BAB II

KAJIAN PUSTAKA

A. Kajian Teori

Teori tentang waktu respon telah mengalami perkembangan, khususnya yang terkait dengan teori pemodelan waktu respon. Teori-teori ini memang sudah lama dibahas oleh beberapa orang peneliti, tetapi pengembangannya tidak sebanyak model IRT (*Item Response Theory*).

1. Model Teori Respon Butir/ *Item Response Theory* (IRT)

Pada pendekatan Teori Tes Klasik/ *Classical Test Theory* (CTT), kemampuan siswa dinyatakan dengan skor total yang diperolehnya. Tingkat kesulitan soal dalam CTT dimaknai sebagai proporsi peserta tes yang menjawab benar pada suatu sampel atau kelompok peserta tes tertentu. Pemaknaan ini menunjukkan bahwa tingkat kesulitan soal bergantung kepada peserta tes yang dikenai butir soal tersebut dan sebaliknya kemampuan peserta tes tergantung pada apakah butir-butir soal tersebut mudah atau sulit. Jika tingkat kesulitan rendah, maka estimasi kemampuan akan tinggi dan sebaliknya. Besarnya daya beda, validitas juga reliabilitas skor tes tergantung pada sampel atau kelompok peserta tes yang dikenai butir soal. Penggunaan CTT untuk analisis butir soal memang relatif mudah, tetapi memiliki beberapa kelemahan. Kelemahan utamanya adalah keterikatan CTT pada sampel. Dalam proses pembelajaran, hal ini akan menimbulkan berbagai macam persoalan, terutama untuk melihat kemampuan peserta tes secara perorangan. Karena respon setiap peserta tes terhadap soal tidak

bisa dijelaskan oleh CTT. Oleh karena itulah hadir pendekatan teori tes modern atau Teori Respons Butir/ *Item Response Theory* (IRT) yang mempunyai kelebihan dapat menjelaskan respon setiap peserta tes terhadap soal. Pendekatan IRT ini merupakan suatu alternatif yang dapat digunakan dalam analisis butir soal sebagai upaya membebaskan alat ukur dari keterikatan terhadap sampel.

Berdasarkan model matematis IRT, didapatkan penjelasan bahwa probabilitas peserta tes untuk menjawab benar butir soal tergantung pada kemampuan peserta tes dan karakteristik butir soal. Hal ini berarti bahwa peserta tes yang mempunyai kemampuan tinggi akan memiliki probabilitas menjawab benar lebih besar jika dibandingkan dengan peserta tes yang mempunyai kemampuan rendah. Pendekatan IRT dapat dibedakan dari pendekatan CTT dengan beberapa karakteristik berikut ini, yaitu: (1) IRT memiliki asumsi yang detail, (2) IRT lebih fokus pada item sebagai unit yang unik dan independen dibandingkan tes secara utuh, (3) IRT memberikan perhatian terhadap hasil ujian individu *versus* hasil ujian kelompok, (4) IRT juga mempertimbangkan berbagai skala/ metrik di luar *raw score* (jumlah benar), (5) IRT mempunyai tipe, jangkauan dan prediksi yang mendalam, dan (6) dalam IRT penting untuk menguji akurasi dari asumsi model dan prediksi model (Brennan, 2006: 111). Hambleton &Swaminathan (2013: 16) dan Hambleton et al. (1991: 9) menyatakan bahwa ada tiga asumsi yang mendasari IRT, yaitu unidimensi, independensi lokal dan invariansi parameter.

Unidimensi, artinya setiap butir tes hanya mengukur satu kemampuan saja. Contohnya, pada ujian mata pelajaran matematika, butir-butir yang termuat di

dalam ujian tersebut hanya mengukur kemampuan siswa dalam mata pelajaran matematika saja, bukan kemampuan dalam bidang yang lainnya. Pada praktiknya, asumsi unidimensi ini sulit untuk dipenuhi karena adanya faktor-faktor lain yang turut mempengaruhi pelaksanaan tes. Oleh karena itu asumsi ini dapat ditunjukkan hanya dengan, jika tes mengandung satu saja komponen dominan yang mengukur kemampuan peserta tes, maka sudah dapat dikatakan memenuhi asumsi unidimensi.

Asumsi independensi lokal ini akan terpenuhi apabila jawaban peserta terhadap satu butir soal tidak mempengaruhi jawaban peserta terhadap terhadap butir soal yang lainnya. Tes untuk memenuhi asumsi independensi lokal dapat dilakukan dengan membuktikan bahwa probabilitas peserta tes dapat menjawab benar butir soal ke-1 sampai dengan butir soal ke-m sama dengan hasil kali probabilitas peserta tes menjawab benar pada setiap butir soalnya.

Invariansi parameter artinya kemampuan seseorang tidak akan berubah hanya karena mengerjakan tes yang berbeda tingkat kesulitannya dan *item parameter* tidak akan berubah hanya karena diujikan pada kelompok peserta tes yang berbeda tingkat kemampuannya. Menurut Hambleton et al. (1991: 18), invariansi parameter kemampuan dapat diselidiki dengan mengujikan dua perangkat tes atau lebih yang memiliki tingkat kesulitan yang berbeda pada sekelompok peserta tes, jika hasil estimasi kemampuan peserta tes tidak berbeda walaupun tes yang dikerjakan berbeda tingkat kesulitannya maka invariansi parameter kemampuan terbukti. Invariansi parameter butir dapat diselidiki dengan mengujikan suatu tes pada dua atau lebih kelompok peserta yang berbeda

kemampuannya, jika hasil estimasi *item parameter* tidak berbeda walaupun diujikan pada kelompok peserta yang berbeda tingkat kemampuannya maka invariansi parameter butir terbukti.

Pendekatan IRT dikembangkan oleh para ahli pengukuran bidang psikologi dan pendidikan sebagai upaya meminimalkan kelemahan-kelemahan yang ada dalam CTT. Pada IRT, hal penting yang juga perlu diperhatikan adalah pemilihan model yang tepat. Karena pemilihan model yang tepat akan dapat mengungkap keadaan yang sebenarnya.

Pada prinsipnya model IRT menggunakan distribusi normal standar N(0,1). Namun karena perhitungan distribusi normal agak rumit (menggunakan integral), maka dicarilah distribusi yang setara dengan distribusi normal standar N(0,1) yaitu distribusi logistik L(0,1.7) yang tidak melibatkan proses integral dalam perhitungan hasil estimasinya (Mardapi, 2012: 202-203). Selain karena proses perhitungan yang lebih sederhana, IRT lebih cenderung mengunakan distribusi logistik karena perbedaan integral dari distribusi normal standar N(0,1) dan distribusi logistik L(0,1.7) lebih kecil dari 0,01 (Mardapi, 2012: 203). Karena perbedaan integral yang cukup kecil ini (< 0,01), maka besarnya peluang yang merupakan luasan distribusi logistik L(0,1.7) dan distribusi normal standar N(0,1) hampir sama, integral akan mendekati 1 jika batas atasnya menuju $+\infty$.

Parameter utama yang menjadi dasar dalam perhitungan pada IRT adalah kemampuan peserta tes (theta/ Θ). Nilai Θ tidak terbatas, terbentang dari $-\infty$ sampai ∞, meskipun demikian nilai Θ dapat ditentukan dalam suatu batasan baku berdistribusi normal dari -4 sampai 4 (Hambleton et al., 1991). Selain Θ, parameter

lainnya yang juga dapat menunjukkan kebaikan suatu item adalah tingkat kesulitan, daya beda dan *pseudo guessing*.

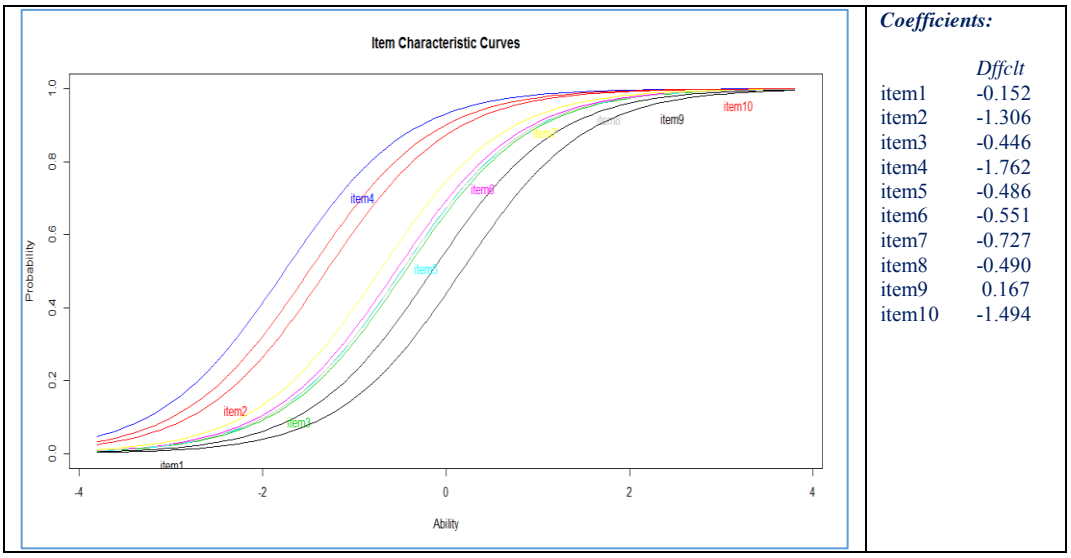
Dalam IRT, ada 3 model yang sudah umum dikenal. Model-model tersebut adalah Model Logistik 1 Parameter/ 1 *Parameter Logistic Model* (Model 1PL), Model Logistik 2 Parameter/ 2 *Parameter Logistic Model* (Model 2PL) dan Model Logistik 3 Parameter/ 3 *Parameter Logistic Model* (Model 3PL) (Kusumawati & Hadi, 2018: 71).

Dalam model logistik (IRT) yang menggunakan 1 parameter (Model 1PL), tingkat kesulitan soal (b) didefinisikan sebagai suatu titik pada skala kemampuan agar seorang peserta tes memiliki probabilitas menjawab benar sebesar 0,5 pada butir tertentu. Misal suatu butir soal mempunyai b=2, maka diperlukan kemampuan minimal 2 pada skala untuk dapat menjawab benar dengan probabilitas 0,5. Semakin besar nilai b maka semakin besar pula kemampuan yang diperlukan untuk dapat menjawab benar dengan probabilitas 0,5. Nilai b terbentang dari - ∞ sampai ∞, namun nilai b dapat dikategorikan baik jika berada pada rentang -2 sampai 2 (Hambleton & Swaminathan, 1985: 107). Dalam Model 1PL, probabilitas kemampuan peserta tes dapat dibuat dalam model matematis sebagai berikut (Hambleton et al., 1991: 12).

$$P(x_{ij}=1|\theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)} \qquad (1)$$

Hubungan antara variabel dependen dengan parameter pada IRT dapat dinyatakan dalam *Item Characteristic Curve* (ICC). Pada dasarnya ICC menggambarkan perubahan tingkat kemampuan peserta tes (theta/Θ) berpengaruh terhadap pergeseran probabilitas peserta tes menjawab benar. Berikut pada Gambar 1 adalah

contoh ICC 10 butir soal untuk Model 1PL (Wulansari, 2017: 4), yang menunjukkan bentuk monoton naik, dimana probablitas benarnya tidak akan sama dengan 1 (hanya mendekati 1).



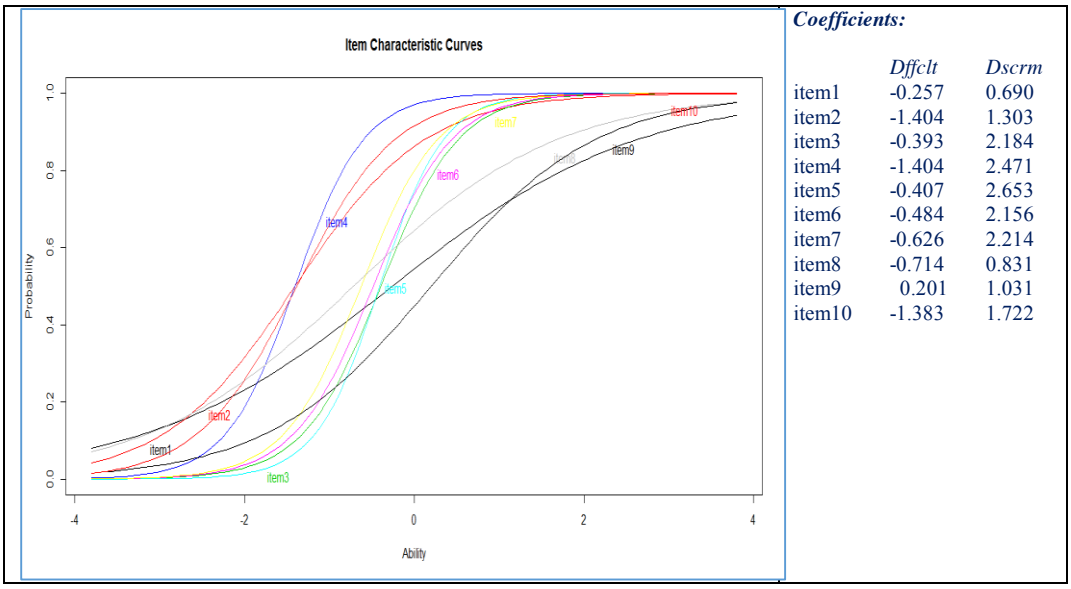
Coefficients:

	Dffclt
item1	-0.152
item2	-1.306
item3	-0.446
item4	-1.762
item5	-0.486
item6	-0.551
item7	-0.727
item8	-0.490
item9	0.167
item10	-1.494

Gambar 1. ICC untuk Model Logistik 1 Parameter

Model 2PL menggunakan parameter tingkat kesulitan soal (b) dan daya beda (a). Parameter a menunjukkan *slope* (kemiringan) dari kurva karakteristik/ICC di titik b pada skala kemampuan tertentu. Daya beda (a) berfungsi untuk menentukan dapat tidaknya suatu butir soal membedakan suatu kelompok dalam aspek yang diukur, sesuai dengan perbedaan yang ada dalam kelompok tersebut. Nilai a terbentang dari $-\infty$ sampai ∞, namun nilai a dapat dikategorikan baik jika berada pada rentang 0 sampai 2 (Hambleton & Swaminathan, 1985: 37). Menurut Hambleton et al. (1991) model matematis untuk Model 1PL apabila ditambahkan konstanta daya beda (a) yang bertindak sebagai arah kemiringan pada *ogive* normal (kurva karakteristik/ICC) adalah sebagai berikut.

$$P(x_{ij}=1|\theta_i, b_j, a_j) = \frac{\exp(a_j(\theta_i - b_j))}{1+\exp(a_j(\theta_i - b_j))} \qquad (2)$$



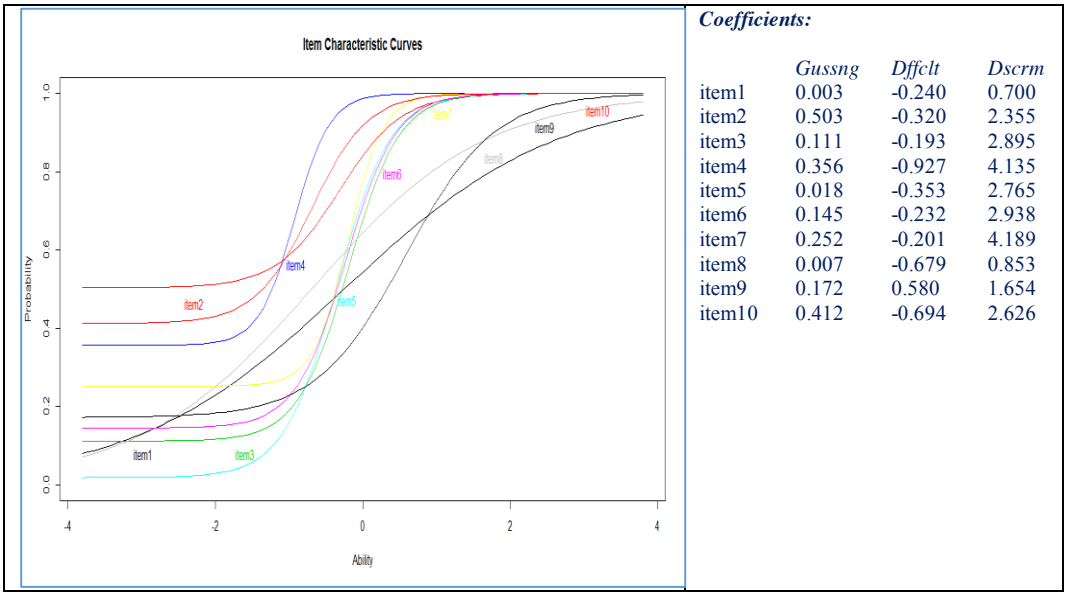
Gambar 2. ICC untuk Model Logistik 2 Parameter

Gambar 2 adalah contoh ICC 10 butir soal untuk Model 2PL (Wulansari, 2017: 7), semakin besar kemiringannya maka semakin besar nilai daya bedanya (Retnawati, 2014: 17). Bentuk *Slope* atau kemiringan garis lengkung ICC adalah efek ditambahkannya parameter a pada model logistik, sehingga bentuk *slope* adalah pembeda antara Model 1PL dan Model 2PL. Asimtot pada ICC Model 2PL untuk Θ → -∞ adalah 0. Jika tidak sama dengan 0, maka nilai tersebut adalah adalah *pseudo guessing* (c), sehingga modelnya akan berubah menjadi Model 3PL.

Model 3PL menggunakan parameter tingkat kesulitan soal (b), daya beda (a), dan *pseudo guessing* (c). *Pseudo guessing* adalah probabilitas peserta tes yang mempunyai kemampuan rendah untuk menjawab dengan benar butir soal yang mempunyai tingkat kesulitan yang tidak sesuai dengan kemampuan peserta tes tersebut (Retnawati, 2014: 80). Nilai c terbentang dari 0 sampai 1, namun nilai c

dapat dikategorikan baik jika nilai c < 1/k (Hulin et al., 1983: 36). Model matematis untuk Model 3PL adalah sebagai berikut (Hambleton et al., 1991: 17).

$$P(x_{ij}=1|\theta_i, b_j, a_j, c_j) = c_j + (1-c_j)\frac{\exp(a_j(\theta_i - b_j))}{1+\exp(a_j(\theta_i - b_j))} \qquad (3)$$



Coefficients:

	Gussng	*Dffclt*	*Dscrm*
item1	0.003	-0.240	0.700
item2	0.503	-0.320	2.355
item3	0.111	-0.193	2.895
item4	0.356	-0.927	4.135
item5	0.018	-0.353	2.765
item6	0.145	-0.232	2.938
item7	0.252	-0.201	4.189
item8	0.007	-0.679	0.853
item9	0.172	0.580	1.654
item10	0.412	-0.694	2.626

Gambar 3. ICC untuk Model Logistik 3 Parameter

Gambar 3 adalah contoh ICC 10 butir soal untuk Model 3PL (Wulansari, 2017: 10), yang menunjukkan asimtot untuk Θ → -∞ tidak sama dengan 0. Parameter c dapat memberikan kemungkinan asimtot bawah tidak sama dengan 0 (*nonzero lower asymptote*). Bentuk *lower asymptote* ini adalah efek ditambahkannya parameter c pada model logistik, sehingga bentuk *lower asymptote* adalah pembeda antara Model 1PL, Model 2PL dan Model 3PL.

Menurut Hambleton et al. (1991) agar estimasi parameter stabil haruslah menggunakan ukuran sampel yang besar. Untuk mengestimasi kemampuan dengan ukuran sampel besar bisa dengan menggunakan *Maximum Likelihood Estimation* (MLE). Apabila ukuran sampel besar tidak tercapai, maka alternatifnya

bisa menggunakan metode estimasi *Bayesian*. Metode *Bayesian* dapat mengestimasi parameter model yang sifatnya kompleks, data yang melanggar asumsi dasar IRT, dan ukuran sampel yang kecil (Natesan, 2011: 550). Pada penelitian ini metode estimasi *Bayesian* diterapkan karena model yang dikembangkan mempunyai parameter model yang kompleks.

Dengan memanfaatkan hasil estimasi parameter ini, besarnya sumbangan atau kekuatan butir soal dalam mengungkap *latent trait* yang ingin diukur dari suatu tes dapat dijelaskan. Nilai Fungsi Informasi (NFI) butir dihitung disini untuk mendapatkan informasi butir soal mana yang sebenarnya cocok dengan model sehingga dapat membantu seleksi butir pada bank soal (Retnawati, 2014: 80-81). Secara matematis, fungsi informasi tes I(Θ) dapat didefinisikan sebagai jumlah dari fungsi informasi butir soal $I_i(\Theta)$ (Hambleton & Swaminathan, 1985: 94).

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{4}$$

dengan

$$I_i(\theta) = \frac{2{,}89a_i^2(1-c_i)}{[c + \exp(1{,}7a_i(\theta - b_i))][1 + \exp(-1{,}7a_i(\theta - b_i))]^2}$$

Besarnya parameter yang dihasilkan dalam IRT merupakan hasil estimasi, maka kebenarannya pengukurannya bersifat probabilistik demikian juga dengan nilai fungsi informasi yang dihasilkan juga tidak bisa terlepas dari kesalahan pengukuran yang dikenal sebagai *Standard Error of Mesurement* (SEM). Menurut Hambleton

et al. (1991: 94) hubungan NFI dengan SEM berbanding terbalik kuadratik. Secara matematis dapat ditulis menjadi :

$$SEM(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{5}$$

2. Pemodelan Waktu Respon

Pemodelan waktu respon hadir untuk memperbaiki konsep IRT, karena selalu ada dua parameter utama yang dapat digunakan untuk menggambarkan bagaimana pengerjaan suatu tes, yaitu akurasi respon dan waktu respon. Kelemahan IRT ada pada informasi yang diberikan yaitu hanya terbatas pada keakuratan saja, yang menunjukkan seberapa benar atau salah jawaban yang diberikan oleh peserta tes saja tanpa mempertimbangkan aspek *speed* (Hambleton & Swaminathan, 2013: 30).

Waktu respon (*response time)* atau yang kerap kali disebut sebagai *response latency* (Abdelfattah & Johanson, 2007: 3; Halkitis, 1996: 1) adalah waktu yang dibutuhkan oleh peserta tes untuk membaca sekaligus menjawab suatu butir soal tes (Verbic, 2010: 1). Waktu respon (t_{ij}) yang dicatat dalam penelitian ini, dihitung dari waktu peserta tes i mulai mengklik butir soal ke j untuk dibaca, dijawab, sampai dengan klik ke soal selanjutnya. Apabila peserta tes i kesulitan atau ragu untuk menjawab butir soal ke j dan memilih untuk kembali lagi menjawab soal ke j tersebut setelah selesai mengerjakan soal lainnya, maka waktu responnya adalah total waktu saat peserta tes i mulai mencoba mengerjakan butir soal ke j untuk pertama kalinya sampai dengan mencoba untuk kedua kalinya, dan seterusnya sampai dengan waktu tes habis. Meskipun sudah lama dibahas, kajian dengan topik

pemodelan waktu respon tidak sebanyak kajian tentang pemodelan IRT karena pencatatan waktu respon sedikit sulit dilakukan. Padahal pemodelan waktu respon dapat memberikan informasi ukuran *speed* dari peserta tes dalam menjawab soal. Ukuran s*peed* di sini mengindikasikan strategi apa yang digunakan atau dipilih oleh peserta tes, apakah hanya sekedar menebak atau memang mengerjakan soal sesuai dengan proses yang sewajarnya.

Kajian waku respon yang pernah dilakukan meliputi tiga cara yang berbeda (Entink & van der Linden, 2009: 47-48). Cara pertama adalah pemodelan waktu respon secara eksklusif, pemodelan seperti ini dilakukan jika soal yang diujikan mudah dan batasan waktu yang diberikan sangatlah ketat, dari hasil ujian ini didapatkan informasi yang sangat sedikit. Cara kedua adalah melakukan analisis terpisah pada model waktu respon dan model keakuratan skor, sehingga didapatkan informasi waktu respon dan keakuratan skor secara terpisah atau independen, beberapa peneliti yang mengembangkan cara kedua ini adalah van der Linden (2006) dan Entink, Fox, et al. (2009). Cara ketiga adalah menggabungkan waktu respon dan keakuratan skor ke dalam satu model secara simultan, beberapa peneliti yang mengembangkan cara kedua ini adalah Thissen (1979), Roskam (1997), Verhelst et al. (1997), Tuerlinckx &de Boeck (2005), Wang &Hanson (2005), Wang (2006), van der Linden (2007), Ingrisone et al. (2008), Ingrisone II et al. (2008), Entink, Fox, et al. (2009), Meyer (2010), Meng et al. (2015) dan (Hidayah & Kumaidi, 2016). Adapun yang menjadi fokus pada penelitian ini adalah cara pemodelan ketiga, yaitu model simultan antara IRT dengan waktu respon. Menurut (Schnipke & Scrams, 2002: 11), pengembangan model simultan

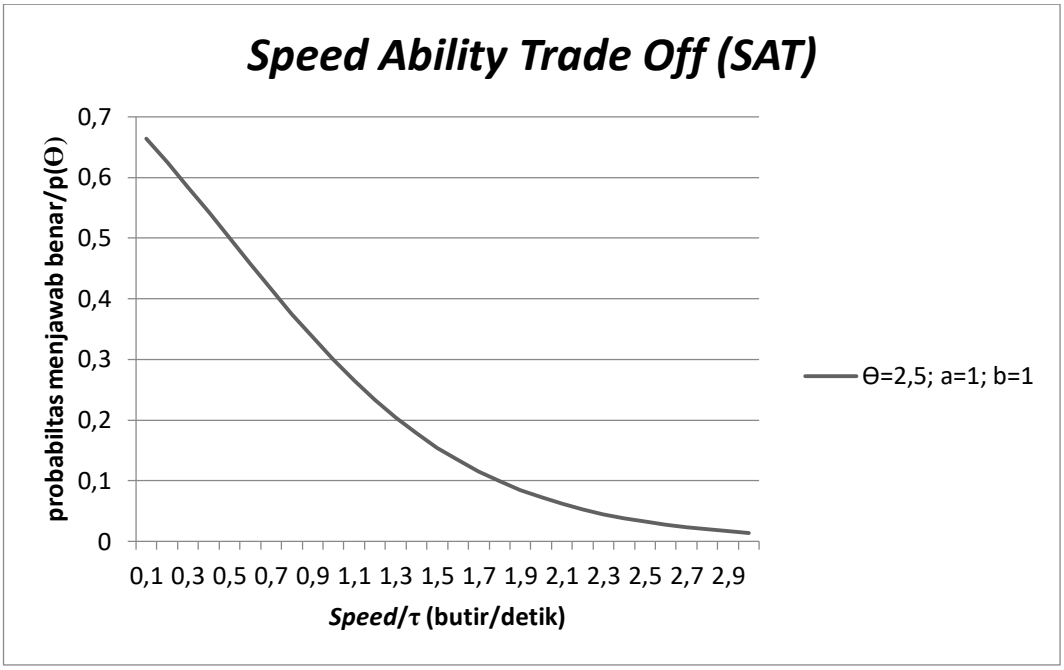
yang dilakukan oleh para ahli pada umumnya didasarkan pada tiga hal yaitu, (1) tipe tes, apakah *speed test* atau *power test*, (2) asumsi distribusi waktu respon yang digunakan, dan (3) asumsi hubungan antara keakuratan respon dan *speed*.

Pertama, pengembangan model simultan dilakukan berdasarkan jenis tes, apakah merupakan *speed test* atau *power test*. Menurut Gulliksen (1950), *power test* murni dan *speed test* murni mempunyai konsep yang berbeda. *Speed test* murni adalah tes yang dengan waktu terbatas dan butir soalnya mudah untuk dijawab dengan benar (tingkat kesulitan sama), tes seperti ini dapat di skor berdasarkan total waktu yang digunakan peserta tes untuk mengisi semua butir soal. *Power test* murni adalah tes dengan waktu yang tidak dibatasi, bedanya dengan *speed test* murni adalah tingkat kesulitan butir soalnya bervariasi, tes seperti ini cara penskorannya adalah dengan menghitung jumlah butir soal yang dapat dijawab dengan benar. Pada praktiknya pelaksanaannya *power test* murni itu tidak ada, karena tiap tes selalu dibatasi waktu (van der Linden & Hambleton, 2013: 166). Adanya batasan waktu tes, membuat adanya aspek lain yang terlibat dalam pengerjaan suatu butir soal yaitu aspek *speed* dari peserta tes dalam menjawab suatu butir soal. Ukuran *speed* peserta tes dapat diketahui dari waktu respon peserta tes dalam menjawab suatu butir soal.

Kedua, menurut Lindsey (2004: 197) ada beberapa karakteristik yang harus dipertimbangkan dalam pemilihan distribusi waktu respon untuk pengembangan model simultan: (1) waktu respon haruslah bernilai positif; (2) waktu respon yang singkat lebih sering terjadi dibandingkan waktu respon yang lama, atau dengan kata lain besarnya probabilitas waktu respon yang singkat sangat besar jika

dibandingkan besarnya probabilitas waktu respon yang lama (*positive skewed*). Distribusi-distribusi yang cocok dengan kedua karakteristik tersebut adalah distribusi *Lognormal*, *Weibull* dan *Gamma* (Lindsey, 2004: 203-206).

Ketiga, hubungan antara kemampuan seorang peserta tes dalam mengerjakan soal (akurasi) dengan *speed*-nya dapat dilihat dalam kurva *Speed Ability Trade Off*/ SAT (van der Linden, 2009: 259). *Speed* dalam menyelesaikan suatu butir soal bisa tergantung dengan strategi yang digunakan oleh seorang peserta tes. Semakin baik strategi yang dipilih oleh peserta tes, maka *speed* yang digunakan akan lebih tinggi, sehingga waktu respon yang dibutuhkan untuk menyelesaikan suatu butir soal bisa semakin pendek. Pada kondisi yang memungkinkan, peserta yang mempunyai strategi yang baik, akan mengerjakan soal-soal yang dianggap mudah (dapat dijawab dengan benar) di awal, dan akan mengerjakan soal-soal yang sulit di akhir. Menurut kurva SAT, seorang peserta tes yang memilih strategi menyelesaikan suatu butir soal dengan lambat, akan memiliki probabilitas menjawab benar lebih besar dibandingkan seorang peserta tes yang memilih strategi menyelesaikan suatu butir soal dengan cepat. Peserta tes yang memilih untuk meningkatkan *speed*-nya, akan berakibat menurunnya ketelitiannya dalam menyelesaikan suatu butir soal, sehingga probabilitas menjawab benar (akurasi jawaban) menjadi rendah.



Gambar 4. Hubungan antara Keakuratan dan *Speed* Peserta Tes

Probabilitas menjawab benar (p(Θ)) dalam tes dipengaruhi oleh *person parameter* seperti kemampuan (Θ), waktu respon (t) dan *speed* (τ) serta dipengaruhi oleh *item parameter* seperti tingkat kesulitan (b), daya beda (a) dan *pseudo guessing* (c). Gambar 4 menunjukkan hubungan antara probabilitas menjawab benar (keakuratan respon) dengan *speed*-nya (butir/detik) untuk persamaan 52 yang dapat dinyatakan dalam kurva monoton turun, yaitu semakin tinggi *speed* seorang peserta tes dalam mengerjakan soal maka semakin turun keakuratan jawabannya. Dari penjelasan di atas, dapat diketahui bahwa *speed* (semakin banyak butir/detik yang dikerjakan oleh peserta tes, maka artinya semakin cepat) yang dipilih oleh seorang peserta tes (mau pilih cepat atau lambat), akan berpengaruh terhadap waktu respon, dan waktu respon tersebut berpengaruh terhadap keakuratan jawaban peserta tes.

Penelitian pengembangan model simultan IRT dengan waktu respon memang tidak terlalu banyak. Beberapa peneliti yang mengembangkan adalah Thissen (1979), Roskam (1997), Verhelst et al. (1997), Wang &Hanson (2005), Wang (2006), van der Linden (2007), Ingrisone et al. (2008), Ingrisone II et al. (2008) dan Hidayah et al. (2016) .

Model simultan antara IRT dan waktu respon pertama kali dikembangkan oleh Thissen (1979). Model tersebut menggabungkan waktu respon ke dalam Model 2PL yang bisa diterapkan pada *power test* dan *speed test*, namun pada penelitian Ingrisone II et al. (2008: 6) terbukti hanya berlaku untuk *speed test*. Model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan Thissen (1979: 258) adalah sebagai berikut :

$$P(r_{ij}=1)=\frac{1}{1+\exp\left(-\left(z_{ij}\right)\right)} \qquad (6)$$

dengan $z_{ij}=a_j\theta_i+c_j$ dan $P(r_{ij}=0)=1-P(r_{ij}=1)$.

Pada persamaan 6, parameter kemampuan peserta tes disimbolkan θ, a adalah parameter daya beda, c adalah parameter tingkat kesulitan dan r_{ij} adalah jawaban peserta ke-i dalam menjawab soal ke-j, r_{ij} = 1 jika jawaban peserta benar, dan r_{ij} = 0 jika jawaban peserta salah.

Thissen (1979: 259) menggunakan distribusi *Lognormal* sebagai marginal waktu respon (t_{ij}). Nilai *mean log natural* waktu respon adalah $\mu_{\ln t_{ij}}=v-s+u-b.z_{ij}+\varepsilon_{ij}$ dengan $\varepsilon_{ij}\approx N(0,\sigma^2)$. *Log natural* waktu respon memiliki distribusi normal, atau dengan kata lain distribusi waktu respon memiliki

distribusi *Lognormal*, sehingga $\ln t_{ij} \approx N\left(\left(v - s + u - b.z_{ij} + \varepsilon_{ij}\right), \sigma^2\right)$, atau dapat ditulis menjadi :

$$f(t_{ij};\beta_j;\tau_i) = \frac{1}{t_{ij}\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(\ln t_{ij} - (v - s + u - b.z_{ij} + \varepsilon_{ij}))}{\sigma} \right]^2 \right) \qquad (7)$$

Pada persamaan 7, *mean* dari keseluruhan waktu respon disimbolkan sebagai v, parameter b adalah suatu koefesien regresi yang merefleksikan hubungan antara parameter kemampuan, daya beda dan tingkat kesulitan dengan waktu respon, kemudian σ^2 adalah varian dari t_{ij}, dan parameter kelambatan disimbolkan sebagai $(s+u)$ yang merupakan parameter tambahan. Menurut Thissen (1979: 259), parameter kelambatan perlu dipertimbangkan. Pada kenyataannya faktor kelambatan peserta tes dalam memencet tombol dan kelambatan waktu membaca karena panjangnya soal, terutama pada soal tes yang bersifat verbal berpengaruh terhadap waktu respon peserta tes, oleh karena itu parameter $(s+u)$ perlu ditambahkan dalam model. Thissen (1979: 261) mengalikan Model 2PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon pada persamaan 6 dengan distribusi *Lognormal* sebagai marginal waktu respon pada persamaan 7 dengan konsep *joint distribution* untuk mendapatkan *joint model*. Kelemahan dari model Thissen (1979) adalah, dapat diterapkan jika respon butir dan waktu respon memenuhi asumsi saling bebas, namun pada kenyataannya asumsi tersebut tidak bisa terpenuhi dalam praktik pengukuran (Wang, 2006: 1).

Pada penelitian selanjutnya, Roskam (1997) dan Verhelst et al. (1997) sama-sama memasukkan waktu respon pada model *Rasch*. Kedua model yang dihasilkan mempunyai keterbatasan yaitu hanya tepat digunakan untuk *speed test*.

Roskam (1997: 194) menggunakan *Weibull* sebagai distribusi untuk waktu respon. Verhelst et al. (1997: 171) menggunakan *Gamma* sebagai distribusi untuk waktu respon. Waktu respon dalam kedua model ini merupakan total waktu yang dibutuhkan untuk menjawab semua soal tes, bukan waktu untuk menjawab sebuah soal tes. Model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan Roskam (1997: 195) adalah sebagai berikut :

$$P(u_{ij}=1|t_i, j, i) = \frac{\theta_j t_{ij}}{\theta_j t_{ij} + \varepsilon_i} = \frac{\exp(\xi_j + \tau_{ij} - \sigma_i)}{1 + \exp(\xi_j + \tau_{ij} - \sigma_i)} \quad . \qquad (8)$$

Kemampuan peserta tes disimbolkan sebagai θ, t adalah waktu respon, dan ε adalah tingkat kesulitan. Simbol ξ, τ dan σ adalah logaritma dari θ, t dan ε. Rasionalisasi dari model di atas adalah kemampuan peserta tes dalam menjawab soal akan naik jika waktu menjawab soal ditambah, dan *speed* dari peserta tes adalah faktor yang mempengaruhi waktu peserta tes dalam menjawab soal (van der Linden & Hambleton, 2013: 193).

Roskam (1997: 194) menggunakan *Weibull* sebagai distribusi untuk waktu respon (t_{ij}). Distribusi *Weibull* karakteristiknya ditunjukkan oleh fungsi *hazard*/ fungsi kegagalan (h(t)), yang merupakan laju kegagalan dari suatu individu untuk mampu bertahan setelah melewati waktu yang ditetapkan (t) (Klein & Moeschberger, 2006).

$$h_{ij}(t) = \frac{\theta_j}{\varepsilon_i \delta_i} t \quad . \qquad (9)$$

Parameter persistensi disimbolkan sebagai δ, yaitu besarnya usaha yang dibutuhkan peserta tes untuk menyelesaikan suatu soal. Parameter kemampuan

peserta tes disimbolkan sebagai θ, dan parameter tingkat kesulitan soal disimbolkan sebagai ε. Berdasarkan fungsi pada persamaan 9, maka *Probability Density Function* (PDF) dari distribusi *Weibull* adalah :

$$f(t) = \lambda t.\exp\left(-\frac{\lambda}{2}t^2\right) \tag{10}$$

dengan $\lambda = \frac{\theta}{\varepsilon\delta}$.

Roskam (1997: 195) mengalikan Model 1PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon pada persamaan 8 dengan distribusi *Weibull* sebagai marginal waktu respon pada persamaan 10 kemudian hasilnya diintegralkan terhadap waktu respon sehingga mendapatkan *joint model* seperti persamaan berikut :

$$P(u_{ij} = 1|j,i) = \int_0^\infty \frac{\theta_j t_{ij}}{\theta_j t_{ij} + \varepsilon_i} \frac{\theta}{\varepsilon_i \delta_j} t.\exp\left(\frac{\theta_j}{2\varepsilon_i \delta_j}\right) dt \tag{11}$$

Kelemahan model Roskam (1997) dan Verhelst et al. (1997) adalah peluang menjawab benar bisa sama dengan 1, jika tidak ada pembatasan waktu tes (lebih cocok untuk *speed test*). Keterbatasan Roskam (1997) dan Verhelst et al. (1997) diperbaiki oleh Wang &Hanson (2005: 4) dengan menggabungkan waktu respon ke dalam Model 3PL yang bisa diterapkan pada *power test*. Model ini mengasumsikan waktu respon sebagai *fixed variable*, sehingga model tersebut dikenal dengan Model 4PL karena memiliki 4 parameter. Wang &Hanson (2005: 4) mengklaim bahwa model yang dihasilkan lebih cocok untuk *power test*.

Wang (2006: 2) memperbaiki asumsinya terkait waktu respon (*fixed variable*) karena asumsi ini sukar dipenuhi, dengan mengalikan Model 4PL sebagai

distribusi bersyarat keakuratan respon terhadap waktu respon dengan distribusi *Weibull* satu parameter sebagai marginal waktu respon (konsep *joint distribution*). Menurut Wang (2006: 2), penggunaan distribusi *Weibull* satu parameter perlu dikaji lebih lanjut karena terlalu sederhana untuk menangkap realita yang ada dalam praktik pengukuran. Berikut adalah model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan Wang (2006: 2):

$$P(u_{ij}=1|\theta_i,\rho_i,a_j,b_j,c_j,t_{ij}) = c_i + \frac{1-c_i}{1+\exp(-1{,}7a_j(\theta_i - \left(\frac{\rho_i d_j}{t_{ij}}\right) - b_j))}. \quad (12)$$

Parameter Model 3PL disimbolkan sebagai a, b, c dan θ, sedangkan t adalah waktu respon, ρ adalah parameter kelambatan karena peserta tes, dan d adalah parameter kelambatan karena soal (Wang, 2006: 2). Menurut Thissen (1979: 259) parameter kelambatan perlu dipertimbangkan, karena pada kenyataannya faktor kelambatan peserta tes dalam memencet tombol dan kelambatan waktu membaca karena panjangnya soal, terutama pada soal tes yang bersifat verbal berpengaruh terhadap waktu respon peserta tes.

Wang (2006: 2) menggunakan distribusi *Weibull* untuk waktu respon (t_{ij}). Distribusi *Weibull* karakteristiknya ditunjukkan oleh fungsi *hazard*/ fungsi kegagalan (h(t)), seperti berikut ini :

$$h_{ij}(t) = \lambda t_{ij}. \quad (13)$$

Fungsi pada persamaan 13 menunjukkan bahwa fungsi *hazard*/ fungsi kegagalan berbanding lurus dengan waktu respon. Berdasarkan fungsi pada persamaan 9, maka PDF dari distribusi *Weibull* adalah :

$$f(t_{ij}|\theta_i,\rho_i,\delta_j) = \lambda t_{ij}.\exp\left(-\frac{\lambda}{2}t^2\right) \quad (14)$$

dengan $\lambda = \rho_i(\theta_i - b_j)^2$.

Besarnya λ menunjukkan *slope* dari fungsi *hazard*/ fungsi kegagalan. Parameter ρ_i adalah *speed* peserta tes, θ_i adalah kemampuan peserta tes dan b_j adalah tingkat kesulitan soal. Semakin tinggi ρ_i maka λ juga akan semakin tinggi, sehingga *mean* dan varian distribusi waktu respon menurun (*mean* dan varian berbanding terbalik dengan *slope*). Semakin besar $(\theta_i - b_j)$, yang berarti kemampuan peserta tes lebih tinggi dari tingkat kesulitan soal, maka λ juga akan semakin tinggi, sehingga *mean* dan varian distribusi waktu respon menurun (*mean* dan varian berbanding terbalik dengan *slope*). Berdasarkan hal tersebut, Wang (2006: 2-3) menyatakan bahwa peserta tes pada umumnya akan menggunakan sedikit waktu untuk mengerjakan soal yang sesuai dengan kemampuannya.

Wang (2006: 2) mengalikan Model 4PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon pada persamaan 12 dengan distribusi *Weibull* sebagai marginal waktu respon pada persamaan 14 kemudian hasilnya diintegralkan terhadap waktu respon sehingga mendapatkan *joint model* seperti persamaan berikut :

$$P(u_{ij}=1|j,i) = \int_0^\infty c_i + \frac{1-c_i}{1+\exp(-1{,}7a_j(\theta_i - \left(\frac{\rho_i d_j}{t_{ij}}\right) - b_j))}\rho_i(\theta_i - b_j)^2 t_{ij}.\exp\left(-\frac{\left(\rho_i(\theta_i - b_j)^2\right)}{2}t^2\right)dt. \quad (15)$$

van der Linden (2007) juga mengembangkan model simultan (gabungan waktu respon dan keakuratan skor ke dalam satu model secara simultan) untuk

mendeteksi perilaku *cheating* yang dikenal sebagai *Hierarchical Framework Model*. Distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Model 3PL, dan marginal waktu responnya menggunakan distribusi *Lognormal*. *Hierarchical Framework Model* memiliki dua level dalam pemodelannya. Pada level pertama, parameter-parameter dalam Model 3PL dan distribusi *Lognormal* diestimasi secara terpisah. Pada level kedua, besaran hasil estimasi pada level satu tersebut dikelompokkan menjadi dua yaitu kelompok vektor *person parameter* dan kelompok vektor *item parameter* kemudian masing-masing kelompok vektor parameter tersebut dimodelkan sendiri-sendiri dengan konsep *joint distribution*, sehingga didapatkan model distribusi *normal multivariate* dengan *person parameter* dan model distribusi *normal multivariate* dengan *item parameter*. Berikut adalah model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan van der Linden (2007: 293) :

$$P(u_{ij}=1|\theta_j, a_i, b_i, c_i) = c_i + (1-c_i)\frac{\exp(a_i(\theta_j - b_i))}{1+\exp(a_i(\theta_j - b_i))} \quad . \qquad (16)$$

Respon peserta tes disimbolkan sebagai u, c adalah parameter *pseudo guessing*, a adalah parameter daya beda, b adalah parameter tingkat kesulitan, dan θ adalah parameter kemampuan peserta tes.

van der Linden (2007: 293) menggunakan distribusi *Lognormal* sebagai marginal waktu respon (t_{ij}). Nilai *mean log natural* waktu respon adalah $\mu_{\ln t_{ij}} = \beta_j - \tau_i$. *Log natural* waktu respon memiliki distribusi normal, atau dengan

kata lain distribusi waktu respon memiliki distribusi *Lognormal*, sehingga $\ln t_{ij} \approx N\left(\left(\beta_j - \tau_i\right), \left(\frac{1}{\alpha_i}\right)^2\right)$, atau dapat ditulis menjadi :

$$f(t_{ij};\tau_j,\alpha_i,\beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(\ln t_{ij} - (\beta_i - \tau_j))}{\left(\frac{1}{\alpha_i}\right)}\right]^2\right). \quad (17)$$

Pada Persamaan 17, lama waktu yang dibutuhkan untuk menyelesaikan soal tes ke-i, disimbolkan menjadi β_i. Semakin besar β_i, maka *time intensity* yang dibutuhkan oleh peserta ke-i juga semakin besar. Parameter τ_j adalah *speed* peserta tes ke-j dalam menyelesaikan soal, semakin besar τ_j, maka semakin tinggi *speed* dari peserta tes ke-j, sehingga semakin sedikit waktu yang dibutuhkan untuk menyelesaikan soal tes ke-i dan $\left(\frac{1}{\alpha_i}\right)^2$ adalah varians dari t_{ij}.

Pada level pertama, proses estimasi menghasilkan vektor *person parameter* ξ_j dan *item parameter* ψ_i. Parameter-parameter pada level satu dimodelkan kembali pada level dua dengan konsep *joint distribution*.

Person parameter ξ_j dimodelkan kembali dengan distribusi *normal multivariate* $\xi_j \approx f\left(\xi_j;\mu_p,\Sigma_p\right)$ seperti berikut ini:

$$f\left(\xi_j;\mu_p,\Sigma_p\right) = \frac{\left|\Sigma_p^{-1}\right|^{1/2}}{2\pi} \exp\left(-\frac{1}{2}\left(\xi_j - \mu_p\right)^T \Sigma_p^{-1}\left(\xi_j - \mu_p\right)\right) \quad (18)$$

dengan :

$\xi_j = \begin{bmatrix} \theta_j \\ \tau_j \end{bmatrix}$ yang merupakan vektor *person parameter* berordo 2x1

$\mu_p = \begin{bmatrix} \mu_\theta \\ \mu_\tau \end{bmatrix}$ yang merupakan vektor rata-rata berordo 2x1

$\Sigma_p = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{bmatrix}$ yang merupakan matriks kovarian berordo 2x2

sehingga $f(\xi_j; \mu_p, \Sigma_p)$ bernilai skalar.

Parameter peserta soal ψ_i dimodelkan kembali dengan distribusi *normal multivariate* $\psi_i \approx f(\psi_i; \mu_l, \Sigma_l)$ seperti berikut ini:

$$f(\psi_i; \mu_l, \Sigma_l) = \frac{\left|\Sigma_l^{-1}\right|^{1/2}}{2\pi} \exp\left(-\frac{1}{2}(\psi_i - \mu_l)^T \Sigma_l^{-1}(\psi_i - \mu_l)\right) \quad (19)$$

dengan :

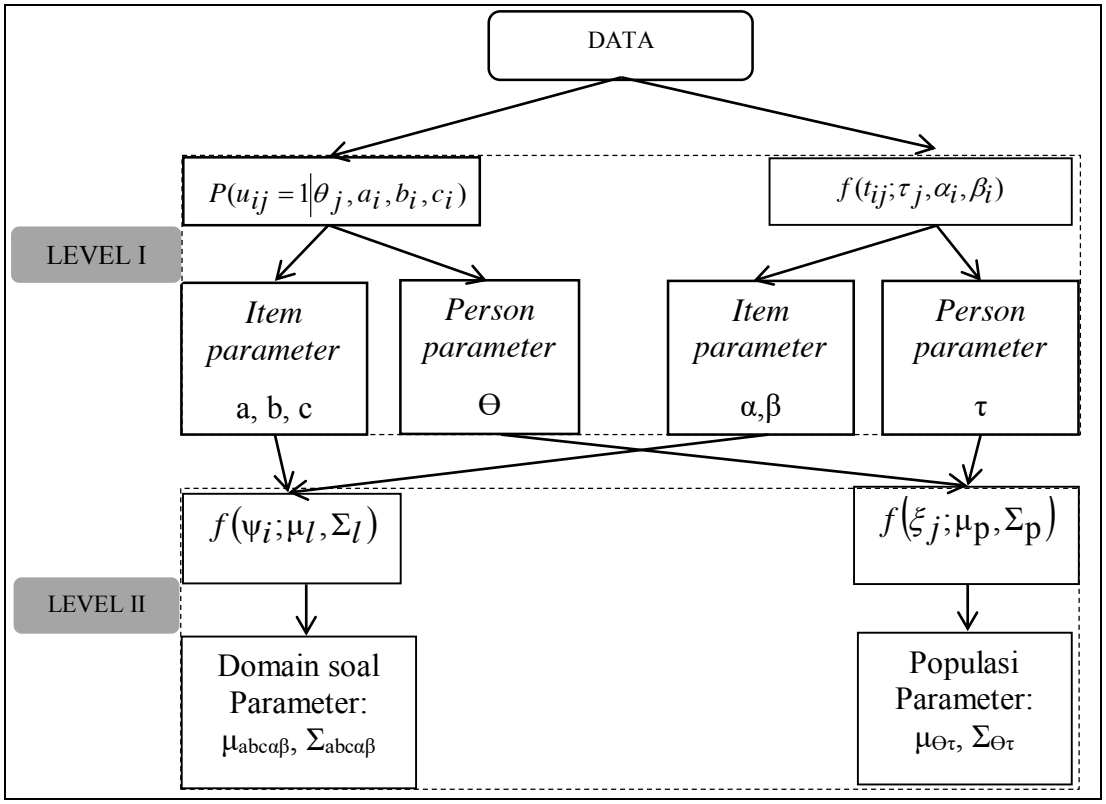
$\psi_i = \begin{bmatrix} a_i \\ b_i \\ c_i \\ \alpha_i \\ \beta_i \end{bmatrix}$ yang merupakan vektor *item parameter* berordo 5x1

$\mu_l = \begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \\ \mu_\alpha \\ \mu_\beta \end{bmatrix}$ yang merupakan vektor rata-rata berordo 5x1

$$\Sigma_l = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma ac & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma bc & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\alpha^2 \end{pmatrix}$$ yang merupakan matriks kovarian berordo 5x5

sehingga $f(\psi_i;\mu_l,\Sigma_l)$ bernilai skalar.

Untuk memperjelas konsep *Hierarchical Framework Model* yang dikembangkan oleh (van der Linden, 2007), berikut adalah bagannya:



Gambar 5. Konsep *Hierarchical Framework Model*

Selanjutnya, Ingrisone et al. (2008) dan Ingrisone II et al. (2008) juga mengembangkan model dengan konsep *joint distribution*. Ingrisone et al. (2008: 15) mengalikan Model 2PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon dengan distribusi *Lognormal* yang dia modifikasi. Ingrisone II et al.

(2008: 3) mengalikan Model 1PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon dengan distribusi *Weibull* dua parameter yang menurutnya model simultan yang dihasilkan cocok untuk *power test*, padahal dengan waktu respon yang sama dengan kemampuan peserta tes yang berbeda model Ingrisone II et al. (2008) menghasilkan probabilitas menjawab benar sebuah soal hampir sama besarnya (*speed test*). Model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan Ingrisone et al. (2008: 23) adalah:

$$P(u_{ij}=1|t_{ij},\theta_i,\rho_i,a_j,b_j)=\frac{\exp\left(-1{,}7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(-1{,}7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)}. \qquad (20)$$

Parameter kemampuan peserta tes disimbolkan sebagai θ, a adalah parameter daya beda, b adalah parameter tingkat kesulitan dan t adalah waktu respon. Parameter η nilainya konstan dan merupakan koefisien regresi dari waktu respon terhadap probabilitas menjawab benar, jika waktu respon tidak berpengaruh terhadap probabilitas peserta tes menjawab benar, maka nilai η akan sama dengan 0 dan η akan ada nilainya jika waktu respon berpengaruh terhadap probabilitas peserta tes menjawab benar Ingrisone et al. (2008: 21).

Ingrisone et al. (2008) menggunakan distribusi *Lognormal* sebagai marginal waktu respon (t_{ij}) dengan memodifikasi model distribusi *Lognormal* yang dikembangkan Thissen (1983). Ingrisone et al. (2008: 27) mengubah tanda negatif menjadi tanda positif dengan mengunakan konsep *Speed Ability Trade-Off (*SAT). Nilai *mean log natural* waktu respon adalah $\mu_{\ln t_{ij}}=v+s_i+r_j+gz_{ij}+\varepsilon_{ij}$, $\varepsilon_{ij}\approx N(0,\sigma^2)$ dengan $z_{ij}=a_j\theta_i+c_j$. *Log natural* waktu respon memiliki

distribusi normal, atau dengan kata lain distribusi waktu respon memiliki distribusi *Lognormal*, sehingga $\ln t_{ij} \approx N\left(\left(v + s_i + r_j + gz_{ij} + \varepsilon\right), \sigma^2\right)$, atau dapat ditulis menjadi :

$$f(t_{ij};\beta_j;\tau_i) = \frac{1}{t_{ij}\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(\ln t_{ij} - (v + s_i + r_j + gz_{ij} + \varepsilon))}{\sigma} \right]^2 \right). \qquad (21)$$

Pada Persamaan 21, *mean* dari keseluruhan waktu respon disimbolkan sebagai v, kemudian parameter σ^2 adalah varians dari t_{ij}, dan parameter g adalah suatu koefesien regresi yang merefleksikan hubungan antara parameter kemampuan, daya beda dan tingkat kesulitan dengan waktu respon. Parameter kelambatan karena peserta tes disimbolkan sebagai s_i dan parameter kelambatan karena soal disimbolkan sebagai r_j. Menurut Thissen (1979: 259) parameter kelambatan perlu dipertimbangkan, karena pada kenyataannya faktor kelambatan peserta tes dalam memencet tombol dan kelambatan waktu membaca karena panjangnya soal, terutama pada soal tes yang bersifat verbal berpengaruh terhadap waktu respon peserta tes. Ingrisone et al. (2008) mengalikan Model 2PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon pada persamaan 20 dengan distribusi *Lognormal* sebagai marginal waktu respon pada persamaan 21 dengan konsep *joint distribution* untuk mendapatkan *joint model*.

Model distribusi bersyarat keakuratan respon terhadap waktu respon yang digunakan Ingrisone II et al. (2008: 25) adalah sebagai berikut :

$$P(u_{ij} = 1 \mid t_{ij}, \theta_i, b_j, \eta) = \frac{\exp 1{,}7\left(\theta_i - \eta t_{ij} - b_j\right)}{1 + \exp 1{,}7\left(\theta_i - \eta t_{ij} - b_j\right)}. \qquad (22)$$

Parameter kemampuan peserta tes disimbolkan sebagai θ, b adalah parameter tingkat kesulitan dan t adalah waktu respon. Parameter η merupakan koefisien regresi dari waktu respon terhadap probabilitas menjawab benar. Jika η sama dengan 0 maka model pada Persamaan 22 akan menjadi Model 1PL.

Ingrisone II et al. (2008: 29) menggunakan distribusi *Weibull* untuk waktu respon (t_{ij}). Distribusi *Weibull* karakteristiknya ditunjukkan oleh fungsi *hazard*/ fungsi kegagalan (h(t)), seperti berikut :

$$h_{ij}(t) = \frac{\gamma}{\beta} t_{ij}^{\gamma-1} \tag{23}$$

dengan $\frac{\gamma}{\beta} = \frac{\rho_i (\theta_i - b_j)^2 \gamma}{2}$.

Besarnya $\frac{\gamma}{\beta}$ menunjukkan *slope* dari fungsi *hazard*/ fungsi kegagalan. Fungsi *slope* berarti, jika ρ_i bertambah maka nilai $\frac{\gamma}{\beta}$ akan bertambah, sehingga *mean* dan varian distribusi waktu respon menurun (*mean* dan varian berbanding terbalik dengan *slope*). Semakin besar $(\theta_i - b_j)$ maka nilai $\frac{\gamma}{\beta}$ akan bertambah, sehingga *mean* dan varian distribusi waktu respon menurun (*mean* dan varian berbanding terbalik dengan *slope*). Berdasarkan fungsi pada persamaan 23, maka PDF dari distribusi *Weibull* adalah :

$$f(t_{ij} | \theta_i, \rho_i, b_j) = \frac{\rho_i (\theta_i - b_j)^2 \gamma}{2} t_{ij}^{(\gamma-1)} . \exp\left(- \frac{\left(\rho_i (\theta_i - b_j)^2\right)}{2} t_{ij}^{\gamma} \right). \tag{24}$$

Parameter ρ_i adalah *speed* peserta tes, θ_i adalah kemampuan peserta tes dan b_j adalah tingkat kesulitan item soal.

Ingrisone II et al. (2008: 31) mengalikan Model 1PL sebagai distribusi bersyarat keakuratan respon terhadap waktu respon pada persamaan 22 dengan distribusi *Weibull* sebagai marginal waktu respon pada persamaan 24 kemudian hasilnya diintegralkan terhadap waktu respon sehingga mendapatkan *joint model* seperti persamaan berikut :

$$P(u_{ij}=1|j,i)=\int_0^{\infty}\frac{\exp 1{,}7\left(\theta_i-\eta t_{ij}-b_j\right)}{1+\exp 1{,}7\left(\theta_i-\eta t_{ij}-b_j\right)}\frac{\rho_i\left(\theta_i-b_j\right)^2\gamma}{2}t_{ij}^{(\gamma-1)}.\exp\left(-\frac{\left(\rho_i\left(\theta_i-b_j\right)^2\right)}{2}t_{ij}^{\gamma}\right)dt \quad .(25)$$

Keterbatasan model simultan Ingrisone II et al. (2008) ini diperbaiki oleh Hidayah et al. (2016) dengan memasukkan waktu respon ke Model 1PL dengan konsep *joint distribution* yang cocok untuk *power test*. Hidayah et al. (2016) menggunakan distribusi *Lognormal*, $\mu_{\ln t_{ij}}$ yang merupakan fungsi dari *speed* dan usaha yang dibutuhkan oleh peserta tes dalam menyelesaikan tes, yang mana faktor-faktor yang mempengaruhi $\mu_{\ln t_{ij}}$ tersebut masih perlu dikaji lagi karena pada praktiknya parameter-parameter tersebut masih sulit untuk didefinisikan.

Model pada Persamaan 26 merupakan Model 1PL yang memasukkan waktu respon sebagai *fixed variable*. Pada Persamaan 27, waktu respon dimodelkan berdasarkan distribusi *Lognormal*, karena distribusi tersebut dianggap sesuai dengan karakteristik waktu respon (nilainya positif dan bentuknya *positive skewed*) dan lebih mudah diinterpretasikan jika dibanding 2 distribusi lain yang sering digunakan dalam penelitian waktu respon yaitu *Weibull* dan *Gamma*. Pada

Persamaan 27, waktu respon dianggap sebagai *random variable* karena diasumsikan waktu respon tidak hanya dikontrol oleh petugas administrasi tes atau teknologi komputer saja tetapi ada faktor lain yang mempengaruhi, sehingga apabila seseorang menjalankan tes yang sama dengan waktu pelaksanaan tes yang berbeda, maka waktu respon yang ditempuh bisa jadi akan berbeda dengan tes yang dilaksanakan sebelumnya. Selanjutnya, model distribusi bersyarat keakuratan respon (Model 1PL) terhadap waktu respon, dan model waktu respon dimodelkan kembali dengan konsep *joint distribution*.

$$P(x_{ij}=1|\theta_j, d_j, t_{ij}, b_i) = \frac{\exp(\theta_j - \frac{d_j}{t_{ij}} - b_i)}{1+\exp(\theta_j - \frac{d_j}{t_{ij}} - b_i)}. \qquad (28)$$

Model di atas menunjukkan probabilitas respon menjawab benar peserta tes dipengaruhi oleh parameter kemampuan peserta tes (θ_j), kelambatan (d_j), waktu respon peserta tes (t_{ij}) dan tingkat kesulitan (b_i). Apabila besarnya t_{ij} tidak terhingga atau tes tidak dibatasi waktu, maka Persamaan 26 sama dengan Model 1PL pada IRT.

Nilai *mean log natural* waktu respon adalah $\mu_{\ln t_{ij}} = \beta_j - \tau_i$. *Log natural* waktu respon memiliki distribusi normal, atau dengan kata lain distribusi waktu respon memiliki distribusi *Lognormal*, sehingga $\ln t_{ij} \approx N\left((\beta_j - \tau_i), \sigma^2\right)$, atau dapat ditulis menjadi :

$$f(t_{ij};\beta_j;\tau_i) = \frac{1}{t_{ij}\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(\ln t_{ij} - (\beta_j - \tau_i))}{\sigma}\right]^2\right). \qquad (27)$$

Pada Persamaan 27, besarnya usaha yang dibutuhkan untuk menyelesaikan soal tes ke-j, disimbolkan menjadi β_j. Semakin besar β_j, maka banyaknya usaha yang dibutuhkan oleh peserta ke-i juga semakin besar. Parameter τ_i adalah *speed* peserta tes ke-i dalam menyelesaikan soal, semakin besar τ_i, maka semakin tinggi *speed* peserta tes ke-i, sehingga semakin sedikit waktu yang dibutuhkan untuk menyelesaikan soal tes ke-j dan σ^2 adalah varians dari t_{ij}.

Kelebihan dari model yang dikembangkan Hidayah et al. (2016) adalah, apabila dilihat dari kurvanya, model simultan yang dikembangkan memang cocok digunakan untuk *power test* (secara simulasi) sesuai dengan tujuan awal pengembangannya. Adapun kelemahannya adalah, alasan pemilihan distribusi *Lognormal* sebagai distribusi marginalnya yang hanya berdasarkan asumsi bukan pembuktian secara uji statistik (empiris), menurut Hidayah et al. (2016) diantara ketiga distribusi waktu respon yang sering digunakan oleh para peneliti (*Lognormal*, *Weibull* dan *Gamma*), distribusi *Lognormal* adalah distribusi yang paling mudah untuk diinterpretasikan. Kelemahan model simultan dari Hidayah et al. (2016) akan diperbaiki dalam penelitian ini dengan cara menerapkan data empiris selain data bangkitan simulasi, dengan data empiris uji kecocokan distribusi marginalnya dapat dilakukan terlebih dahulu sehingga didapatkan distribusi marginal yang lebih cocok.

3. Estimasi Parameter dengan Pendekatan *Bayesian*

Dalam teori estimasi, dikenal dua pendekatan yaitu pendekatan klasik dan pendekatan *Bayesian*. Pendekatan klasik sepenuhnya mengandalkan proses

inferensia pada data sampel yang diambil dari populasi. Pendekatan *Bayesian*, disamping memanfaatkan data sampel yang diperoleh dari populasi juga memperhitungkan suatu distribusi awal yang disebut *prior*. Statistika Inferensia dengan pendekatan *Bayesian* berbeda dengan pendekatan klasik. Pendekatan klasik memandang parameter θ sebagai parameter bernilai tetap, sedangkan pendekatan *Bayesian* memandang parameter θ sebagai *random variable* yang memiliki distribusi, disebut distribusi *prior*. Dari distribusi *prior*, selanjutnya dapat ditentukan distribusi *posterior* sehingga diperoleh estimator *Bayesian* yang merupakan *mean* atau modus dari distribusi *posterior*. Informasi yang diketahui tentang parameter θ sebelum pengamatan dilakukan disebut sebagai *prior* θ atau p(θ). Selanjutnya untuk menentukan distribusi *posterior* θ, yaitu $p(\theta|x)$ didasarkan pada aturan probabilitas dalam teorema *Bayes* sebagai berikut (Box & Tiao, 2011):

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \tag{28}$$

dengan $f(x) = E(f(x|\theta)) = \begin{cases} \int f(x|\theta)f(\theta)d\theta & jika\ \ \theta\ \ kontinu \\ \sum f(x|\theta)p(\theta) & jika\ \ \theta\ \ diskret \end{cases}$

dimana $f(x)$ adalah suatu konstanta yang disebut sebagai *normalized constant* (Gelman et al., 2013) yang nilainya sulit untuk dihitung, sehingga Persamaan 28 dapat ditulis sebagai suatu proposionalitas seperti berikut ini :

$$p(\theta|x) \propto f(x|\theta)p(\theta). \tag{29}$$

Proposionalitas diatas menunjukkan bahwa *posterior* adalah proporsional terhadap *likelihood* dikalikan dengan *prior* dari parameter model.

Penyelesaian masalah melalui pendekatan *Bayesian* mempunyai kelebihan dari pendekatan klasik (*likelihood*), karena pendekatan ini mengintegrasikan kondisi *prior*-nya ke dalam perhitungan selanjutnya. Keuntungan menggunakan metode *Bayesian* dibandingkan statistik secara konvensional adalah:

a. Menggunakan informasi kondisi *prior* dalam proses pengelolaan atau inferensia data.
b. Pendekatan *Bayesian* menggunakan prinsip distribusi probabilitas langsung pada parameternya. Hal ini memberikan kepercayaan yang lebih dibanding cara klasik pada umumnya.
c. Teori *Bayesian* adalah alat bantu estimasi parameter model yang dapat digunakan untuk menyelesaikan berbagai persoalan untuk berbagai situasi.
d. Pendekatan *Bayesian* merupakan cara yang sederhana untuk mempelajari parameter yang bermasalah dalam model.
e. Teori *Bayesian* memberikan cara untuk mendapatkan distribusi prediksi untuk masa mendatang. Hal ini tidak selalu mudah dikerjakan dengan cara klasik pada umumnya.

Dalam IRT, metode *Bayesian* dianggap lebih mempunyai keuntungan jika dibandingkan dengan metode *Maximum Likelihood Estimation* (MLE). Menurut Natesan (2011: 550), metode *Bayesian* mampu mengestimasi parameter pada model yang sifatnya kompleks, data yang melanggar beberapa asumsi dasar IRT, dan ukuran sampel yang sedikit. Ada banyak penelitian dengan sampel kecil dilakukan, estimasi parameter dengan *Bayesian* menunjukkan keunggulannya di

sini. Estimator *Bayesian* terbukti dapat mengestimasi parameter lebih baik daripada metode MLE (Pandey et al., 2011).

Berdasarkan teorema *Bayes*, informasi awal yang digunakan sebagai distribusi *prior* dan informasi sampel yang dinyatakan dengan fungsi *likelihood* dikombinasikan untuk membentuk distribusi *posterior*. Box &Tiao (2011) menyatakan ada beberapa tipe distribusi *prior* yang dikenal dalam metode *Bayesian*:

a. *Conjugate prior VS non conjugate prior* (Gelman et al., 2013; Tanner, 1991; Zellner, 1971) adalah *prior* yang terkait dengan pola model *likelihood* dari data.

b. *Proper prior VS Improper prior* (*Jeffreys prior*), yaitu *prior* yang terkait dengan pemberian bobot di setiap titik apakah terdistribusi secara *Uniform* atau tidak.

c. *Informative prior VS Non-Informative Prior*, yaitu *prior* yang terkait dengan diketahui atau belum diketahuinya pola/ frekuensi distribusi dari data.

d. *Pseudo Prior* menjabarkan *prior* yang terkait dengan pemberian nilai yang disetarakan dengan hasil elaborasi dari pendapat kaum *frequentist* (klasik).

Markov Chain Monte Carlo (MCMC) memudahkan permodelan yang cukup kompleks sehingga dianggap sebagai suatu terobosan dalam penggunaan metode estimasi *Bayesian* (Carlin & Chib, 1995: 473). Ada beberapa teknik yang tersedia untuk integrasi numerik, dan sebagian besar metode yang ada sangat berhubungan dengan ide yang ada pada integral *Monte Carlo* yaitu sebuah teknik integrasi yang dapat dilakukan untuk memperoleh sebuah nilai ekspektasi (μ). Jika

$\hat{\mu}$ adalah estimator *unbias* dari μ, maka persamaan matematisnya dapat ditulis sebagai :

$$\mu = \int g(x)dx = \int f(x)p(x)dx = E\left[f(x)\right]$$

$$\mu \approx \hat{\mu} \cong \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,. \qquad (30)$$

Disini p(x) didefinisikan sebagai sebuah PDF sehingga $\int p(x)dx = 1$ dan $p(x) \geq 0$. Sampel ke i disimbolkan sebagai x_i, dengan n >> 1. Nilai $x_1, x_2, ..., x_n$ dapat diperoleh secara bebas pada p(x) dalam interval (a,b) dalam bentuk yang paling sederhana dapat menggunakan distribusi *Uniform* (a,b).

Pada *Bayesian*, penggunaan MCMC dapat mempermudah analisisnya, sehingga keputusan yang diambil dari hasil analisis akan dapat dilakukan dengan cepat dan tepat. Ada dua kemudahan yang diperoleh dari penggunaan MCMC pada analisis *Bayesian* (Iriawan, 2000). Pertama, MCMC dapat menyederhanakan bentuk integral yang kompleks dengan dimensi besar menjadi bentuk integral yang sederhana dengan satu dimensi. Kedua, dengan menggunakan MCMC, estimasi densitas data dapat diketahui dengan cara membangkitkan suatu rantai *Markov* yang berurutan sebanyak n.

Salah satu algoritma dalam pendekatan *Bayesian* MCMC adalah *Gibbs Sampling* (Gelfand et al., 1990: 972). *Gibbs Sampling* merupakan algoritma untuk membangkitkan *random variable* dari distribusi marginal secara tidak langsung tanpa harus menghitung densitasnya. Algoritma *Gibbs Sampling* sangat membantu dalam proses perhitungan, dimana dapat dihindari perhitungan marginal yang sulit. Oleh karena itu algoritma *Gibbs Sampling* sangat berguna untuk perhitungan dalam

metode *Bayesian* dan metode klasik (Casella & George, 1992: 167). Penggunaan algoritma *Gibbs Sampling* pada suatu analisis data ditujukan untuk mendapatkan data tiap parameter, Θ_k secara individual dari bentuk distribusi *full conditional* semua parameter terhadap data $p(\theta_k|\theta_{k-}, x)$, dengan demikian untuk mendapatkan sampel dari tiap parameter dilakukan dengan membentuk semua parameter model menjadi sebuah vektor parameter dalam bentuk partisi yang khusus, yaitu : $\theta = (\theta_k, \theta_{k-})$.

B. Kajian Penelitian yang Relevan

Ada beberapa penelitian terdahulu yang dianggap relavan oleh peneliti. Pertama, Thissen (1979: 259) mengembangkan distribusi bersyarat keakuratan respon terhadap waktu respon Model 2PL dengan konsep *joint distribution* yang bisa diterapkan pada *power test* dan *speed test*, namun pada penelitian Ingrisone II et al. (2008: 6) terbukti model Thissen (1979: 259) hanya berlaku untuk *speed test*. Thissen (1979: 259) menggunakan distribusi *Lognormal* sebagai marginal waktu respon. Untuk metode estimasi parameternya, Thissen (1979: 260-261) menggunakan metode *Maximum Likelihood Estimation* (MLE). Data yang digunakan dalam penelitian ini adalah data empiris yaitu berupa respon dari 3 macam tes kemampuan kognitif (27 item, 36 item, 19 item) dari 78 mahasiswa S1 (36 laki-laki dan 42 perempuan) yang berasal dari Universitas Kansas. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) untuk melihat akurasinya. Dari sini dapat diketahui bahwa proses estimasi dengan pendekatan MLE berjalan dengan baik.

Penelitian Thissen (1979: 259) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Thissen (1979) adalah model 1PL sedangkan pada penelitian ini adalah Model 2PL. Kemudian untuk marginal waktu responnya sama-sama menggunakan distribusi *Lognormal*. Metode estimasi parameter yang digunakan Thissen (1979: 260-261) adalah MLE sedangkan pada penelitian ini menggunakan pendekatan *Bayesian* dengan kriteria akurasi parameter yang sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*). Kedua penelitian sama-sama melibatkan data empiris untuk pengujian model.

Kedua, Wang &Hanson (2005: 4) menggabungkan parameter waktu respon ke dalam Model 3PL yang bisa diterapkan pada *power test*, sehingga model tersebut dikenal dengan Model 4PL. Pada penelitian Wang &Hanson (2005: 8), keakuratan respon saja yang diasumsikan merupakan *random variable* sedangkan waktu respon sebagai *fixed variable*. Untuk estimasi parameternya, Wang &Hanson (2005: 8) menggunakan algoritma *Expectation Maximization* (EM). Data yang digunakan dalam penelitian ini adalah data simulasi yang dibangkitkan berdasarkan banyaknya peserta tes (1000, 2000 dan 4000) dan banyaknya soal tes (20 dan 60), dengan replikasi skenario pembangkitan data sebanyak 100 kali. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) untuk melihat akurasinya. Dari sini dapat diketahui bahwa proses estimasi dengan algoritma EM berjalan dengan baik.

Penelitian Wang &Hanson (2005) dianggap relevan karena sama-sama bisa diterapkan pada *power test*, walaupun pada penelitian Wang &Hanson (2005) hanya keakuratan respon saja yang diasumsikan merupakan *random variable* sedangkan waktu respon sebagai *fixed variable*. Untuk estimasi parameternya Wang &Hanson (2005: 18) menggunakan algoritma EM sedangkan penelitian ini menggunakan algoritma *Gibbs Sampling* dengan kriteria akurasi parameter yang sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*), tetapi jenis data yang digunakan berbeda yang mana dalam penelitian Wang &Hanson (2005: 4) menggunakan data simulasi saja sedangkan pada penelitian ini mengunakan kombinasi data simulasi dan data empiris.

Ketiga, Wang (2006: 2) memperbaiki asumsi model Wang &Hanson (2005) terkait dengan waktu respon adalah *fixed variable*, dengan cara mengalikan Model 4PL (menambahkan waktu respon ke dalam Model 3PL) sebagai distribusi bersyarat keakuratan respon terhadap waktu respon dengan distribusi *Weibull* satu parameter sebagai marginal waktu respon (konsep *joint distribution*) untuk *power test*. Menurut Wang (2006: 3), penggunaan distribusi *Weibull* satu parameter perlu dikaji lebih lanjut karena terlalu sederhana untuk menangkap realita yang ada dalam praktik pengukuran. Kemudian untuk metode estimasi parameter yang digunakan adalah dengan pendekatan *Bayesian*. Data yang digunakan dalam penelitian ini adalah data simulasi yang dibangkitkan berdasarkan banyaknya peserta tes (1000, 2000 dan 4000) dan banyaknya soal tes (20 dan 60), dengan replikasi skenario pembangkitan data sebanyak 100 kali. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*)

untuk melihat akurasinya. Dari sini dapat diketahui bahwa proses estimasi dengan pendekatan *Bayesian* berjalan dengan baik.

Penelitian Wang (2006: 2) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, tetapi model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan berbeda karena pada penelitian ini yang dikembangkan adalah Model 2PL bukan Model 3PL. Kemudian untuk marginal waktu responnya, Wang (2006: 3) menggunakan distribusi *Weibull* tetapi pada penelitian ini menggunakan distribusi *Lognormal*. Metode estimasi parameter yang digunakan sama-sama dengan pendekatan *Bayesian*, dengan kriteria akurasi parameter yang juga sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*), tetapi jenis data yang digunakan berbeda, Wang (2006: 8) dalam penelitiannya menggunakan data simulasi saja, sedangkan pada penelitian ini menggunakan kombinasi data simulasi dan data empiris.

Keempat, van der Linden (2007) mengembangkan *Hierarchical Framework Model*. Model ini dikembangkan untuk mendeteksi perilaku *cheating*. Distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Model 3PL. Kemudian untuk marginal waktu responnya menggunakan distribusi *Lognormal*. *Hierarchical Framework Model*, memiliki dua level dalam pemodelannya. Pada level pertama, parameter-parameter dalam Model 3PL dan distribusi *Lognormal* diestimasi secara terpisah. Pada level kedua, besaran hasil estimasi pada level satu tersebut dikelompokkan menjadi dua yaitu kelompok vektor *person parameter* dan kelompok vektor *item parameter* kemudian masing-

masing kelompok vektor parameter tersebut dimodelkan sendiri-sendiri dengan konsep *joint distribution*, sehingga didapatkan model distribusi *normal multivariate* dengan *person parameter* dan model distribusi *normal multivariate* dengan *item parameter*. Metode estimasi parameter yang digunakan untuk estimasi parameter tes adalah *Bayesian Markov Chain Monte Carlo* dengan algoritma *Gibbs Sampling*. Data yang digunakan dalam penelitian ini adalah data empiris yaitu hasil tes ujian CPA yang dilakukan oleh AICPA.

Penelitian van der Linden (2007: 287) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, tetapi model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan berbeda karena pada penelitian ini yang dikembangkan adalah Model 2PL bukan 3PL. Selain itu konsep pemodelannya berbeda karena *Hierarchical Framework Model*, memiliki dua level seperti yang sudah dijelaskan sebelumnya sedangkan pada penelitian ini hanya satu level. Kemudian untuk marginal waktu responnya sama-sama menggunakan distribusi *Lognormal*. Metode estimasi parameter yang digunakan juga sama-sama menggunakan pendekatan *Bayesian*. Kedua penelitian sama-sama melibatkan data empiris untuk pengujian model.

Kelima, Ingrisone et al. (2008: 15) mengembangkan model simultan untuk *power test* dengan konsep *joint distribution*, yaitu mengalikan Model 2PL sebagai distribusi bersyarat respon butir terhadap waktu respon dengan marginal waktu respon, yaitu distribusi *Lognormal* yang merupakan hasil modifikasi dari marginal waktu respon Thissen (1979: 259). Untuk estimasi *item parameter* Ingrisone et al. (2008: 47) menggunakan *Marginal Maximum Likelihood* (MML) dan untuk

estimasi *person parameter* Ingrisone et al. (2008: 47) menggunakan *Maximum a Posteriori* (MAP). Data yang digunakan dalam penelitian ini adalah data simulasi yang dibangkitkan berdasarkan banyaknya peserta tes (1000 dan 2000) dan banyaknya soal tes (20 dan 40), dengan replikasi skenario pembangkitan data sebanyak 100 kali. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) untuk melihat akurasinya. Dari sini dapat diketahui bahwa nilai bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) mendekati 0 (sangat kecil) sehingga dapat dimaknai bahwa estimasi parameter cukup akurat (besaran parameter yang diestimasi mendekati besaran parameter hasil bangkitan).

Penelitian Ingrisone et al. (2008: 15) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, dengan model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan sama-sama Model 2PL. Kemudian untuk marginal waktu responnya sama-sama menggunakan distribusi *Lognormal* hasil modifikasi. Untuk estimasi parameter yang digunakan berbeda, Ingrisone et al. (2008: 47) menggunakan MML dan MAP sedangkan pada penelitian ini menggunakan pendekatan *Bayesian Markov Chain Monte Carlo* dengan algoritma *Gibbs Sampling*. Kriteria akurasi parameter yang digunakan sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*). Untuk jenis data yang digunakan berbeda, Ingrisone et al. (2008: 47) menggunakan data simulasi saja namun pada penelitian ini data yang digunakan adalah kombinasi data simulasi dan data empiris.

Keenam, Ingrisone II et al. (2008: 3) juga mengembangkan model dengan konsep *joint distribution*. Tetapi Ingrisone II et al. (2008: 3) mengalikan Model 1PL sebagai distribusi bersyarat respon butir terhadap waktu respon dengan marginal waktu responnya yaitu distribusi *Weibull* dua parameter yang menurut Ingrisone II et al. (2008: 3) model simultan yang dihasilkan cocok untuk *power test*, padahal dengan waktu respon yang sama dan dengan kemampuan peserta tes yang berbeda menghasilkan probabilitas menjawab benar hampir sama besarnya, sehingga lebih cocok untuk *speed test* (Hidayah et al., 2016: 86). Untuk estimasi *item parameter* Ingrisone et al. (2008: 47) menggunakan *Marginal Maximum Likelihood* (MML) dan untuk estimasi *person parameter*-nya menggunakan *Maximum a Posteriori* (MAP). Data yang digunakan dalam penelitian ini adalah data simulasi yang dibangkitkan berdasarkan banyaknya peserta tes (1000 dan 2000) dan banyaknya soal tes (20 dan 40), dengan replikasi skenario pembangkitan data sebanyak 100 kali. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) untuk melihat akurasinya. Dari sini dapat diketahui bahwa nilai bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) mendekati 0 (sangat kecil) sehingga dapat dimaknai bahwa estimasi parameter cukup akurat (besaran parameter yang diestimasi mendekati besaran parameter hasil bangkitan).

Penelitian Ingrisone II et al. (2008: 3) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, tetapi model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Model 1PL sedangkan pada penelitian ini adalah Model 2PL. Kemudian

untuk marginal waktu responnya, Ingrisone II et al. (2008: 3) menggunakan distribusi *Weibull*, sedangkan penelitian ini menggunakan distribusi *Lognormal*. Untuk estimasi parameter yang digunakan berbeda, Ingrisone II et al. (2008: 24) menggunakan MML dan MAP sedangkan pada penelitian ini menggunakan pendekatan *Bayesian Markov Chain Monte Carlo* dengan algoritma *Gibbs Sampling*. Kriteria akurasi parameter juga sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*), tetapi jenis data yang digunakan berbeda, Ingrisone II et al. (2008: 13) menggunakan data simulasi sedangkan pada penelitian ini data yang digunakan adalah kombinasi data simulasi dan data empiris.

Ketujuh, Hidayah et al. (2016) mengembangkan distribusi bersyarat keakuratan respon terhadap waktu respon Model 1PL dengan konsep *joint distribution* untuk *power test*. Pada penelitian Hidayah et al. (2016), waktu respon diasumsikan sebagai *random variable* dengan marginal waktu responnya adalah distribusi *Lognormal*, dan keterbatasannya terletak pendefinisian $\mu_{\ln t_{ij}}$ yang merupakan fungsi dari *speed* dan besarnya usaha yang dibutuhkan oleh peserta tes dalam menyelesaikan tes, yang mana faktor-faktor yang mempengaruhi kedua paramater tersebut masih perlu dikaji lagi. Kemudian metode estimasi parameter yang digunakan adalah dengan *Bayesian*. Data yang digunakan dalam penelitian ini adalah data simulasi yang dibangkitkan berdasarkan banyaknya peserta tes dan banyaknya soal tes. Hasil estimasi parameternya dianalisis dengan menggunakan bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*) untuk melihat akurasinya. Dari sini dapat diketahui bahwa proses estimasi dengan pendekatan *Bayesian* berjalan dengan baik.

Penelitian Hidayah et al. (2016) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, tetapi model distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Model 1PL sedangkan pada penelitian ini adalah Model 2PL. Untuk marginal waktu responnya sama-sama menggunakan distribusi *Lognormal* tetapi pada penelitian ini nilai $\mu_{\ln t_{ij}}$ akan didefinisikan kembali untuk mengatasi kelemahan pada model simultan dari Hidayah et al. (2016). Metode estimasi parameter yang digunakan sama-sama dengan pendekatan *Bayesian* dengan kriteria akurasi parameter yang sama yaitu bias, SE (*Standard Error*), RMSE (*Root Mean Square Error*). Jenis data yang digunakan berbeda, Hidayah et al. (2016) menggunakan data simulasi saja namun pada penelitian ini data yang digunakan adalah kombinasi data simulasi dan data empiris.

Kedelapan, Fox &Marianti (2017: 244) mengembangkan *joint model* dengan konsep *Hierarchical Framework* yang dikembangkan oleh van der Linden (2007), model dikembangkan untuk mendeteksi perilaku menyimpang. Distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan adalah Model 3PL, kemudian untuk marginal waktu responnya menggunakan distribusi *Lognormal*. Metode estimasi parameter yang digunakan untuk estimasi *item parameter* adalah *Bayesian Markov Chain Monte Carlo* dengan algoritma *Gibbs Sampling*. Data yang digunakan dalam penelitian ini adalah kombinasi data simulasi dan data empiris.

Penelitian Fox &Marianti (2017: 244) dianggap relevan karena sama-sama menggunakan model simultan dengan konsep *joint distribution*, tetapi model

distribusi bersyarat keakuratan respon terhadap waktu respon yang dikembangkan berbeda karena pada penelitian ini yang dikembangkan adalah Model 2PL. Untuk konsep pemodelannya berbeda, karena model *hierarchical framework*, memiliki dua level seperti yang sudah dijelaskan sebelumnya, sedangkan pada penelitian ini hanya satu level. Persamaannya, marginal waktu responnya sama-sama menggunakan distribusi *Lognormal*, kemudian metode estimasi parameter yang digunakan juga sama-sama menggunakan pendekatan *Bayesian* dan data yang digunakan sama-sama kombinasi dari data simulasi dan data empiris.

C. Kerangka Pikir

Di Indonesia tes terkomputerisasi sudah mulai diterapkan, seperti pada pelaksanaan ujian nasional yang lebih dikenal sebagai UNBK (Ujian Nasional Berbasis Komputer) yang sudah mulai dilaksanakan pada tahun 2014 secara online, dan juga ada tes seleksi masuk perguruan tinggi seperti di Program Pascasarjana Unversitas Negeri Yogyakarta (PPs UNY) yang beralih dari *Paper Based Testing* (PBT) menjadi *Computerized-Based Testing* (CBT) mulai tahun 2017. Penerapan tes terkomputerisasi memiliki banyak keuntungan, seperti fleksibilitas manajemen tes, peningkatan keamanan tes, peningkatan motivasi dalam peningkatan literasi komputer, dan efesiensi waktu (Georgiadou et al., 2006: 3). Pelaksanaan tes terkomputerisasi juga dapat memberikan informasi yang sifatnya kompleks, karena selain informasi tentang pola respon peserta tes, tes terkomputerisasi juga dapat memberikan informasi tentang waktu respon yang digunakan peserta tes dalam menjawab suatu butir soal. Catatan waktu respon yang

dihasilkan oleh tes terkomputerisasi ini akan menjadi penting untuk dipertimbangkan dalam suatu ujian atau tes, apalagi tujuannya adalah seleksi, karena catatan waktu respon dapat memberikan informasi tambahan bagi lembaga penyelenggara tes selain ukuran keakuratan jawaban peserta tes, yaitu berupa informasi ukuran *speed* peserta tes mengerjakan soal, sehingga dapat mempermudah pelaksanaan seleksi, peserta tes diputuskan lulus atau gagal atau masuk kategori cepat atau lambat. Waktu respon di sini juga diperhitungkan sebagai indikator kemampuan peserta tes selain keakuratan menjawab soal (Bolsinova et al., 2017: 1126).

Ukuran keakuratan jawaban peserta tes dapat diperoleh dengan konsep IRT yang memodelkan *person parameter* dan *item parameter* (tingkat kesulitan, daya beda, dan *pseudo guessing*) hanya berdasarkan pola respon peserta tes yaitu berupa jawaban salah atau benar dari peserta tes. Kemudian pemodelan waktu respon hadir untuk memperbaiki hasil estimasi parameter kemampuan peserta tes pada IRT. Oshima (1994: 200) menyatakan bahwa *speed* dari peserta tes dalam mengerjakan suatu butir soal dapat mempengaruhi estimasi parameter kemampuan peserta tes dan estimasi butir soalnya. Hornke (2000: 175) juga berpendapat bahwa waktu respon penting untuk dipertimbangkan dalam suatu tes, karena waktu respon adalah indikator untuk ciri-ciri kepribadian peserta tes yang harus dibedakan dari skor.

Pemodelan waku respon yang pernah dilakukan meliputi tiga cara yang berbeda, (1) pemodelan waktu respon secara eksklusif, (2) melakukan analisis terpisah pada model waktu respon dan model keakuratan skor, dan (3) menggabungkan waktu respon dan keakuratan skor ke dalam satu model secara

simultan (Entink 2009: 47-48). Dari ketiga cara tersebut, yang menjadi fokus pada penelitian ini adalah cara pemodelan ketiga, yaitu model simultan antara IRT dengan waktu respon.

Setiap model yang dikembangkan tersebut mempunyai spesifikasi tersendiri dalam penerapannya, ada kekurangan juga kelebihannya, sehingga pengembangan model simultan yang memperhitungkan waktu respon perlu terus dilakukan untuk memperbaiki model yang sudah ada sehingga penerapannya menjadi lebih realistis. Pengembangan model simultan yang dilakukan oleh para ahli menurut Schnipke &Scrams (2002: 11), pada umumnya didasarkan pada tiga hal yaitu, (1) tipe tes, apakah *speed test* atau *power test*, (2) asumsi distribusi waktu respon yang digunakan, dan (3) asumsi hubungan antara keakuratan respon dan *response speed*.

Tidak ada model standar yang dapat digunakan untuk memodelkan semua kasus waktu respon dalam berbagai tes. Karena faktanya distribusi waktu respon dari berbagai item soal bisa beragam, sehingga model tertentu tidak akan berlaku untuk segala macam tes (Ranger & Kuhn, 2012: 31). Setiap model waktu respon yang dikembangkan pasti ada kelebihan dan kekurangannya. Oleh karena itu pengembangan model yang memperhitungkan waktu respon perlu terus dilakukan untuk memperbaiki model yang sudah ada, agar didapatkan model yang semakin realistis dalam penerapannya.

Penelitian ini bermaksud untuk untuk mengembangkan Model Logistik 2 Parameter dengan variabel random waktu respon secara simultan yang cocok untuk tes terkomputerisasi. Pengembangan model dilakukan dengan cara mengkaji kembali model dari Hidayah et al. (2016). Pada awalnya, Hidayah et al. (2016)

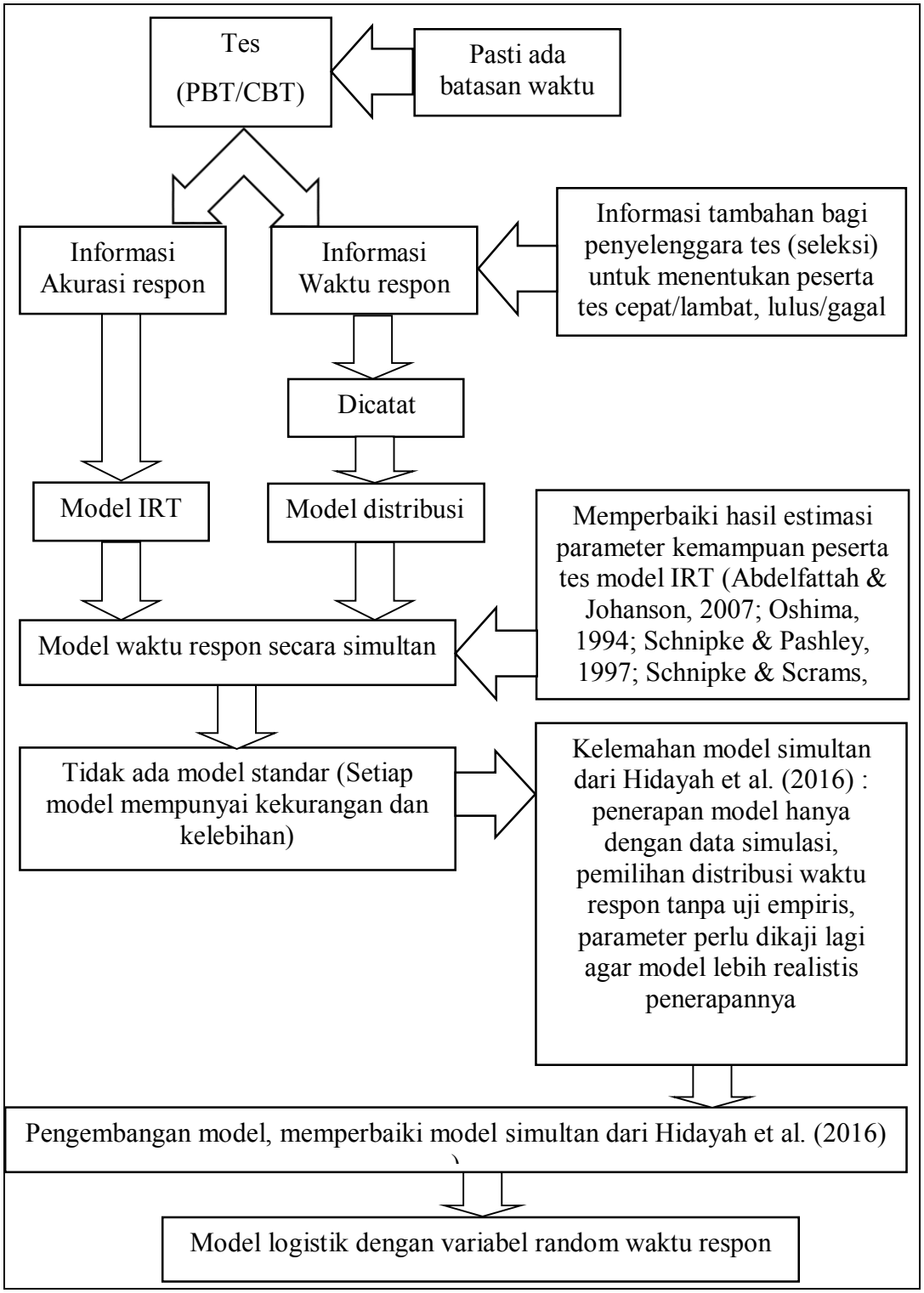
mengembangkan model dengan cara memperbaiki kelemahan dari model Ingrisone II et al. (2008: 3) yaitu penggunaan model untuk *power test*, padahal menurut Hidayah et al. (2016: 86) model Ingrisone II et al. (2008) terbukti lebih cocok untuk *speed test* bukan *power test* karena modelnya lebih cocok atau lebih tepat untuk memprediksi butir-butir soal yang mudah untuk dijawab dengan benar dengan waktu terbatas (cepat).

Kelebihan sekaligus kekurangan dari penelitian Hidayah et al. (2016) adalah pengujian modelnya dengan data simulasi. Di satu sisi, *output* dari data simulasi ini bersifat ideal karena dibangkitkan dengan algoritma tertentu sesuai dengan model. Tetapi di sisi lain, hasil dari simulasi ini bisa jadi tidak akurat karena teknik ini bukan proses optimasi dan tidak menghasilkan jawaban tetapi hanya menghasilkan sekumpulan *output* dari sistem pada berbagai kondisi yang berbeda sehingga keakuratannya masih harus dipertanyakan. Oleh karena itu pada penelitian ini akan dilihat bagaimana analisis model dengan menggunakan data empiris. Hal ini akan menjadi menarik karena hasil data empiris sangat memungkinkan tidak seideal hasil dari data simulasi. Penggunaan data empiris pada penelitian akan sangat berguna dalam pengujian keakuratan model simultan yang dihasilkan, sejauh apa model tersebut mampu menjelaskan kondisi *real* dalam praktik pengukuran yang mempertimbangkan waktu respon. Adapun data empiris yang akan digunakan dalam penelitian ini, berasal dari tes terkompuerisasi (CBT) pada seleksi masuk PPs UNY tahun 2017. Pada tes ini selain jawaban benar-salah, juga dicatat waktu respon dari setiap mahasiswa dalam menyelesaikan tiap butir soal yang dia kerjakan.

Adanya beberapa parameter model yang dikembangkan Hidayah et al. (2016) yang pada praktiknya (secara empiris) juga perlu dikaji lagi, misalnya parameter kelambatan (d_j), dan parameter besarnya usaha yang dibutuhkan untuk menyelesaikan soal tes ke-j (β_j) pada pendefinisian *mean log natural* waktu respon ($\mu_{\ln t_{ij}}$) model *Lognormal*. Oleh karena itu pada penelitian ini perlu diseleksi lagi parameter-parameter apa saja yang dapat dimasukkan ke dalam model, karena secara teoritis waktu respon dipengaruhi oleh banyak faktor. Pada penelitian Halkitis (1996: 1) menunjukkan bahwa adanya hubungan positif antara waktu respon dengan tingkat kesulitan soal, daya beda soal dan panjang soal. Zenisky &Baldwin (2006: 2) juga menunjukkan dalam penelitiannya bahwa ada hubungan antara waktu respon dengan tingkat kesulitan soal, daya beda, tingkat kompleksitas soal, konten soal dan perbedaan kelompok soal. Dari beberapa parameter tersebut dapat dipertimbangkan lagi untuk ditambahkan atau dicantumkan sebagai faktor yang berpengaruh terhadap waktu respon pada model simultan antara model logistik dengan variabel random waktu respon, tetapi tentu saja harus logis pada praktiknya, sehingga didapatkan model yang lebih realistis penerapannya.

Kelemahan selanjutnya, Hidayah et al. (2016) menentukan secara langsung distribusi marginal dari waktu responnya adalah *Lognormal* dengan $\ln t_{ij} \approx N\left(\left(\beta_j - \tau_i\right), \sigma^2\right)$, berdasarkan pertimbangan bahwa nilainya positif, bentuknya *positive skewed* dan lebih mudah diinterpretasikan, walaupun masih ada distribusi lain yang dianggap cocok dengan karakter waktu respon seperti *Weibull* dan *Gamma* (Lindsey, 2004: 197). Kelemahan dari model (Hidayah et al., 2016)

akan diperbaiki dalam penelitian ini dengan cara menerapkan data empiris selain data bangkitan simulasi, dengan data empiris uji kecocokan distribusi marginalnya dapat dilakukan terlebih dahulu sehingga didapatkan distribusi marginal yang lebih cocok.



Gambar 6. Kerangka Berpikir Pengembangan Model

D. Pertanyaan Penelitian

Lingkup pembahasan dalam disertasi ini akan menjawab pertanyaan-pertanyaan berikut:

1.a Bagaimana formulasi model distribusi bersyarat keakuratan respon terhadap waktu respon?

1.b Bagaimana formulasi model waktu respon berdasarkan karakteristik distribusi marginal yang sesuai?

1.c Bagaimana bentuk *joint distribution* antara model distribusi bersyarat keakuratan respon terhadap waktu respon dengan model distribusi marginal dari waktu respon?

2.a Distribusi *likelihood* apakah yang sesuai dengan data hasil respon peserta tes?

2.b Distribusi *prior* apakah yang sesuai dengan setiap parameter yang ada dalam model?

2.c Bagaimana bentuk distribusi *posterior* yang dihasilkan?

2.d Algoritma apakah yang digunakan untuk membangkitkan nilai parameter berdasarkan distribusi *posterior* yang dihasilkan?

2.e Bagaimana hasil estimasi parameter dari model simultan yang dihasilkan?

2.f Kriteria apa saja yang digunakan untuk keakuratan hasil estimasi parameter model simultan yang dihasilkan?

2.g Bagaimana keakuratan hasil estimasi parameter model simultan yang dihasilkan untuk masing-masing skenario?

2.h Bagaimana perbandingan keakuratan hasil estimasi parameter model simultan yang dihasilkan berdasarkan jumlah soal (m) dan jumlah peserta tes (n)?

3.a Kriteria apa saja yang digunakan dalam membandingkan keakuratan hasil estimasi parameter antara model simultan yang dihasilkan, model simultan dari Hidayah et al. (2016), dengan model *Item Response Theory* (IRT)?

3.b Bagaimana perbandingan keakuratan hasil estimasi parameter kemampuan (theta/Θ) antara model simultan yang dihasilkan, model simultan dari Hidayah et al. (2016), dengan model *Item Response Theory* (IRT)?

4.a Kriteria apa saja yang digunakan untuk membandingkan kecocokan antara model simultan yang dihasilkan, model simultan dari Hidayah et al. (2016), dengan model IRT pada data empiris?

4.b Bagaimana perbandingan kecocokan antara model simultan yang dihasilkan, dengan model simultan dari Hidayah et al. (2016) dan model IRT?

4.c Bagaimana hasil analisis butir soal dengan menggunakan model yang memiliki kecocokan dengan data empiris paling tinggi?