RMSE. Sedangkan pada kriteria *standard error*, interaksi ketiga variabel tersebut tidak berpengaruh secara signifikan. Dari hasil figur EMM RMSE, interaksi antara rentang *threshold*, jumlah kategori dan metode estimasi menunjukkan bahwa interaksi antara rentang *threshold* yang *equal* dengan jumlah kategori 4 menghasilkan nilai RMSE yang ideal atau kecil bila dibandingkan dengan interaksi antara kategori 3 atau 4 dengan rentang *threshold* yang *unequal*. Hal tersebut berlaku pada seluruh metode estimasi. Namun demikian, jika berdasarkan metode estimasi pada seluruh kondisi interaksi antara rentang *threshold* dengan jumlah kategori, metode estimasi MLR dan WLSMV yang menghasilkan nilai RMSE terkecil daripada EAP dan MLE. Jika penulis amati secara detail, bahkan dengan rentang *threshold* yang *unequal* sekalipun, nilai RMSE yang dihasilkan oleh metode estimasi MLR ataupun WLSMV lebih kecil daripada nilai RMSE yang dihasilkan oleh

metode estimasi EAP ataupun MLE. Hal tersebut berlaku baik pada kategori 3 dan 4.

Interaksi antara variabel rentang *threshold*, model politomi dan metode estimasi menghasilkan nilai F pada masing-masing kriteria RMSE dan *standard error* sebesar 10,76 ($p<0.05$) dan 7,748 ($p<0.05$). Dengan demikian dapat diartikan bahwa interaksi antara rentang *threshold*, model politomi dan metode estimasi berpengaruh secara signifikan terhadap kriteria RMSE dan *standard error*. Dari figur EMM RMSE penulis mendapati bahwa rentang *threshold equal* menghasilkan nilai RMSE yang lebih kecil dibandingkan nilai RMSE dari *threshold unequal* pada seluruh kondisi model politomi dan metode estimasi. Perbandingan model GRM dan GPCM hanya dapat dilakukan pada metode estimasi EAP dan MLE. Dari figur EMM atas kondisi tersebut penulis mendapati bahwa pada metode estimasi EAP, nilai RMSE model GRM lebih kecil dibandingkan dengan nilai RMSE model GPCM. Hal ini berlaku baik pada rentang *threshold equal* maupun *unequal*. Namun kondisi tersebut berbeda ketika pada metode estimasi MLE, yaitu model GRM menghasilkan nilai RMSE yang kecil daripada model GPCM hanya pada kondisi *threshold equal*. Sedangkan pada kondisi *threshold unequal*, justru model GPCM yang mengasilkan nilai RMSE lebih kecil dibandingkan dengan nilai RMSE yang dihasilkan oleh model GRM.

Selanjutnya pada figur EMM *standard error*, penulis mendapati bahwa rentang *threshold equal* menghasilkan nilai *standard error* yang kecil dibandingkan nilai *standard error* pada rentang *threshold unequal*. Hal ini berlaku untuk model GRM dan GPCM, serta kondisi metode estimasi MLR, EAP dan MLE. Kemudian untuk model GRM menghasilkan nilai *standard error* lebih kecil daripada nilai *standard error* pada model GPCM. Hal tersebut juga berlaku untuk seluruh kondisi. Namun jika dilihat berdasarkan metode estimasi, justru metode estimasi MLE yang menghasilkan nilai *standard error* paling kecil diantara metode estimasi MLR dan EAP. Dan hal tersebut juga berlaku untuk seluruh kondisi model politomi dan rentang *threshold*.

Pengaruh interaksi antara jumlah kategori, model politomi dan metode estimasi terhadap nilai RMSE dan *standard error* menghasilkan nilai F masing-masing sebesar 13,988 ($p<0.05$) dan 1,092 ($p>0.05$). Dengan demikian variabel jumlah kategori, model politomi dan metode estimasi hanya berpengaruh signifikan terhadap nilai RMSE. Berdasarkan hasil figur EMM RMSE dari interaksi ketiga variabel tersebut, dapat dituliskan bahwa jumlah kategori 4 menghasilkan nilai RMSE yang lebih kecil dibandingkan dengan nilai RMSE dari jumlah kategori 3. Hal ini berlaku untuk seluruh kondisi model politomi dan metode estimasi.

Kemudian model politomi GRM menghasilkan nilai RMSE yang lebih kecil daripada model GPCM pada kondisi kategori 3 dan 4 dengan metode estimasi MLR, WLSMV dan EAP. Sedangkan pada kondisi metode estimasi MLE, model GRM menghasilkan nilai RMSE yang lebih kecil daripada model GPCM hanya pada jumlah kategori 4. Sedangkan pada jumlah kategori 3 justru model GPCM yang menghasilkan nilai RMSE lebih kecil dibandingkan nilai RMSE model GRM. Namun demikian, nilai RMSE pada metode estimasi MLE lebih besar daripada nilai RMSE pada metode EAP. Artinya meskipun kategori 3 memiliki nilai RMSE lebih kecil daripada kategori 4 pada metode estimasi MLE, namun kenyataanya kedua nilai tersebut lebih besar bila dibandingkan dengan metode estimasi EAP untuk kategori yang sama. Disamping itu, tentunya model politomi ikut menentukan besar kecilnya nilai RMSE tersebut.

Interaksi seluruh variabel independen yaitu jumlah kategori, rentang *threshold*, model politomi dan metode estimasi menghasilkan nilai F pada masing-masing kriteria RMSE dan *standard error* yaitu 7,791 ($p<0.05$) dan 2,555 ($p>0.05$). Artinya interaksi seluruh variabel independen tersebut berpengaruh signifikan terhadap nilai RMSE yang dihasilkan. Berdasarkan hasil tabel *comparisons mean* untuk semua level yang penulis peroleh, dapat ditulis sebagai berikut. Pada rentang *threshold equal*, kategori 3 menghasilkan nilai RMSE yang lebih besar daripada kategori 4. Hal ini berlaku baik untuk seluruh model GRM dan GPCM, serta metode estimasi MLR, WLSMV, EAP dan MLE. Kemudian pada kondisi rentang *threshold equal* dan jumlah kategori 3, metode GRM menghasilkan nilai RMSE yang lebih kecil daripada metode GPCM. Hal ini berlaku pada seluruh metode estimasi. Selanjutnya pada kondisi *threshold equal* dan model GRM, metode estimasi MLR menghasilkan nilai RMSE yang lebih kecil pada kategori 4 dibandingkan dengan nilai RMSE pada kategori 3. Pada kondisi yang sama, hasil yang sama juga terjadi pada metode estimasi WLSMV, EAP dan MLE. Dan juga pada kondisi *threshold equal* dan model GPCM, metode estimasi EAP dan MLE menghasilkan nilai RMSE yang lebih kecil pada jumlah kategori 4 dibandingkan dengan jumlah kategori 3.

Pada kriteria *standard error*, kondisi rentang *threshold equal* dan jumlah kategori 3, model GRM menghasilkan nilai *standard error* yang lebih kecil bila dibandingkan dengan nilai *standard error* pada model GPCM. Hal ini berlaku untuk seluruh metode estimasi. Kemudian seluruh metode estimasi menghasilkan nilai *standard error* yang kecil terjadi pada kondisi rentang *threshold equal*, dengan jumlah kategori 4 dan model politomi GRM. Untuk metode estimasi sendiri, metode yang menghasilkan nilai

standard error terkecil ialah metode estimasi MLE, kemudian MLR dan EAP. Sedangkan dari sisi jumlah kategori, pada seluruh model GRM dan GPCM, serta seluruh kondisi metode estimasi, jumlah kategori 4 menghasilkan nilai *standard error* yang lebih kecil bila dibandingkan dengan nilai *standard error* yang dihasilkan oleh jumlah kategori 3. Khusus pada model metode estimasi MLE dan model GPCM, nilai *standard error* yang dihasilkan lebih kecil daripada nilai *standard error* yang dihasilkan oleh model GRM dengan metode estimasi MLR dan EAP. Hal ini berlaku baik untuk jumlah kategori maupun jumlah kategori 4.

Selanjutnya pada rentang *threshold unequal*, jumlah kategori 3 menghasilkan nilai RMSE yang lebih besar bila dibandingkan dengan nilai RMSE yang dihasilkan pada jumlah kategori 4. Hal tersebut terjadi untuk seluruh model politomi dan metode estimasi. Kemudian pada rentang *threshold unequal* dan jumlah kategori 3, model GRM menghasilkan nilai RMSE yang lebih kecil bila dibandingkan dengan nilai RMSE yang dihasilkan oleh model GPCM.

Perbandingan ini hanya berlaku untuk metode estimasi EAP dan MLE. Sebab hanya kedua metode estimasi tersebut yang memiliki nilai RMSE baik pada model GRM maupun GPCM. Selanjutnya pada kondisi rentang *threshold equal*, jumlah kategori 3 dan model GRM, metode estimasi MLR dan WLSMV yang menghasilkan nilai RMSE paling kecil dibandingkan dengan nilai RMSE yang dihasilkan oleh metode estimasi EAP dan MLE. Sedangkan pada kondisi *threshold unequal* dan jumlah kategori 3, pada model GPCM metode estimasi EAP menghasilkan nilai RMSE yang lebih kecil daripada nilai RMSE pada metode estimasi MLE.

Pada kondisi rentang *threshold unequal* dan seluruh metode estimasi, model GRM menghasilkan nilai RMSE yang lebih kecil apabila dikalibrasi pada jumlah kategori 4 dibandingkan dengan kalibrasi pada jumlah kategori 3. Begitupun juga dengan model GPCM dengan kondisi yang sama. Terakhir, jika dibandingkan secara keseluruhan antara rentang *threshold equal* dan *unequal*, nilai RMSE rentang *threshold equal* lebih kecil daripada nilai RMSE rentang *threshold unequal*. Hal ini berlaku baik untuk seluruh model politomi, jumlah kategori dan metode estimasi baik pada rentang *equal* maupun *unequal*.

Pada kondisi rentang *threshold unequal* untuk kriteria nilai *standard error*, jumlah kategori 3 juga memiliki nilai *standard error* yang lebih besar dibandingkan dengan jumlah kategori 4. Hal tersebut berlaku untuk seluruh kondisi model politomi dan metode estimasi. Kemudian pada kondisi rentang *threshold unequal* dan jumlah kategori 3, model politomi GRM memiliki nilai *standard error* yang lebih kecil dibandingkan dengan model politomi GPCM. Namun

kondisi tersebut hanya berlaku pada metode estimasi EAP. Pada metode estimasi MLE, model politomi GPCM memiliki nilai *standard error* yang lebih kecil daripada model politomi GRM. Hal tersebut juga terjadi pada rentang *threshold unequal* dan jumlah kategori 4, yaitu yang mana model GRM memiliki nilai *standard error* yang kecil dibandingkan model GPCM hanya pada metode estimasi EAP. Sedangkan pada metode estimasi MLE, nilai *standard error* GRM lebih besar dibandingkan dengan nilai *standard error* GPCM. Baik untuk seluruh kondisi model politomi dan metode estimasi, nilai *standard error* yang lebih kecil diperoleh pada kondisi jumlah kategori 4 dibandingkan pada kondisi jumlah kategori 3. Perbandingan antara nilai *standard error* baik pada rentang *threshold equal* dan *unequal*, yaitu pada seluruh kondisi model politomi, jumlah kategori dan metode estimasi, nilai *standard error* pada rentang *threshold equal* lebih kecil daripada nilai *standard error* yang dihasilkan pada rentang *threshold unequal*.

Diskusi RMSE dan *Standard error*

Pada kriteria RMSE, seluruh hasil interaksi antar variabel berpengaruh signifikan terhadap nilai RMSE tersebut. Artinya perbedaan varian pada RMSE dapat dideteksi sebagai dampak dari interaksi antar variabel independen. Menurut Harwell dkk (1996) kriteria seperti RMSE merupakan kriteria yang cukup banyak digunakan untuk studi simulasi khususnya perbandingan *true score* dengan hasil estimasi. Namun begitu, untuk penelitian dengan data empiris perbandingan antara *true score* dengan hasil estimasi tidak dapat dilakukan. Oleh sebab itu kriteria yang digunakan untuk ukuran keakurasian pada data empiris ialah *standard error*. Menurut Umar (2013, *personal communication*) perbandingan antara *true score* dengan hasil estimasi lebih tepat jika disebut sebagai kriteria penyimpangan, sedangkan *standard error* memang digunakan untuk keakurasian skor tes. Disamping itu Asparouhov dan Muthen (2010) menuliskan bahwa *mean square error* merupakan indeks yang ideal untuk digunakan sebagai ukuran penyimpangan apabila memang terdapat *true score*, sehingga skor tersebut dapat dijadikan sebagai skor pembanding terhadap skor hasil estimasi.

Pada penelitian ini penulis tidak menggunakan kriteria lain yang dapat digunakan untuk perbandingan antara *true score* dengan hasil estimasi, misalnya *recovery rate* dan *ideal observer index* (IOI). Kedua kriteria tersebut belakangan ini cukup banyak digunakan oleh para peneliti terkait dengan perbandingan *true score* dengan hasil estimasi. Misalnya saja Asparouhov dan Muthen (2010), Levine dkk (1992), Fung (2002), Bastari (1998). Untuk ke depannya, kedua kriteria tersebut dapat dijadikan

acuan untuk menentukan penyimpangan dan atau keakurasian skor kemampuan peserta.

Rentang *Threshold* dan Jumlah Kategori.

Terkait dengan rentang *threshold* dan jumlah kategori pada hasil penelitian ini, penulis mendapati bahwa rentang *threshold equal* dan jumlah kategori 4 yang menghasilkan nilai RMSE dan *standard error* terkecil. Hasil penelitian ini agak berbeda dengan yang ditemukan oleh Dodd dan Koch (1987). Mereka menemukan bahwa *step values* (*threshold*) yang berupa *unequal with small range* menghasilkan *item information function* yang tinggi. Dan sebaliknya item *step unequal with large range* menghasilkan pengaruh terkecil terhadap *item information function.* Namun demikian terdapat perbedaan antara penelitian Dodd & Koch (1987) dengan penelitian ini, yaitu model politomi yang digunakan pada data penelitian mereka yaitu model *partial credit* (PCM). Menurut mereka, pada model PCM memungkinkan untuk sebuah *step values* bersifat *unordered*. Sebab data penelitian mereka adalah data politomi yang diperoleh dari *open-ended items*. Tentu *unordered step* tersebut tidak sesuai dengan model data yang penulis miliki yaitu *categorical-ordinal*. Asumsi penulis ialah data *categorical-ordinal* merupakan data dari skala Likert.

Fung (2002) pada disertasinya melakukan penelitian tentang *threshold distance* terhadap estimasi kemampuan peserta tes. Dalam penelitiannya, variasi pada *threshold distance* yaitu *unequal-close at the low end, equal threshold* dan *unequal at the high end*. Urutan tersebut juga menunjukkan hasil dari *recovery rates* yang tertinggi hingga ke yang terendah dan RMSE yang terendah hingga ke yang tertinggi. Namun perbedaan *recovery rates* dan RMSE yang dihasilkan kecil dan tidak signifikan. Fung menyatakan bahwa jarak antar *threshold* mempengaruhi keakurasian estimasi kemampuan orang. Selanjutnya, dengan empat kombinasi distribusi kemampuan (*ability distributions*), ketiga variasi *threshold* tidak menunjukkan perbedaan *recovery rate* yang signifikan, tetapi pada kriteria RMSE justru menunjukkan perbedaan yang signifikan. Menurut penulis alasan mengapa rentang *threshold equal* lebih presisi dibandingkan dengan rentang *threshold unequal* dapat disebabkan karena data distribusi baik parameter item maupun theta dibangkitkan pada kondisi distribusi yang normal. Oleh sebab itu, antara parameter item dan parameter orang berada pada kondisi distribusi yang sama. Maka itu salah satu saran penulis, penelitian lain dapat menggunakan interaksi antara rentang *threshold equal* dan *unequal* dengan kondisi distribusi theta yang *skewed,* atau bimodal dsb.

Mengenai jumlah kategori, kategori 4 menghasilkan nilai RMSE yang lebih kecil daripada nilai RMSE dari kategori 3. Hasil tersebut umumnya sesuai dengan hasil penelitian lain yang juga meneliti tentang banyaknya kategori terhadap keakurasian kemampuan peserta tes. Dodd dan Koch (1987) menuliskan bahwa format respon sejumlah 4 (*threshold* berjumlah 3) akan menghasilkan *information function* yang lebih optimum dibandingkan dengan format respon 3 (*threshold* berjumlah 2). Fabiola dkk (2012) juga menemukan bahwa dari format respon 3, 5 dan 7 ternyata format respon 5 jauh lebih baik secara psikometris khususnya pada reliabilitas dan validitas. Justru semakin banyak format respon (misal: 7) maka akan semakin memperkecil koefisien item, khususnya pada jumlah sampel yang kurang dari 500.

Model Politomi GRM dan GPCM

Pada hasil penelitian ini model GRM menghasilkan nilai penyimpangan yang lebih kecil dan akurat dibandingkan model GPCM. Hal tersebut terlihat dari kriteria RMSE dan *standard error*. Model GPCM unggul secara *standard error* dibandingkan model GRM hanya ketika dikombinasikan dengan metode estimasi *maximum likelihood* . Selebihnya model GRM yang lebih kecil menyimpang dan akurat dibandingkan model GPCM. Namun demikian perbedaan mean RMSE dan *standard error* yang dihasilkan dari kedua model politomi tersebut tidak berbeda secara signifikan. Hal ini dapat diartikan bahwa meskipun model GRM menghasilkan nilai RMSE dan *standard error* yang lebih kecil daripada model GPCM, namun secara statistik nilai tersebut tidak jauh berbeda. Hasil ini sama seperti yang ditemui oleh Wang (2002) dalam penelitiannya mengenai *computerized adaptive testing*. Wang menemukan bahwa pada berbagai metode estimasi (WLE, EAP, MAP, MLE) model politomi baik GRM dan GPCM menghasilkan nilai RMSE dan *standard error* yang sama. Sebab tidak ada pengaruh yang signifikan interaksi kedua model tersebut terhadap ketiga variabel dependen. Menurut penulis, hasil yang sama tersebut dapat terjadi pada model GRM dan GPCM. Sebab baik GRM dan GPCM sebetulnya sama-sama mengakomodasi parameter *item discrimination*, sehingga skor orang dihitung dengan mempertimbangkan parameter a tiap item tersebut. Hanya saja, umumnya GPCM lebih populer digunakan untuk tipe item yang sifatnya *open-ended* atau *essay* (Embretson & Reise, 2000). Dalam penelitian ini terdapat kekurangan mengenai interaksi antara model politomi dengan metode estimasi, yaitu tidak adanya interaksi antara model politomi GPCM dengan metode estimasi MLR dan WLSMV. Hal ini dikarenakan software yang penulis gunakan baik Mplus maupun

Parscale untuk saat ini belum mampu mengkalibrasi soal dengan kondisi tersebut. Oleh sebab itu, untuk penelitian kedepannya perlu dicari atau bahkan disusun sendiri cara untuk mengkalibrasi soal dengan kondisi tersebut. Disamping itu pula, model p o l i t o m i l a i n n y a j u g a d a p a t diikutsertakan sebagai variabel yaitu misalnya *rating scale model* (RSM) dan *nominal response model* (NRM).

Metode Estimasi

Hasil penelitian yang penulis peroleh mengenai metode estimasi yaitu metode estimasi MLR dan WLSMV menghasilkan nilai RMSE terkecil dibandingkan dengan metode estimasi lainnya. Namun begitu, dampak metode estimasi hanya berpengaruh secara signifikan terhadap nilai RMSE. Artinya metode estimasi MLR, WLSMV, EAP dan MLE hanya memiliki perbedaan keakurasian yang signifikan pada kriteria RMSE. Sedangkan pada kriteria *standard error*, metode estimasi menghasilkan nilai *standard error* yang lebih kecil dibandingkan dengan metode estimasi EAP dan MLE. Perbedaan nilai *standard error* yang dihasilkan oleh metode estimasi tersebut signifikan. Hasil penelitian Wang (2002) juga menemukan bahwa metode estimasi *weighted least square* menghasilkan nilai RMSE yang terkecil dibandingkan dengan metode estimasi lainnya yaitu EAP, MAP dan MLE. Muthen dkk (1997) mengatakan bahwa metode estimasi WLSMV sangat tepat apabila digunakan untuk data yang bersifat kategorik dan akan sangat stabil apabila digunakan pada sampel (orang) lebih dari 200.

Kemudian metode estimasi MLE dapat menghasilkan nilai *standard error* yang lebih kecil dibandingkan dengan metode estimasi MLR dan EAP. Menurut penulis, hal ini dapat terjadi sebab distribusi terhadap *latent variable* dalam penelitian ini mengikuti kurva normal (mean 0; sd 1) dan juga sampel dalam penelitian ini relatif cukup besar yaitu 500. Disamping itu pula, karena penelitian ini menggunakan studi simulasi, sehingga amat sedikit yang memiliki *abberant response*. Dari kondisi yang normal tersebut, maka itu MLE dapat bekerja sebagaimana mestinya. Namun apabila distribusi *latent variable* berupa *skewed* dan atau terdapat *abberant response*, maka nampaknya metode estimasi MLE tidak akan menghasilkan nilai *standard error* yang kecil seperti dalam penelitian ini. Oleh sebab itu, salah satu saran penulis yaitu untuk mendukung hasil studi simulasi, maka diperlukan juga data empiris. Agar dapat dilihat kesesuaian hasil kalibrasi pada data simulasi dengan data empiris dilapangan.

5.3 Saran

Saran disusun berdasarkan kekurangan penelitian ini dengan maksud untuk perbaikan penelitian

lainnya diwaktu yang akan datang. Adapun sarannya sebagai berikut:

1. Selain meneliti tentang parameter orang (theta responden), sebaiknya diteliti juga *item* atau *test information function*. Dari variabel dependen tersebut, tentu akan menghasilkan infromasi yang berbeda pula.
2. Disamping menggunakan kriteria seperti RMSE dan *standard error*, sebaiknya digunakan juga kriteria yang tidak kalah penting yaitu misalnya *ideal observer index* (IOI) dan *recovery rate*. Kedua kriteria tersebut masih jarang digunakan secara bersamaan khususnya untuk penelitian dengan data simulasi.
3. Disamping menggunakan data simulasi, sebaiknya sangat diperlukan juga data empiris. Dengan begitu, hasil antara data simulasi dengan data empiris dapat dibahas secara mendalam sesuai dengan kondisi data.
4. Agar lebih kaya analisis terhadap variabel dependen, maka perlu ditambahkan variabel independen seperti *prior distribution* terhadap *latent variable*, apakah itu normal, *skewed to the left* atau *skewed to the right*, *bimodal* dst. Kemudian dapat juga kondisi tes dibuat berupa *mixture types of item formats*, yang terdiri dari multiple choice, soal essay dan atau skala Likert.
5. Seiring dengan berkembangnya *multidimensional item response theory* (MIRT) dan *computerized adaptive testing* (CAT) saat ini, sebaiknya penelitian mengenai parameter item, kemampuan responden (theta), dsb lebih diarahkan kepada dua topik tersebut. Dengan begitu, hasil penelitian lebih menjadi aplikatif untuk penerapannya.

DAFTAR PUSTAKA

Alagoz, C. (2000). *Scoring tests with dichotomous and polytomous items*. Unpublished Master of Arts Thesis, Georgia.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. California: Wadsworth, Inc.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.

Asparouhov, T & Muthen, B.O. (2010). Plausible values for latent variable using Mplus. Mplus Handout at www.statmodel.com/download/Plausible.pdf

Bastari. 1998. Comparisons of IRT Models that handle dichotomous and polytomous response data simultaneuosly. Makalah dipresentasikan pada seminat NCME di UMASS.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal

categories. *Psychometrika, 37*, 29-51.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Comrey, L. A., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement*, 6, 3, 285 – 289.

Diao, Q., & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Dodd, B. G., & Koch, W. R. (1985). Item and scale informations functions for the partial credit model. *Paper presented at the annual meeting of the American Educational Research Association, Chicago*.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11*, 371-384.

Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19 – 29.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fabiola, G. B., Iwin, L., Jennifer, L. M., & Zaira, V. V. (2012). The effect of the number of answer choices on the psychometric properties of stress measurement in an instrument applied to children. *Evaluar, 12*, 43-59.

Fung, C. B. (2002). *Ability estimation under differential item parameterization and scoring models*. Unpublished Doctoral Dissertation, University of North Texas.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. California: Sage Publications, Inc.

Harwell, C. A., Stone, T. C., Hsu., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101 – 125.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 2, 149 – 174.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994).

Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, 18, 245-256.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muthen, B.O,. & Muthen, L. (1998 – 2010). Mplus version 6.1 student edition (Computer Software). Los Angeles, CA: Muthen & Muthen.

Muthén , B.O, du Toit, S., & Spisic,, D. (1997). *Robust infoerince using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthen, B.O., & Muthen. L. (2002). How to use a monte carlo study to decide on sample size and determine power. UCLA, Graduate School of Education & Information Studies.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta analysis of 80 years of research. *Educational Measurement: Issues and Practices*, *24*, 2, 3-13.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *6*, 255-270.

Seong, T.J. (1990). Sensitivity of marginal maximum likelihood estimation of item and abilty parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, 7, 456 – 461.

Thissen, D. M. (1976). Information in wrong responses to Raven Progressive Matrices. *Journal of Educational Measurement*, *13*, 201 – 214.

van der Linden, W.J., & Pashley, P.J. (2010). Item selction and ability estimation adaptive testing. In W.J van der Linden & A.W. Glas (Eds.). *Elements of adaptive testing* (hal. 3-30). New York: Springer.

Wang, T. (2002). Relative precision of ability estimation in polytomous CAT: A comparison under the Generalized Partial Credit Model and Graded Response Model. *Paper presented at the annual meeting of AERA Associations*, New Orleans.