Jurnal Penelitian dan Evaluasi Pendidikan
Volume 19, No 2, Desember 2015 (145-155)
Tersedia Online: http://journal.uny.ac.id/index.php/jpep

PERBANDINGAN ESTIMASI KEMAMPUAN *LATEN* ANTARA METODE MAKSIMUM LIKELIHOOD DAN METODE BAYES

Heri Retnawati
Pend. Matematika FMIPA UNY
retnawati_heriuny@yahoo.co.id

Abstrak

Studi ini bertujuan untuk membandingkan ketepatan estimasi kemampuan laten (*latent trait*) pada model logistik dengan metode maksimum likelihood (ML) gabungan dan bayes. Studi ini menggunakan metode simulasi Monte Carlo, dengan model data ujian nasional matematika SMP. Variabel simulasi adalah panjang tes dan banyaknya peserta. Data dibangkitkan dengan menggunakan SAS/IML dengan replikasi 40 kali, dan tiap data diestimasi dengan ML dan Bayes. Hasil estimasi kemudian dibandingkan dengan kemampuan yang sebenarnya, dengan menghitung *mean square of error* (MSE) dan korelasi antara kemampuan laten yang sebenarnya dan hasil estimasi. Metode yang memiliki MSE lebih kecil dikatakan sebagai metode estimasi yang lebih baik. Hasil studi menunjukkan bahwa pada estimasi kemampuan laten dengan 15, 20, 25, dan 30 butir dengan 500 dan 1.000 peserta, hasil MSE belum stabil, namun ketika peserta menjadi 1.500 orang, diperoleh akurasi estimasi kemampuan yang hampir sama baik estimasi antara metode ML dan metode Bayes. Pada estimasi dengan 15 dan 20 butir dan peserta 500, 1.000, dan 1.500, hasil MSE belum stabil, dan ketika estimasi melibatkan 25 dan 30 butir, baik dengan peserta 500, 1.000, maupun 1.500 akan diperoleh hasil yang lebih akurat dengan metode ML.

Kata kunci: *estimasi kemampuan, metode maksimum likelihood, metode Bayes*

THE COMPARISON OF ESTIMATION OF LATENT TRAITS USING MAXIMUM LIKELIHOOD AND BAYES METHODS

Heri Retnawati
Pend. Matematika FMIPA UNY
retnawati_heriuny@yahoo.co.id

Abstract

This study aimed to compare the accuracy of the estimation of latent ability (latent trait) in the logistic model using maximum likelihood (ML) and Bayes methods. This study uses a quantitative approach that is the Monte Carlo simulation method using students responses to national examination as data model, and variables are the length of the test and the number of participants. The data were generated using SAS/IML with replication 40 times, and each datum is then estimated by ML and Bayes. The estimation results are then compared with the true abilities, by calculating the mean square of error (MSE) and correlation between the true ability and the results of estimation. The smaller MSE estimation method is said to be better. The study shows that on the estimates with 15, 20, 25, and 30 items with 500 and 1,000 participants, the results have not been stable, but when participants were upto 1,500 people, it was obtained accuracy estimation capabilities similar to the ML and Bayesian methods, and with 15 items and participants of 500, 1,000, and 1,500, the result has not been stable, while using 20 items, the results have not been stable, and when estimates involve 25 and 30 items, either by participants 500, 1,000, and 1,500 it will obtain more accurate results with ML method.

Keywords: *estimation ability, maximum likelihood method, bayes method*

Pendahuluan

Dalam dunia real, berbagai fenomena tidak dapat diukur secara langsung. Fenomena ini dapat diukur, melalui serangkaian indikator yang tampak, baru kemudian diestimasi gejala ni menjadi suatu ukuran. Contohnya adalah fenomena intelegensi, kompetensi, kesetiaan, kemahiran, dan lain-lain. Untuk mengestimasi suatu kemampuan yang tidak tampak, seperti intelegensi, kesetiaan, kemahiran, dan lain-lain, diperlukan suatu indikator dari suatu perilaku atau suatu kompetensi peserta tes yang dapat diukur (*observable*). Indikator-indikator ini kemudian disusun menjadi suatu instrument yang kemudian digunakan untuk mengumpulkan respons peserta tes. Dengan menggunakan respons inilah, kemampuan laten dapat diestimasikan.

Dalam penilaian yang dilaksanakan dalam pendidikan, siswa menjawab butir soal suatu tes yang berbentuk pilihan ganda dengan benar, biasanya diberi skor 1 dan yang menjawab salah diberi skor 0. Pada penyekoran dengan pendekatan teori tes klasik, kemampuan siswa dinyatakan dengan skor total yang diperolehnya. Prosedur ini kurang memperhatikan interaksi antara setiap peserta tes dengan butir. Alternatif model penyekoran yang dapat digunakan adalah dengan menggunakan teori respons butir.

Pendekatan teori respons butir merupakan pendekatan alternatif yang dapat digunakan dalam menganalisis suatu tes. Ada dua prinsip yang digunakan pada pendekatan ini, yakni prinsip relativitas dan prinsip probabilitas (Keeves dan Alagumalai, 1999, p.24).

Berdasarkan prinsip ini, dapat disusun model logistik dengan menghubungkan antara probabilitas seseorang untuk menjawab benar dengan skala kemampuan (θ), tingkat kesulitan (b), daya pembeda butir (a), dan tebakan semu (*pseudo guessing*, c). Jika probabilitas dinyatakan dengan P, kemampuan dengan θ dan tingkat kesulita dengan b, maka hubungan keempat besaran tersebut dinyatakan dengan persamaan (Hambleton, Swaminathan, dan Rogers, 1991, p.12). Persamaan tersebut, yang merupakan model yang memuat 3 parameter butir yang kemudian disebut dengan model 3PL secara matematis dapat dinyatakan sebagai berikut (Hambleton, Swaminathan, dan Rogers, 1991, p.17; Hambleton dan Swaminathan, 1985, p.49; Van der Linden dan Hambleton, 1997, p.13).

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \qquad (1)$$

Model yang hanya memuat 2 parameter butir yaitu b dan a sering disebut dengan model 2PL, dan yang memuat 1 parameter butir yaitu b saja sering disebut dengan model 1PL. Model-model ini, yang menyatakan hubungan antara peluang terjadinya suatu fenomena dan kemampuan laten yang paling sering dipakai dalam psikometri, kesehatan, maupun pendidikan. Dengan model-model ini, kemampuan laten milik seseorang dapat diestimasi.

Ada dua metode estimasi kemampuan laten yang dapat digunakan, yakni metode maksimum likelihood dan Bayes. Beberapa pakar berpendapat bahwa dari kedua metode tersebut, belum diketahui metode mana yang lebih efektif dan pengguna maupun praktisi pengukuran yang menggunakan teori respons butir sering memperdebatkan pemakaian kedua metode tersebut. Terkait dengan hal tersebut, pada studi ini dibandingkan akurasi kedua metode tersebut untuk mengestimasi kemampuan laten. Studi ini bertujuan untuk mengetahui efektivitas perbandingan metode estimasi kemampuan dengan metode maximum likelihood (ML) dengan metode Bayes dilihat dari panjang tes dan mengetahui efektivitas perbandingan metode kemampuan ML dengan Bayes dilihat dari banyaknya peserta tes.

Agar informasi yang diperoleh berguna dalam penskoran tes, parameter butir perlu diestimasi. Estimasi parameter butir dan mengecek kecocokan model sering disebut sebagai kaliberasi butir. Kaliberasi ini dapat dilakukan jika data respons peserta terhadap tes telah diperoleh. Paling tidak

ada 2 pendekatan yang dapat digunakan untuk estimasi parameter butir atau melakukan kaliberasi butir, yakni esimasi Marginal Maximum Likelihood (MML) dan estimasi Marginal Maximum A Posteriori (MMAP) (Du Toit, 2006; Retnawati, 2014).

Dalam metode MML, nilai parameter butir dipilih yang dapat memaksimumkan logaritma dari fungsi marginal ma-ximum likelihood yang didefinisikan sebagai

$$\log L_m = \sum_{i=1}^{S} r_i log_e \overline{P}(x_i) \quad (2)$$

Dengan $r_i$merupakan frekuensi dari pola x_i yang diamati dari ukuran sampel sebesar N peserta dan S merupakan banyaknya pola yang berbeda. Pada model 3 parameter, kondisi maksimum diberikan dalam persamaan likelihood

$$\sum_{k=1}^{q} \left(\frac{\overline{r_{jk}} - \overline{N_{jk}} P_j(X_k)}{P_j(X_k)[1-P_j(X_k)]} \right) \frac{\partial P_j(X_k)}{\partial \begin{pmatrix} c_j \\ a_j \\ g_j \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad (3)$$

dengan

$$\overline{r_{jk}} = \sum_{l}^{S} r_l x_{lj} P(x_i|X_k) A(X_k) / \overline{P_{xl}}$$

$$\overline{N_k} = \sum_{l}^{S} r_l P(x_l|X_k) A(X_k) / \overline{P_{xl}}$$

Keduanya berturut-turut merupakan ekspektasi posterior dari banyaknya jawaban benar dan banyaknya usaha menjawab benar di titik X_k dan x_{lj} skor 0-1 butir ke-j pada pola ke-l.

Langkah untuk mengestimasi tersebut disebut algoritma E dan lanjutannya disebut algoritma M, sehingga keseluruhannya disebut dengan EM algoritma (Du Toit, 2006; Retnawati, 2014).

Pada tes klasik, kemampuan peserta tes diestimasi berdasarkan kemampuan menjawab benar. Pada teori respons butir, kemampuan diestimasi dengan fungsi non linear yang kemudian disebut dengan skor. Ada 3 metode estimasi yang sering digunakan, yakni estimasi maksimum likelihood, estimasi Bayes, dan estimasi Modal Bayes.

Likelihood maksimum dari skor kemampuan peserta diestimasi dengan memaksimumkan fungsi

$$\log L_i(\theta) = \sum_{j=1}^{n} \{ x_{ij} log_e P_j(\theta) + (1 - x_{ij}) log_e [1 - P_j(\theta)] \} \quad (4)$$

Dengan fungsi yang cocok dengan butir j.

Selanjutnya, diselesaikan persamaan implisit likelihood

$$\frac{\partial log L_i(\theta)}{\partial \theta} = \sum_{j=1}^{n} \frac{x_{ij} - p_{j(\theta)}}{P_j(\theta)[1-P_j(\theta)]} \cdot \frac{\partial P_{j(\theta)}}{\partial(\theta)} = 0$$

Estimasi $\hat{\theta}$ dihitung dengan metode penskoran Fisher, yang biasa disebut dengan Informasi dari Fisher, misalnya pada model 2 parameter memenuhi rumus:

$$I(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)] \quad (5)$$

Iterasi dari penyelesaian penskoran Fisher yakni

$$\hat{\theta}_{i+1} = \hat{\theta}_i + I^{-1}(\hat{\theta}) \left(\frac{\partial log L_i(\hat{\theta})}{\partial \theta} \right) \quad (6)$$

Kesalahan standar dari estimator MLmerupakan kebalikan dari akar kuadrat nilai informasi pada $\hat{\theta}$

$$\text{SE}\left(\hat{\theta}\right) = \sqrt{\frac{1}{I(\hat{\theta})}} \quad (20)$$

Estimasi Bayes merupakan rerata dari distribusi posterior $\boldsymbol{\theta}$, setelah diberikan pola respons peserta hasil tes x_i. Menurut Du Toit (2006) $\boldsymbol{\theta}$ dapat didekati secara akurat dengan persamaan:

$$\bar{\theta}_i = \frac{\sum_{k=1}^{q} X_k P(x_i|X_k) A(X_k)}{\sum_{k=1}^{q} P(x_i|X_k) A(X_k)} \quad (7)$$

Fungsi dari pola respons x_i sering disebut estimator dari *Expected a Posteriori* (EAP). Ukuran ketepatan dari $\bar{\theta}_i$merupakan standar deviasi posterior (*Posterior standard deviation*, PSD) yang didekati dengan

$$PSD(\bar{\theta}_i) = \frac{\sum_{k=1}^{q} (X_k - \bar{\theta}_i)^2 P(x_i|X_k) A(X_k)}{\sum_{k=1}^{q} P(x_i|X_k) A(X_k)} \quad (8)$$

Pembobotan $(A(X_k))$ pada formula tersebut didasarkan pada asumsi dari distribusi θ.

Estimasi dengan Bayes Modal mirip dengan Estimasi Bayes, namun dengan kesalahan rerata yang lebih besar. Estimasi ini sering disebut juga dengan Maximum a

Posteriori (MAP). Estimator MAP merupakan nilai θ yang memaksimumkan

$$P(\theta|x_i) = \sum_{j=1}^{n}\{x_{ij} log_e P_j(\theta) + (1 - x_{ij}) log_e[1 - P_j(\theta)]\} + log_e g(\theta) \quad (9)$$

Dengan $g(\theta)$fungsi kerapatan dari suatu distribusi populasi yang kontinu θ. Persamaan likelihoodnya yaitu

$$\sum_{j=1}^{n} \frac{x_{ij} - p_{j(\theta)}}{P_j(\theta)[1-P_j(\theta)]} \cdot \frac{\partial P_{j(\theta)}}{\partial(\theta)} + \frac{\partial \log_e g(\theta)}{\partial \theta} = 0 \quad (10)$$

Analog dengan estimasi maksimum likelihood, MAP dihitung dengan penyekoran Fisher dengan menggunakan informasi porterior

$$J(\theta) = I(\theta) + \frac{\partial^2 \log_e g(\theta)}{\partial \theta^2}$$

Pada kasus 2PL, dengan distribusi normal dari kemampuan θ yang memiliki varians σ^2, informasi posteriornya

$$I(\theta) = \sum_{j=1}^{n} a_j^2 \, p_j(\theta)[1 - P_j(\theta)] + \frac{1}{\sigma^2}$$

Dengan PSD dari estimasi MAP$\hat{\theta}$didekati dengan

$$PSD(\hat{\theta}) = \sqrt{\frac{1}{I(\hat{\theta})}} \quad (11)$$

Tentunya estimasi butir dan kemampuan dilakukan jika telah dibuktikan asumsi teori respons butir yaitu unidimesi, invariansi parameter butir dan kemampuan (Hambleton, Swaminathan, dan Rogers, 1991, p.17, Hambleton, dan Swaminathan, 1985, p.49, Retnawati, 2014). Asumsi independensi lokal tidak perlu dibuktikan jika telah memenuhi asumsi unidimensi. Asumsi unidimensi diketahui dengan melihat adanya faktor dominan dengan menggunakan analisis faktor, invariansi parameter butir dilihat dengan membuat plot parameter butir dari dua kelompok, dan invariansi kemampuan dilihat dengan membuat plot parameter kemampuan yang diestimasi dengan menggunakan separuh butir dan separuh butir lainnya (Retnawati, 2014).

Untuk menyelidiki sifat-sifat atau akibat penggunaan teori respons butir, sering digunakan teknik Monte Carlo. Teknik ini sangat populer dan penting dalam pemodelan respons butir. Dalam perkembangannya, berbagai studi dilakukan untuk berbagai kasus, misalnya berbagai panjang tes dan berbagai jumlah peserta tes, maupun data-data yang multidimensional. Terkait dengan studi ini, teknik Monte Carlo banyak digunakan dalam simulasi data dalam menerapkan teori respons butir (Harwell, et al., 1996). Studi Monte Carlo merupakan penelitian dengan desain seperti penelitian eksperimen. Studi ini juga merupakan studi yang dilakukan seperti keadaan di dunia riil, dengan data yang dibangkitkan dengan program komputer.

Aplikasi teknik Monte Carlo dalam teori respons butir, digunakan untuk mengevaluasi prosedur estimasi atau penemuan parameter, mengevaluasi sifat-sifat statistik pada teori respons butir, dan membandingkan metodologi yang terkait dengan teori respons butir. Studi Monte Carlo memiliki beberapa kelebihan. Studi ini dilaksanakan ketika solusi analitik untuk suatu permasalahan tidak ada atau tidak praktis karena kompleksitasnya. Selain itu, dalam suatu penelitian eksperimen kadang-kadang dikehendaki kesalahan standar kurang dari 15%. Metode Monte Carlo mempunyai kemampuan menyediakan nilai khusus dan dapat dimanipulasikan, dan merupakan cara yang adil untuk membandingkan tindakan-tindakan alternatif yang perlu biaya mahal jika melibatkan manusia. Namun demikian, metode ini mempunyai kelemahan memodelkan keaadaan seperti dunia riil. Bagaimana kondisi riil dimodelkan merupakan keterbatasan penelitian simulasi. Selain itu, pembangkit bilangan acak sulit diakses dan hasilnya memuat bias, karena hasil bisa bervariasi tergantung banyaknya replikasi, dan banyaknya iterasi jika menggunakan rantai akan mempengaruhi konvergensi hasil (Sinharay, 2004).

Agar data simulasi yang dibangkitkan seperti data yang ada di lapangan, pada simulasi perlu model data. Untuk mengurangi banyaknya replikasi yang dapat mempengaruhi hasil, banyaknya replikasi ditentukan sesuai dengan tujuan penelitian simulasi seperti yang dinyatakan Harwell, et

al. (1996). Pada penelitian tentang uji statistik, diperlukan 500 replikasi. Jika akan membandingkan metode terkait aplikasi teori respons butir, sejumlah kecil replikasi telah cukup, misalnya 10 kali.

Metode

Studi ini merupakan studi simulasi, dengan menggunakan data model respons siswa terhadap perangkat Ujian Nasional untuk mata pelajaran matematika di Sekolah Menengah Pertama/Madrasah Tsanawiyah (SMP/MTs) di DI Yogyakarta tahun 2006. Dipilihnya mata pelajaran matematika ini terkait bahwa matematika merupakan ilmu yang mendukung perkembangan teknologi. Dipilihnya jenjang SMP/MTs karena hasil ujian ini digunakan untuk seleksi masuk ke SMA/MA sehingga peserta tes mengerjakan tes dengan sebaik-baiknya dalam rangka mengukur kemampuannya. Pada studi simulasi, tes pilihan ganda dengan penskoran dikotomi dipilih sebagai objek simulasi, dengan model parameter butir dan kemampuan yang telah ditemukan pada studi pendahuluan.

Studi simulasi digunakan untuk menjawab kedua permasalahan yakni mengetahui pengaruh panjang tes dan pengaruh ukuran sampel peserta tes terhadap estimasi kemampuan menggunakan metode estimasi ML dan bayes. Untuk menjawab permasalahan ini, dilakukan studi simulasi dengan data rekaan yang dibangkitkan dengan kondisi yang telah ditentukan, seperti keadaan di lapangan. Studi seperti ini disebut studi simulasi Monte Carlo. Studi simulasi dipilih karena penelitian dengan studi simulasi dapat melibatkan sejumlah kasus yang diinginkan, dengan parameter seperti yang ada di lapangan. Selain itu, studi simulasi juga dapat memberikan sumbangan pada perkembangan teori respons butir yang lebih besar dibandingkan dengan menggunakan data riil.

Studi ini mendeskripsikan efek panjang tes, banyaknya peserta tes, penggunaan metode estimasi kemampuan terhadap ketepatan estimasi kemampuan. Ketepatan estimasi ditentukan dengan *mean square of error* (MSE) yang dihitung dengan setiap kasus pada r replikasi sebagai berikut:

$$MSE(\theta_r) = \frac{\sum_{i=1}^{n}(\theta_T - \theta_E)^2}{n} \qquad (12)$$

Dengan θ_T adalah kemampuan yang sebenarnya dan θ_E adalah kemampuan hasil estimasi, dan n banyaknya replikasi (Cohen, Kane, & Kim, 2001). Hasil ini diperkuat dengan mempertimbangkan pula korelasi kemampuan peserta tes yang sebenarnya dengan kemampuan hasil estimasi yang dihitung dengan korelasi *product-moment*.

Panjang tes yang digunakan pada studi ini adalah 15, 20, 25, dan 30 butir. Tes dengan panjang 15 dan 20 butir mewakili tes pendek dan tes dengan panjang 25 dan 30 butir mewakili tes panjang. Hal ini sesuai dengan pernyataan Mislevy & Bock (1990), tes dikelompokkan menjadi dua, tes panjang dan tes pendek. Tes pendek merupakan tes dengan banyaknya butir 11-20 dan tes panjang merupakan tes yang banyak butirnya lebih dari 20.

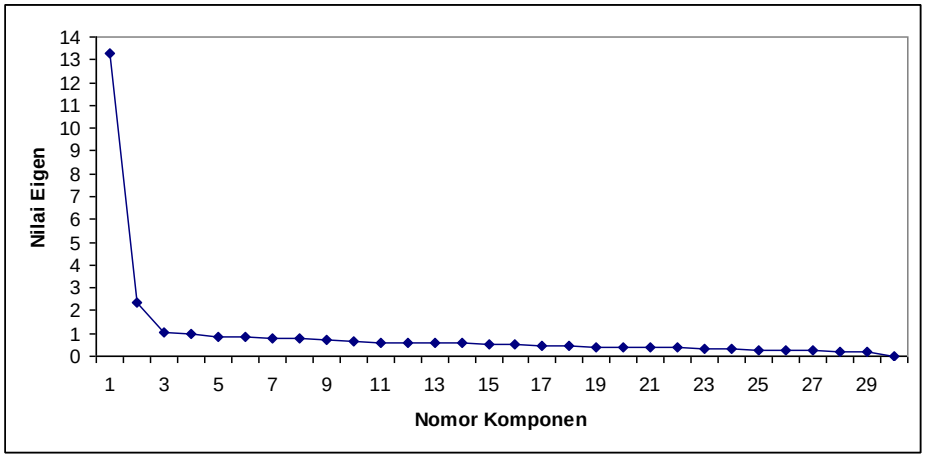
Pada studi ini, ada 3 ukuran banyaknya peserta tes yang diperhatikan, yakni 500 peserta, 1.000 dan 1.500 peserta. Variabel ini dipilih mempertimbangkan hasil penelitian Hambleton & Cook (1983) yang menunjukkan bahwa ukuran peserta tes juga stabilitas estimasi parameter butir, sedangkan hasil penelitian Segall (2000) menunjukkan bahwa pada teori respons butir multidimensi, panjang tes dan banyaknya peserta mempengaruhi estimasi parameter.

Pada studi ini, langkah-langkah yang dilakukan merujuk pada langkah-langkah penelitian simulasi Monte Carlo dalam teori respons butir menurut Harwell, Stone, Hsu & Kirisci (1997). Langkah-langkahnya yaitu: (1) menetapkan variabel yang menjadi fokus studi (panjang tes, banyaknya peserta tes, dan metode estimasi), (2) mendesain studi dengan menentukan banyaknya replikasi 40 kali, (3) memilih model teori respons butir dan menentukan distribusi parameter sesuai dengan data sebenarnya, (4) memilih program untuk membangkitkan data berupa sintaks makro dengan program SAS/IML merujuk pada makro yang ditulis Komrey, et al. (2006), (5) membangkitkan data, (6)

mengetimasi parameter kemampuan menggunakan program BILOGMG, dan selanjutnya (7) menganalisis hasil studi Monte Carlo, sesuai dengan tujuan studi, yakni mengestimasi kemampuan peserta tes, dalam hal ini menggunakan metode ML dan Bayes, yang hasilnya diperbandingkan untuk tiap kasus dengan melihat *mean square of error* (MSE) atau root *mean square of error* (RMSE) dari kemampuan siswa yang sebenarnya dan hasil estimasi. Analisis dilakukan secara deskriptif, sesuai dengan yang disarankan Harwell (1997). Analisis dilakukan untuk mengetahui pola MSE atau RMSE dengan menggunakan pendekatan teori respons butir pada tiap kasus yang dibandingkan, baik dengan perbedaan MSE dan korelasinya secara kuantitatif maupun dengan grafik.

Hasil

Agar dapat menggunakan pendekatan teori respons butir, terlebih dahulu dibuktikan asumsi unidimensi. Berdasarkan hasil analisis faktor eksploratori dengan menggunakan SAS/IML, dapat diperoleh *scree-plot* yang disajikan pada Gambar 1. Berdasarkan gambar ini, diperoleh penurunan nilai eigen dari yang pertama ke yang kedua sebesar 10,91 (sangat curam) yang menunjukkan bahwa perangkat ujian nasional matematika mengukur satu dimensi utama saja atau dapat dikatakan unidimensi. Karena memenuhi asumsi unidimensi, asumsi independensi lokal sekaligus terpenuhi.



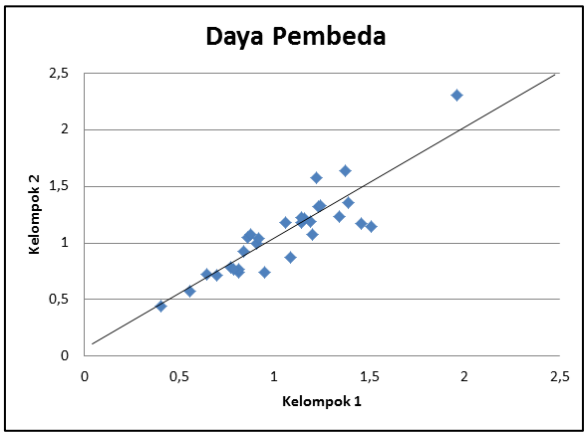
Gambar 1. Scree Plot Hasil Analisis Faktor Eksploratori

Dengan menggunakan plot kurva karakteristik butir dari model dengan titik-titik kuadratur pada estimasi parameter butir, dapat diperoleh perbandingan pada model 1PL, 2PL, dan 3Pl. Dari 30 butir soal, sebanyak 4 butir cocok dengan model 1PL, 6 butir cocok dengan model 2PL, dan 20 butir cocok dengan 3PL. Berdasarkan pertimbangan dengan statistik menggunakan plot ini, ditetapkan pada analisis studi ini dilakukan dengan model 3P.

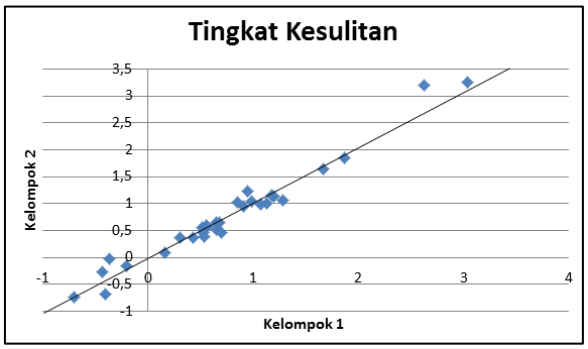
Dengan mengelompokkan peserta menjadi 2 kelompok secara acak, respons masing-masing kelompok terhadap butir-butir tes dianalisis untuk mengestimasi parameter a, b, dan c butir. Hasil estimasi tiap parameter dibuat *scater-plot*, kemudian dibandingkan dengan garis lurus. Berdasarkan Gambar 2 diperoleh bahwa titik-titik konvergen mendekati garis lurus, yang berarti bahwa daya pembeda bersifat invarian. Demikian pula halnya dengan parameter tingkat kesulitan dan tebakan yang disajikan pada Gambar 3 dan Gambar 4.

Selanjutnya, butir dikelompokkan menjadi 2, separuh pertama dan separuh kedua yang masing-masing dianalisis terpisah. Kedua kelompok butir ini kemudian digunakan untuk mengestimasi kemampuan laten (θ) peserta. Hasilnya kemudian dibuat *scater-plot* pada Gambar 5. Gambar tersebut menunjukkan bahwa titik-titik konvergen ke garis

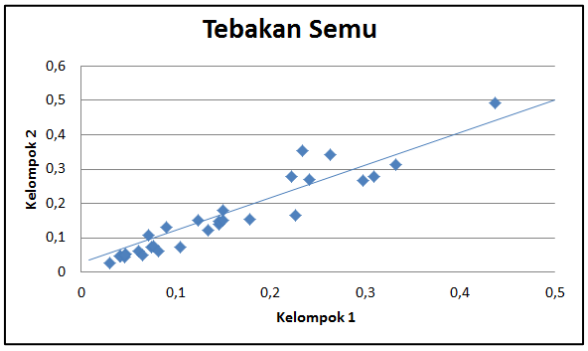
lurus, yang menunjukkan bahwa parameter kemampuan bersifat invarian. Hasil-hasil ini menunjukkan bahwa data respons peserta tes terhadap ujian nasional cocok dianalisis dengan model 3PL karena memenuhi asumsi unidimensi, independensi lokal, dan invariansi parameter baik butir maupun kemampuan.



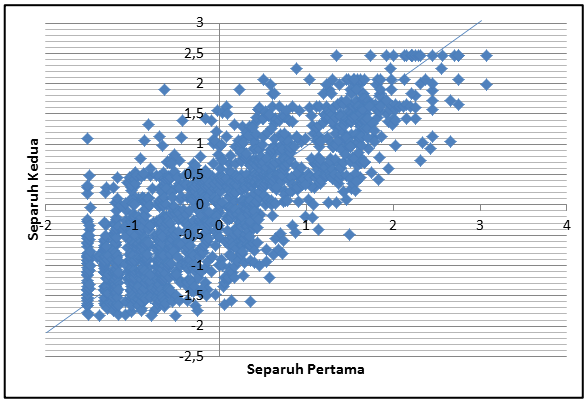
Gambar 2. *Scater-plot* Hasil Estimasi Parameter Daya Pembeda



Gambar 3. *Scater-plot* Hasil Estimasi Parameter Daya Pembeda



Gambar 4. *Scater-plot* Hasil Estimasi Parameter Tebakan Semu



Gambar 5. *Scater-plot* Hasil Estimasi Parameter Kemampuan

Dengan menggunakan model 3PL, data selanjutnya dianalisis untuk mengestimasi parameter sehingga karakteristik butir dapat diketahui. Demikian pula halnya dengan distribusi kemampuan peserta. Kemampuan peserta berdistribusi normal dengan rerata 49,449 dan varians sebesar 16,347. Adapun karakteristik butir disajikan pada Tabel 1.

Dengan bantuan 2 guru matematika yang senior, dosen pendidikan matematika, dosen matematika dan psikolog dalam forum FGD dilakukan pengurangan butir, dengan tetap memperhatikan keterwakilan isi dan karakteristik butir. Pengurangan butir tahap yang pertama sebanyak 5 butir (dari 30 butir menjadi 25 butir). Butir yang dikurangi yakni butir nomor 13 (Perbandingan—soal cerita), 15 (SPL—soal cerita), 5 (jaring-jaring kubus), 7 (sudut segitiga), 29 (trigonometri). Pengurangan butir tahap kedua sebanyak 5 butir lagi (dari 25 butir menjadi 20 butir). Butir yang dikurangi yakni butir nomor 1 (persentase—soal cerita), 6 (simetri lipat), 19 (refleksi), 21 (perbandingan segitiga), dan 9 (teorema Phytagoras). Pengurangan butir tahap ketiga sebanyak 5 butir (dari 20 butir menjadi 15 butir). Butir-butir yang dikurangi yaitu butir nomor 30 (logaritma), 8 (pemetaan), 25 (suku dan faktor), 22 (segitiga kongruen), dan 23 (juring lingkaran).

Selanjutnya, dengan menggunakan model-model data tersebut, dibangkitkan data dengan bantuan program SAS, dengan menggunakan makro dari Komrey, et al.

(2006). Dari makro ini, dimodifikasi nama file *output*, nama *file* masukan, variabel yang dilibatkan, banyaknya replikasi, banyaknya peserta, banyaknya dimensi kemampuan, dan format letak pada *file output*, sesuai dengan kasus yang diharapkan. Pada studi ini, ada 3 variabel yang menjadi perhatian yakni panjang tes (n=15, n=20, n=25 dan n=30), banyaknya peserta (N=500, N= 1000, dan N=1500), dan banyaknya dimensi (k=1, k=2 dan k=3). Proses pembangkitan data ini menggunakan model parameter butir hasil estimasi pada Tabel 1 dan juga distribusi kemampuan peserta.

Tabel 1. Karakteristik Perangkat Tes UN Matematika SMP 2006 Berdasarkan Teori Respons Butir Unidimensi Model 3 Parameter

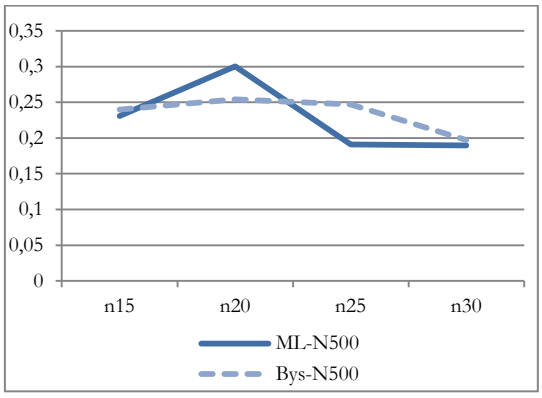
Butir	Materi	a	b	c	Keterangan
1	Persentase (soal cerita)	1,229	-1,242	0,017	Baik
2	Diagram Venn	0,996	-1,089	0,018	Baik
3	Persentase	0,667	-0,023	0,045	Baik
4	HP bil bulat	0,526	0,400	0,035	Baik
5	Jaring-jaring kubus	0,930	-2,552	0,144	Kurang baik (b<-2.0)
6	Simetri lipat	0,520	-1,927	0,070	Baik
7	Sudut segitiga	0,887	-1,034	0,500	Kurang baik (c>0,25)
8	Pemetaan	1,030	-1,048	0,026	Baik
9	Akar dan pangkat	1,083	-1,311	0,033	Baik
10	Sifat garis sejajar	0,677	-0,181	0,036	Baik
11	Keliling belah ketupat	1,321	-0,709	0,022	Baik
12	Luas Jajar genjang	1,104	-0,517	0,026	Baik
13	Perbandingan (soal cerita)	1,175	-1,999	0,159	Baik
14	Persamaan garis lurus	0,762	0,018	0,046	Baik
15	SPL (soal cerita)	1,150	-1,296	0,041	Baik
16	Median data	0,824	-0,397	0,019	Baik
17	Volume limas	0,868	-0,182	0,014	Baik
18	Luas permukaan prisma	1,135	-0,656	0,105	Baik
19	Refleksi	1,231	-0,640	0,255	Kurang baik (c>0,25)
20	Dilatasi	0,803	-0,741	0,094	Baik
21	Perbandingan segitiga	2,847	0,659	0,352	Kurang baik (c>0,25)
22	Segitiga kongruen	0,658	-1,250	0,038	Baik
23	Juring lingkaran	1,076	0,051	0,359	Kurang baik (c>0,25)
24	Persekutuan lingkaran	1,114	-0,667	0,136	Baik
25	Suku dan faktor	0,880	-0,792	0,049	Baik
26	Fungsi kuadrat	1,064	-0,502	0,116	Baik
27	Phytagoras dan luas segitiga	1,247	-0,690	0,292	Kurang baik (c>0,25)
28	Barisan dan deret	1,328	-1,522	0,035	Baik
29	Trigonometri	0,717	-0,265	0,500	Kurang baik (c>0,25)
30	Logaritma	1,170	0,045	0,500	Kurang baik (c>0,25)

Data hasil bangkitan berupa skala kemampuan peserta yang dianggap sebagai kemampuan sebenarnya (θ_T) dan pola respons peserta. Pola respons peserta kemudian dianalisis dengan BILOGMG untuk tiap replikasi pada tiap kasus untuk memperoleh kemampuan hasil estimasi (θ_T). Hasil keduanya dibandingkan dengan menghitung MSE maupun korelasinya. Pada tiap kasus, dihitung rerata dari MSE. Setelah

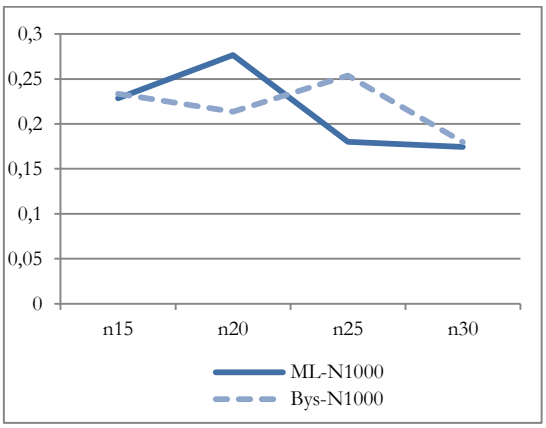
mengestimasi kemampuan 4 x 3 x 40 x 2 = 960 set data dan analisis MSE maupun korelasi, diperoleh hasil yang disajikan pada Tabel 2. Untuk mengamati polanya, hasil disajikan dengan menggunakan grafik pada Gambar 6 dan Gambar 7.

Tabel 2. Hasil Penghitungan MSE dan Korelasi

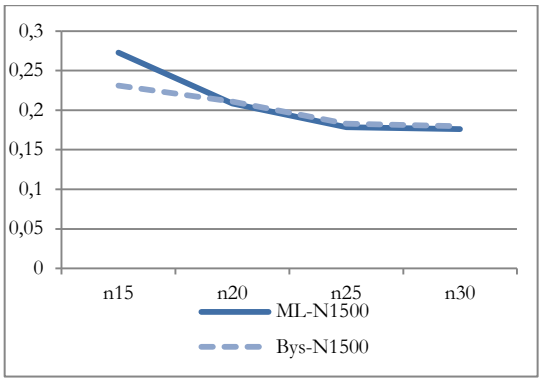
	MSE		Korelasi	
Kasus	MLE	Bayes	MLE	Bayes
n15N500	0,231	0,240	0,887	0,882
n15N1000	0,228	0,234	0,886	0,884
n15N1500	0,273	0,231	0,866	0,886
n20N500	0,300	0,254	0,853	0,877
n20N1000	0,276	0,213	0,864	0,893
n20N1500	0,208	0,211	0,897	0,896
n25N500	0,191	0,247	0,906	0,878
n25N1000	0,180	0,254	0,910	0,873
n25N1500	0,179	0,183	0,911	0,909
n30N500	0,190	0,197	0,907	0,903
n30N1000	0,175	0,180	0,913	0,911
n30N1500	0,176	0,180	0,912	0,910



(a)



(b)



(c)

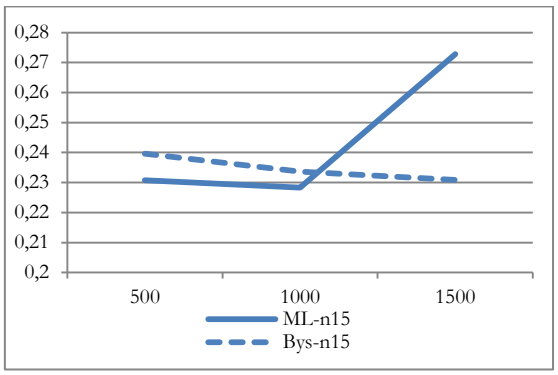
Gambar 6. Pengaruh Panjang Tes Terhadap Estimasi Kemampuan Menggunakan ML dan Bayes

Mencermati Gambar 6(a), diperoleh bahwa dengan peserta 500, perhitungan MSE kemampuan dari metode ML dan Bayes diperoleh hasil hampir sama dengan butir 15 dan 30, MSE lebih tinggi ketika estimasi dilakukan dengan 20 butir menggunakan ML, dan 25 butir meggunakan Bayes. Hasil ini mirip dengan kasus ketika mengguanakan 1000 peserta, yang disajikan pada Gambar 6(b). Pada Gambar 6(c), ketika menggunakan 1.500 peserta, dengan 15 butir pada awalnya ML menghasilkan MSE yang lebih tinggi, namun pada kasus yang lain diperoleh hasil yang stabil sama.

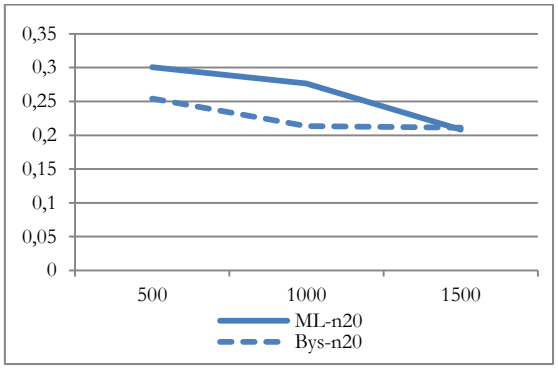
Hasil ini menunjukkan bahwa estimasi kemampuan dengan metode ML dan Bayes dengan peserta 500 dan 1.000 orang, diperoleh hasil hampir sama dengan butir 15 dan 30, dengan 20 butir estimasi lebih akurat menggunakan Bayes, dan jika dengan 25 butir, estimasi akan lebih akurat menggunakan ML. Ketika menggunakan 1.500 peserta, dengan 15 butir pada awalnya Bayes lebih akurat, namun dengan butir lebih banyak diperoleh hasil yang stabil kedua metode memiliki tingkat akurasi yang mirip.

Pengaruh banyaknya peserta tes terhadap estimasi kemampuan disajikan pada Gambar 8. Pada estimasi dengan menggunakan 15 butir yang disajikan pada Gambar 7(a), estimasi akan lebih tepat menggunakan ML dengan melibatkan peserta sebanyak 500 dan 1000. Namun, ketika pesertanya 1.500, hasil estimasi kemampuan akan lebih akurat dengan menggunakan metode Bayes.

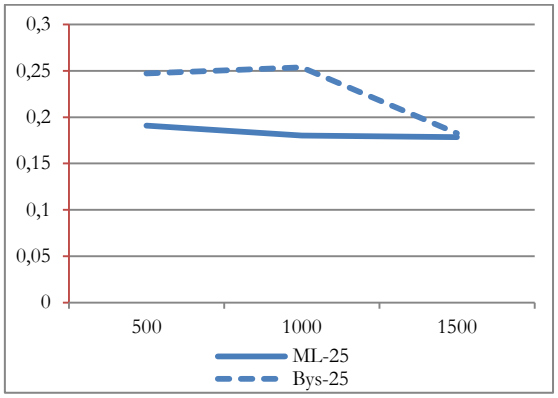
Pada estimasi dengan 20 butir pada Gambar 7(b), diperoleh hasil konsisten metode Bayes lebih akurat baik pada 500, 1.000, maupun 1.500 peserta. Namun sebaliknya pada estimasi dengan 25 dan 30 butir, diperoleh hasil konsisten lebih akurat menggunakan metode ML. Hasil tersebut menunjukkan bahwa estimasi dengan 25 dan 30 butir pada Gambar 7(c) dan 7(d), akan lebih tepat dilakukan dengan metode ML, namun metode Bayes akan lebih akurat untuk estimasi dengan 20 butir, dan belum stabil untuk estimasi dengan 15 butir.



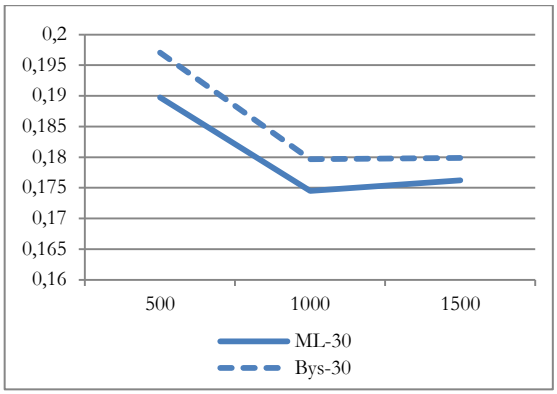
(a)



(b)



(c)



(d)

Gambar 7. Pengaruh Banyaknya Peserta Terhadap Estimasi Kemampuan Menggunakan ML dan Bayes

Hasil analisis mengenai pengaruh panjang tes dan banyaknya peserta terhadap estimasi kemampuan menggunakan ML dan Bayes dapat dipahami. Hal ini dapat disebabkan karena memperhatikan sifat stabilitas. Semakin panjang tes dan semakin banyak peserta, estimasi parameter butir dan parameter kemampuan menjadi semakin stabil. Mengenai perbandingan hasil antara estimasi dengan ML dan Bayes, meskipun pada dasarnya ada perbedaan nilai MSE, namun jika dicermati nilai-nilai tersebut pada Tabel 2 hasilnya tidak terlalu jauh berbeda. Hasil ini diperkuat juga oleh korelasi antara kemampuan sebenarnya dengan kemampuan hasil estimasi, yang juga disajikan pada Tabel 2. Hasil analisis korelasi ini menunjukkan baik pada estimasi dengan ML dan Bayes, korelasi kemampuan sebenarnya dengan kemampuan hasil estimasi berada pada kategori yang sangat tinggi (lebih dari 0,85).

Simpulan

Hasil studi menunjukkan bahwa pada estimasi kemampuan laten dengan 15, 20, 25, dan 30 butir dengan 500 dan 1.000 peserta, hasil MSE belum stabil, namun ketika peserta menjadi 1.500 orang, diperoleh akurasi estimasi kemampuan yang hampir sama baik estimasi antara metode ML dan metode Bayes. Pada estimasi dengan 15 dan 20 butir dan peserta 500, 1.000, dan 1.500, perbandingan hasil MSE belum stabil, dan

ketika estimasi melibatkan 25 dan 30 butir, baik dengan peserta 500, 1.000, maupun 1.500 akan diperoleh hasil yang lebih akurat dengan metode ML.

Daftar Pustaka

Anonim. (2005). Monte-*Carlo simulation*. Bahan kuliah Universitas Alberta. Diambil dari http://www.ualberta.ac/ tanggal 2 Januari 2006.

Cohen, A.S., Kane, M.T., & Kim, S. (2001). The precision of simulation study results**.** *Applied Psychological Measurement Journal*. Vol. 25 No. 2. pp. 136-145.

Du Toit, M. (2003). *IRT from SSi: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood: SSi.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

Hambleton, R.K. & Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. Dalam D. Weiss (Ed.) *New Horizon in Testing* (pp. 31-49). New York: Academic Press.

Harwell, M.R., Stone, C.A., Hsu, T.C., et al. (1996). Monte-Carlo studies in item response theory. *Applied Psichological Measurement*, 20, 101-125.

Harwell, M., (1997). Analyzing the result of Monte-Carlo studies in item response theory. *Educational and Psychological Measurement*, 20, 266-279.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Parama.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item response theory : Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Keeves, J.P. dan Alagumalai,S. (1999). New appoaches to measurement. Dalam Masters, G.N. dan Keeves, J.P.(Eds). Advances in measurement in educational research and assesment. Amsterdam: Pergamon.

Komrey, J.D., Parshall, C.G., Chason, W.M., & Yi, Q. (2006). *Generating responses based on multidimensional item response theory*. Diambil dari http://www2.sas.com/procedings/sugi19/posters/ tanggal 1 September 2006.

Mislevy, R.J. & Bock, R.D. (1990). *BILOG 3: Item analysis & test scoring with binary logistic models*. Moorseville: Scientific Sofware Inc.

Segall, D.O. (2000). General ability measurement: An application of multidimensional item response theory. *Psychometrica*, Vol. 66, 79-97.

Sinharay, S. (2004). Experiences with Marcov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*. Vol . 29. No. 4, 461-488.