MODEL RASCH SEBAGAI KERANGKA ACUAN PENYUSUNAN ALAT UKUR

THE RASCH MODEL AS A FRAME OF REFERENCE IN CONSTRUCTING INSTRUMENTS

Asrijanty
Pusat Penilaian Pendidikan Kemdikbud
Jalan Gunung Sahari Raya No 4 Jakarta Pusat
e-mail: asrijanty@gmail.com

Naskah diterima tanggal: 31/12/2013; Dikembalikan untuk revisi tanggal: 15/01/2014; Disetujui tanggal: 12/02/2014

Abstrak: *Artikel ini merupakan kajian fungsi model Rasch sebagai kerangka acuan penyusunan alat ukur dalam ilmu sosial, khususnya dalan bidang pendidikan dan psikologi. Kajian ini didasari argumen bahwa usaha untuk memperoleh alat ukur yang memberi informasi yang valid dapat dilakukan dengan memanfaatkan model Rasch dalam analisis data. Tujuan kajian ini dimaksudkan untuk mengkaji: 1) karakteristik model Rasch sebagai model pengukuran; 2) penggunaan model Rasch dalam pengembangan tes. Kajian dilakukan dengan membahas karakteristik dan paradigma model Rasch disertai dengan perbandingan dengan model pengukuran lain, khususnya model logistik dua parameter (2PL) dan model logistik tiga parameter (3 PL), termasuk kritik yang sering diajukan terhadap model Rasch. Kajian penggunaan model Rasch dalam pengembangan tes dilakukan dengan membahas implikasi dan aplikasi model Rasch dalam analisis data untuk pengembangan instrumen. Hasil kajian menunjukkan bahwa: 1) model Rasch mempunyai karakteristik dan paradigma yang berbeda dari model 2 PL dan model 3 PL; 2) sesuai dengan karakteristik dan paradigma model Rasch, fungsi model Rasch dalam analisis data pengembangan instrumen, yaitu untuk memberi arah dan mendeteksi atau mendiagnosa adanya masalah pada instrumen.*

Kata kunci: *Model Rasch, Pengukuran Sosial, Model 2 PL, Model 3 PL, Item Response Theory*

Abstract: *This article is a review of the function of the Rasch model as a frame of reference in constructing instruments in social sciences, particularly for education and psychology. The argument of this article is that efforts to obtain an instrument which provides valid information can be done by utilizing the Rasch model in data analysis. The aims of this article are to examine: 1) the characteristics of the Rasch model as a measurement model; 2)the utilization of the Rasch model in test development. The examination of the Rasch model encompasses its characteristics and its paradigm in comparison with other measurement models, namely two-parameters logistic (2 PL model) model and three parameters logistic models (3 PL model;) and criticism of the Rasch model. The examinaton utilization of the Rasch model in test development includes its implication and application of the Rasch model in test development. This study shows that: 1) the characteristics and the paradigm of the Rasch model differ from the 2 PL model and 3 PL model ; 2)in line with its characteristics and its paradigm, the function of the Rasch model in test development is to guide and to diagnose problems in instrument.*

Keywords: *Rasch model, Social Measurement, 2 PL model, 3 PL model, Item Response Theory*

Pendahuluan

Pengukuram dalam bidang ilmu sosial pada umumnya dipandang lebih rumit daripada pengukuran dalam ilmu nonsosial seperti Fisika. Hal ini karena hal yang diukur dalam ilmu sosial tidak sejelas seperti pada Fisika. Pada ilmu sosial hal yang diukur pada umumnya merupakan hal yang abstrak, suatu variabel yang diperkirakan ada dan mempengaruhi perilaku manusia. Contoh populer hal yang diukur dalam dunia pendidikan

dan psikologi adalah inteligensi. Inteligensi tidak dapat dilihat, namun diyakini ada dan bervariasi serta mempengaruhi beberapa aspek perilaku manusia, misalnya dalam pemecahan masalah, mempelajari hal baru, dan dalam melakukan penyesuaian sosial. Oleh karena itu, alat ukur atau instrumen yang digunakan merupakan faktor penting yang menentukan kerberhasilan pengukuran dan bagaimana memperoleh alat ukur yang memberi informasi yang valid merupakan topik penting pada pengukuran dalam dunia pendidikan dan psikologi.

Salah satu alat ukur yang sering digunakan dalam dunia pendidikan dan psikologi adalah tes tertulis. Tes tertulis secara umum dibedakan menjadi dua kategori, yaitu tes dengan pilihan jawaban tersedia, dan tes tanpa pilihan jawaban, sehingga peserta tes harus memberi jawaban secara tertulis. Untuk pengukuran dalam skala besar, tes dengan format beberapa pilihan jawaban, dikenal sebagai tes pilihan ganda atau *multiple choice test*, lebih sering digunakan karena pertimbangan faktor efisiensi dan konsistensi hasil penskoran. Namun menghasilkan tes pilihan ganda yang memberi informasi yang valid tidak mudah; usaha yang serius perlu diupayakan.

Berbagai faktor dapat mempengaruhi kualitas tes sehingga suatu tes tidak memberi informasi yang valid. Misalnya untuk tes prestasi belajar, bila soal menguji materi yang belum diajarkan, maka soal menjadi sulit untuk siswa, sehingga dapat terjadi siswa menjawab dengan menebak, atau bahasa yang digunakan tidak dipahami oleh sebagian siswa, sehingga soal demikian akan merugikan sebagian siswa. Bila hal demikian terjadi, maka tes yang diberikan tidak memberikan informasi mengenai prestasi atau kemampuan peserta tes yang sebenarnya. Faktor-faktor tersebut disebut gangguan pengukuran atau *measuerment disturbances*(Smith & Plackner, 2009). Para ahli (Andrich, 1988; 2004; Smith & Plackner, 2009; Wright & Stone, 1979) menunjukkan bahwa model Rasch dapat digunakan dalam analisis data untuk mengidentifikasi masalah-masalah tersebut.

Analisis data dapat dilakukan menggunakan pendekatan teori tes klasik dan teori tes modern. Model Rasch merupakan salah satu model dengan pendekatan teori tes modern yang sering digunakan dalam analisis data. Namun, sering dijumpai model Rasch dipilih hanya dengan pertimbangan lebih mudah diaplikasikan daripada model lain, misalnya model logistik dua parameter (model 2 PL) atau model logistik tiga parameter (model 3 PL) dan bukan dengan pertimbangan kekhususan atau kelebihannya. Hal ini menyebabkan ketika model Rasch digunakan, analisis yang dilakukan tidak optimal memanfaatkan kelebihan model Rasch, sehingga informasi mengenai soal tes yang juga kurang maksimal. Hal ini sangat disayangkan karena penyusunan suatu alat ukur merupakan tahapan kritikal dan menyerap banyak sumber daya dan dana. Di sisi lain, model Rasch sering dikritik karena kesederhanannya, yaitu karena kurang kompleksnya model Rasch dibandingkan model 2 PL atau 3 PL.

Masalah yang dikemukakan sebelumnya tampaknya bersumber dari kurangnya pengenalan terhadap model Rasch. Untuk itu, pengkajian model Rasch perlu dilakukan agar karakteristiknya dan kelebihannya sebagai model pengukuran dapat diketahui dan diapresiasi. Tujuan kajian ini adalah: 1) mengkaji karakteristik model Rasch sebagai model pengukuran dengan membandingkan model 2 PL dan 3 PL; 2) mengkaji penggunaan model Rasch dalam pengembangan tes. Kajian dilakukan dengan membahas karakteristik dan paradigma model Rasch disertai dengan perbandingan dengan model pengukuran lain, khususnya model logistik dua parameter (2PL) dan model logistik tiga parameter (3 PL), termasuk kritik yang sering diajukan terhadap model Rasch. Kajian penggunaan model Rasch dalam pengembangan tes dilakukan dengan membahas implikasi penggunaan model Rasch dan paradigmanya dalam analisis data disertai dengan ilustrasi aplikasi analisis data dengan meng-gunakan model Rasch.

Kajian Literatur

Model Rasch dan *Item Response Theory*

Secara umum pengukuran dalam bidang ilmu sosial mengenal dua pendekatan, yaitu teori tes klasik dan teori tes modern, atau sering juga disebut model *latent trait* (*latent trait models)*. Terdapat beberapa model *latent trait,* namun model yang populer adalah model Rasch, model 1 PL, 2 PL dan 3 PL. Istilah lain untuk merujuk model

latent trait adalah *Item Reponse Theory* (*IRT*). Sebagian ahli, misalnya Hambleton, Swaminathan & Rogers (1991), Embretson & Reise (2000), mengelompokkan model Rasch dalam IRT, namun sebagian lain membedakan model Rasch dari IRT, misalnya Ryan (1993) dan Andrich (2004). Dalam tulisan ini, untuk memudahkan pembahasan, model Rasch tidak dikelompokkan sebagai model IRT, sedangkan model 1 PL, 2 PL, dan 3 PL dikategorikan sebagai model IRT.

Model Rasch dan model 1 PL mempunyai kesamaan, yaitu hanya satu parameter yang diestimasi sehingga sering dianggap sebagai model yang sama. Namun sebenarnya hal ini kurang tepat karena kedua model tersebut merupakan dua model yang berbeda (Andrich, 2004; Ryan, 1983). Model Rasch dikembangkan oleh Georg Rasch sedangkan IRT dikembangkan antara lain oleh Frederick Lord dan Alan Birnbaum (Lord & Novick, 1968; Bock, 1997). Model 1 PL merupakan model khusus dari 2 PL bila parameter daya beda dijadikan konstan.

Georg Rasch, seorang matematikawan Denmark, mengembangkan model tersebut pada awal tahun 1950-an ketika ia ditugaskan untuk meneliti perkembangan kemampuan membaca sejumlah anak yang mengalami kesulitan membaca. Tugas ini merupakan tugas yang sulit karena tes yang digunakan untuk setiap anak pada beberapa kali pengetesan berbeda dan pada saat itu tidak ada metode analisis yang memungkinkan dilakukan perbandingan dengan tes yang berbeda. Rasch kemudian mengembangkan model yang dapat mengukur perkembangan kemampuan membaca anak meskipun tes digunakan pada beberapa kali pengetesan berbeda. Model ini kemudian disebutnya sebagai *Multiplicative Poisson Model* (Andrich, 2005; Rasch 1960/1980)

Menurut Rasch (1977) model yang dikembangkannya sesuai dengan konsep pengukuran yang ilmiah, yaitu pengukuran yang melibatkan perbandingan yang objektif atau dikenal sebagai *invariant comparison*. Dalam konteks pengukuran bidang pendidikan dan psikologi, hal ini berarti perbandingan antarindividu tidak tergantung pada soal yang digunakan untuk membandingkan individu tersebut, dan perbandingan antarsoal tidak tergantung pada subjek atau individu yang digunakan untuk membandingkan soal. Secara lengkap perbandingan yang dikemukakan Rasch sebagai berikut.

"The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion" (Rasch, 1961).

Perbandingan objektif tersebut sesuai dengan persyaratan instrumen ideal yang dikemukakan oleh Louis Thurstone. Menurut Thurstone (1928) fungsi suatu instrumen atau alat ukur hendaknya tidak dipengaruhi oleh objek yang diukur. Sebagai contoh, meteran (alat ukur panjang) diharapkan tetap menunjukkan panjang yang konsisten, tidak tergantung pada objek yang diukur. Meteran menjadi tidak valid bila memberi informasi panjang yang berbeda untuk objek yang berbeda.

Namun, perbandingan objektif tersebut tidak berarti bahwa karakteristik individu atau peserta tes, misalnya jenis kelamin, usia, tingkat pendidikan tidak penting. Perbandingan yang objektif terbatas pada kerangka acuan yang ditetapkan (Andrich, 1985; 1988; Rasch 1977). Yang dimaksud kerangka acuan yang ditetapkan adalah suatu spesifikasi dari gabungan beberapa elemen, yaitu agen atau kumpulan soal, objek atau kelompok individu, dan reaksi sebagai hasil dari kontak antara suatu agen dan suatu objek. Dengan demikian perbandingan objektif berlaku hanya untuk suatu kelompok individu pada sekelompok soal pada situasi tes tertentu. Bila suatu tes dimaksudkan untuk sekelompok individu pada kelompok umur tertentu (misalnya 17-50) yang meliputi perempuan dan laki, maka perbandingan yang objektif tidak dapat diharapkan untuk individu yang di luar batasan umur tersebut. Meskipun demikian, secara empiris tetap perlu dilakukan pengecekan apakah data menunjukkan perbandingan yang objektif pada keluarga yang dimaksud. Bila hasil analisis menunjukkan, misalnya, estimasi tingkat kesulitan soal untuk kelompok perempuan dan laki-laki berbeda secara

signifikan maka berarti perbandingan yang objektif tidak terbukti. Ini yang kemudian dikenal sebagai *DIF (Differential Item Functioning)*.

Rasch memformulasi suatu model yang memenuhi persyaratan perbandingan objektif dengan menggunakan fungsi probabilitas (Andrich, 1988). Untuk data dengan respon dikotomi, probabilitas seseorang menjawab suatu soal dengan benar merupakan suatu fungsi dari perbedaan parameter orang (*ability* atau kemampuan) dan parameter soal (*difficulty* atau tingkat kesulitan). Fungsi tersebut diekspresikan dalam skala logaritma (logits).

Model untuk data dengan respon dikotomi disebut *Simple Logistic Model* atau *Dichotomous Rasch Model*, yang dieskpresikan sebagai berikut:

$$\Pr\{X_{ni} = \chi\} = \frac{\exp\ \chi(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \qquad (1)$$

di mana $\chi = 1$ atau 0.

Formula (1) dapat dijabarkan dalam dua bentuk:

$$\Pr\{X_{ni} = 1\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)};$$

$$\Pr\{X_{ni} = 0\} = \frac{1}{1 + \exp(\beta_n - \delta_i)} \qquad (2)$$

di mana $\Pr\{X_{ni} = 1\}$ adalah probabilitas seorang *n* menjawab soal *i* dengan benar, $\Pr\{X_{ni} = 0\}$ adalah probabilitas seorang *n* menjawab soal *i* tidak benar, β_n adalah lokasi atau *ability* seorang *n* pada suatu variabel laten, dan δ_i adalah lokasi atau tingkat kesulitan suatu soal pada suatu variabel.

Formula di atas menunjukkan hubungan antara parameter orang dan soal (β_n and δ_i) adalah aditif dan yang menentukan probabilitas seorang n untuk menjawab benar soal i hanya perbedaan antara β_n and δ_i. Hal ini berbeda dengan model 2 PL dan 3 PL. Pada kedua model ini probabilitas menjawab benar tidak hanya ditentukan oleh perbedaan antara *ability* orang dan tingkat kesulitan soal, tetapi juga oleh parameter lain (daya beda dan menebak).

Pada model 3 PL menebak dipandang sebagai parameter soal. Oleh karena itu, menebak diestimasi bersama dengan parameter soal lainnya (Birnbaum, 1968). Formula model 3PL sebagai berikut:

$$\Pr\{X_{ni=1}\} = c_i + (1 - c_i)P \qquad (3)$$

di mana $P = [\exp(\alpha_i(\beta_n - \delta_i))] / [1 + \exp(\alpha_i(\beta_n - \delta_i))]$, α_i merupakan daya beda soal *i* dan c_i adalah parameter menebak soal *i*.

Dari formula (3) dapat dilihat bahwa individu atau peserta tes dengan *ability* yang paling rendah mempunyai probabilitas sebesar c_i untuk menjawab benar. Oleh karena $c_i = 1/C$ di mana C adalah jumlah pilihan jawaban, maka semakin banyak jumlah pilihan jawaban semakin kecil probabilitas untuk menjawab benar soal i.

Bila $c_i = 0$, i=1,2,...,I, persamaan (3) akan menjadi model 2 PL, yaitu :

$$\Pr\{X_{ni} = 1\} = \frac{\exp \alpha_i(\beta_n - \delta_i)}{1 + \exp \alpha_i(\beta_n - \delta_i)} \qquad (4)$$

dan bila kemudian $\alpha_i = 1$, i=1,2,...,I, persamaan tersebut akan menjadi model 1 PL, yang secara matematis sama dengan model Rasch.

Karakteristik Model Rasch

Karaketristik model Rasch pada bagian ini meliputi *item charactertistic curve (ICC),* daya beda, perbandingan yang objektif, daya beda, dan menebak (*guessing*). Dalam diskusi pada masing-masing aspek tersebut, juga disampaikan tinjauan aspek dari model 2 PL dan 3 PL.

Item *Characteristic Curve (ICC)*

Probabilitas seseorang untuk menjawab benar suatu soal berdasarkan formula 1-4 dapat ditunjukkan oleh grafik yang disebut *Item Characteristic Curve (ICC).* Pada model Rasch, ICC dari soal-soal berbentuk paralel. Hal ini berbeda dengan model IRT, khususnya 2 PL dan 3 PL di mana *ICC* dari beberapa soal dapat menyilang atau tidak paralel. Ketika *ICC* dari soal-soal tersebut tidak paralel, maka urutan tingkat kesulitan soal berbeda untuk kelompok dengan kemampuan yang berbeda. Hal ini berarti persyaratan perbandingan objektif tidak terpenuhi karena perbandingan kesulitan antara dua soal tidak independen dari kemampuan orang. Wright (1997) memberi ilustrasi keadaan tersebut seperti terlihat pada Gambar 1 dan Gambar 2.

Gambar 1 menunjukkan *ICC* soal-soal paralel yang diperoleh dari model Rasch. Dapat dilihat bahwa urutan tingkat kesulitan soal tetap sama

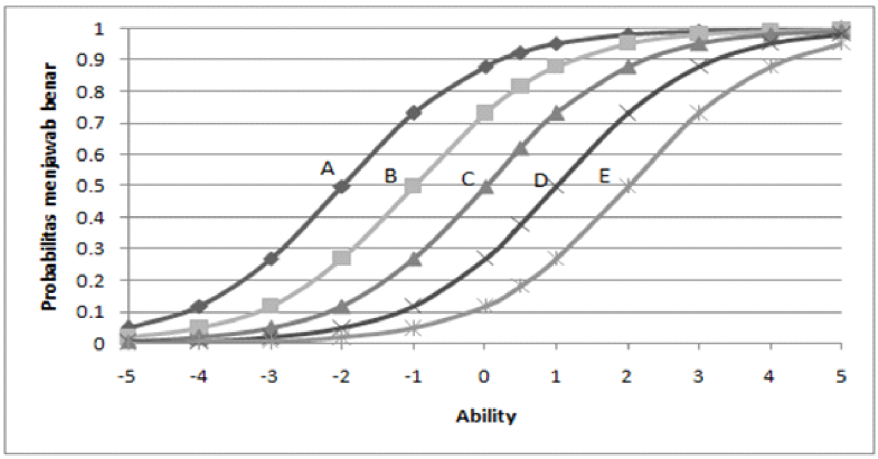
pada setiap lokasi *ability*. Sebagai contoh untuk semua peserta baik dengan kemampuan rendah, misalnya logit -2.0, kemampuan sedang, misalnya logit 0.5 maupun kemampuan tinggi, misalnya logit 3.0 urutan soal dari yang paling sulit untuk ketiga kelompok dengan *ability* yang berbeda adalah sama, yaitu A, B, C, D, E.

Gambar 2 menunjukkan *ICC* soal-soal yang tidak paralel dari model 3 PL. Dapat dilihat bahwa urutan tingkat kesulitan soal tidak sama pada setiap lokasi. Untuk kelompok dengan kemampuan rendah, misalnya logit -2.0 urutan tingkat kesulitan soal adalah A, B, C, D, E; urutan tingkat kesulitan soal untuk kemampuan sedang, misalnya logit 0.5 adalah B, A, C, D, E ; sedangkan untuk kemampuan tinggi, misalnya logit 3.0 adalah B, C, A, D, E. Bila kelima soal tersebut dirakit sebagai satu tes, maka soal-soal berdasarkan model Rasch akan membentuk satu variabel, sementara soal-soal dengan model 3 PL yang tidak menunjukkan arah yang sama tidak dapat dipandang sebagai satu variabel.

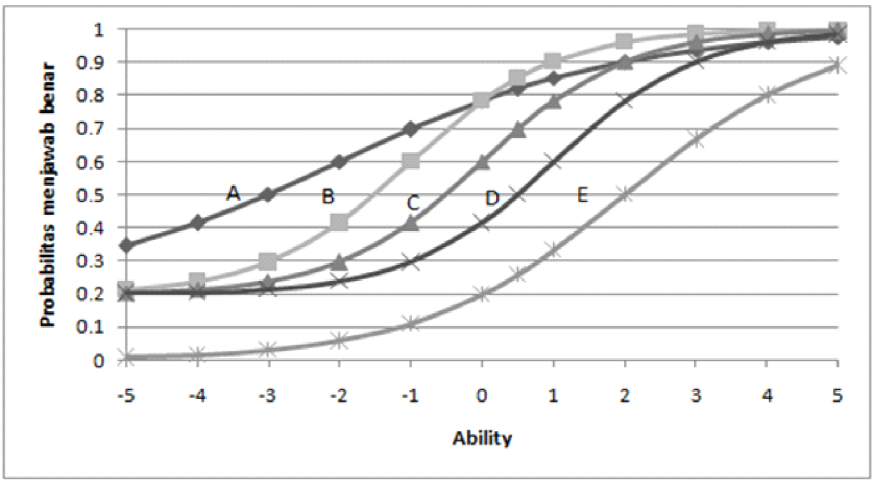
Oleh karena itu, *ICC* yang paralel untuk soal-soal dengan respon dikotomus merupakan properti model Rasch yang khusus dan mencerminkan properti perbandingan yang objektif (Wright, 1997*)* dan properti tersebut merupakan syarat untuk mencapai pengukuran yang fundamental (Embretson & Reise, 2000).

Perbandingan yang Objektif (*Invariant Comparison*)

Perbandingan yang objektif dalam model Rasch didukung oleh dua komponen, yaitu kecukupan statistik (*statistical sufficiency*) dan pemisahan parameter orang dan soal (Andrich, 2005). Realisasi kecukupan statistik pada model, yaitu total skor seseorang merupakan statistik yang



Gambar 1 ICC Soal dengan Respon Dikotomus pada Model Rasch



Gambar 2 ICC soal dengan respon dikotomus pada model 3 PL

mencukupi untuk mengestimasi *ability* orang tersebut dan total skor soal merupakan statistik yang mencukupi untuk mengestimasi tingkat kesulitan soal (Andrich, 1988). Dengan kata lain, total skor seseorang merupakan satu-satunya informasi yang diperlukan untuk mengestimasi *ability* orang tersebut. Demikian pula total skor soal merupakan satu-satunya informasi yang diperlukan untuk mengestimasi tingkat kesulitan soal. Sebagai konsekuensi dari kecukupan statistik ini individu yang menyelesaikan soal yang sama dan memperoleh total skor yang sama akan memperoleh estimasi *ability* yang sama. Demikian pula soal dengan total skor yang sama akan memperoleh estimasi tingkat kesulitan soal yang sama.

Pemisahan parameter orang dan soal berarti perbandingan tingkat kesulitan antara dua soal tidak tergantung pada kemampuan individu yang menjawab soal-soal tersebut dan perbandingan kemampuan antarindividu tidak tergantung pada tingkat kesulitan soal yang digunakan (Andrich, 1988). Pemisahan ini juga berasal dari kecukupan statistik.

Properti perbandingan yang objektif tidak dimiliki oleh model 2 PL dan 3 PL. Hal ini karena estimasi *ability* tidak hanya ditentukan oleh total skor individu, dan estimasi tingkat kesulitan soal tidak hanya ditentukan oleh total skor soal, pola respon juga mempengaruhi estimasi. Artinya estimasi *ability* individu tergantung pada soal mana ia menjawab benar. Dua individu yang memperoleh skor benar pada soal yang berbeda mungkin akan memperoleh estimasi yang berbeda, tergantung pada daya beda soal yang dijawab benar.

Pendukung model Rasch memandang kecukupan statistik sebagai properti yang esensial suatu model pengukuran. Sementara pendukung model 2 PL dan 3 PL meskipun mengakui kecukupan statistik sebagai properti yang berguna, namun memandang properti tersebut tidak esensial dan bukan satu-satunya dasar untuk menghasilkan teori tes. Menurut mereka, konsep yang utama adalah informasi tes dan memaksimalkan informasi dari suatu tes merupakan tujuan utama. Untuk mencapai hal tersebut soal-soal perlu diskor dengan memperhatikan bobot masing-masing (Hambleton, 1994).

Daya Beda

Dalam hal daya beda soal, model *item response theory* menganut paham teori tes klasik, yaitu soal dengan daya beda tinggi adalah yang diharapkan dan menjadi dasar pemilihan soal. Semakin tinggi daya beda soal semakin baik soal tersebut. Sebaliknya pada model Rasch soal yang dianggap baik adalah yang *fit* dengan model, yaitu sesuai dengan prediksi model, tidak rendah dan tidak terlalu tinggi. Soal dengan daya beda yang rendah maupun terlalu tinggi tidak dikehendaki, karena dapat mengindikasikan adanya dimensi atau faktor lain yang mempengaruhi kemampuan yang diukur atau dependensi antarsoal (Andrich, 1988; Masters, 1988).

Daya beda dikatakan terlalu tinggi bila performa atau skor kelompok bawah (kelompok dengan kemampuan rendah) lebih rendah dari yang diprediksi model dan performa kelompok atas lebih tinggi dari yang diprediksi model. Dalam tes prestasi, hal ini dapat terjadi karena adanya perbedaan kesempatan belajar antara kelompok atas dan bawah. Sebagai contoh, dalam studinya Masters (1988) menemukan soal matematika yang diujikan pada kelompok siswa yang mengambil mata pelajaran matematika umum dan matematika murni mempunyai daya beda yang sangat tinggi, karena materi yang dipelajari siswa pada matematika umum lebih rendah tingkatnya dibandingkan materi pada matematika murni.

Faktor lain yang juga dapat mempengaruhi adalah kelelahan, kecepatan, dan *test wiseness*. Soal yang diletakkan pada akhir tes dapat menunjukkan daya beda yang sangat tinggi karena kelompok bawah tidak hati-hati atau lelah dalam menjawab soal-soal akhir. Hal ini seperti yang ditemukan Yen (1980) dalam studinya, yang membandingkan daya beda soal ketika diletakan di awal dan di akhir tes. Daya beda soal yang tinggi juga dapat terjadi pada soal yang kurang jelas atau menimbulkan multiinterpretasi yang merugikan kelompok bawah dan menguntungkan kelompok atas (Masters, 1988).

Contoh-contoh tersebut menunjukkan daya beda yang lebih tinggi daripada yang diprediksi oleh model memberikan informasi kemungkinan adanya masalah pada soal. Masters (1988) mengemukakan soal dengan daya beda yang

terlalu tinggi menunjukkan bias yang memberi keuntungan pada kelompok atas.

Pada perkembangan terakhir, daya beda juga diperhitungkan dalam model Rasch. Namun daya beda yang diestimasi bukan daya beda soal, melainkan daya beda sejumlah soal dalam suatu perangkat tes atau secara lebih spesifik dikatakan daya beda dalam suatu kerangka acuan (*specified frame of reference*). Konsep ini dikemukakan dan diteliti oleh Stephen Humphry (2010; 2011). Humphry menunjukkan daya beda tersebut dapat muncul karena faktor lingkungan seperti karakteristik soal yang digunakan, karakteristik peserta tes dan situasi tes. Oleh karena itu, daya beda ini disebut juga *group person discrimination*. Daya beda ini dapat diperhitungkan dan diestimasi sehingga tidak mencemari estimasi kemampuan orang dan soal. Humphry dan Andrich (2008) menunjukkan bahwa, memasukkan parameter tersebut tidak mengubah properti model Rasch; kecukupan statistik tetap dapat dipertahankan. Seperti telah dikemukakan sebelumnya secara matematis model Rasch merupakan bentuk khusus dari model 2PL, dalam hal ini nilai daya beda (α) adalah 1. Namun sebenarnya nilai tersebut dapat berbeda, tidak harus 1. Bila akan dilakukan penyetaraan dari dua pengukuran (dari kerangka acuan yang berbeda), parameter tersebut perlu diperhi-tungkan karena tiap pengukuran mungkin mempunyai daya beda yang tidak sama. Pembahasan secara rinci mengenai parameter tersebut tidak akan dilakukan dalam tulisan ini. Pembaca yang berminat dapat merujuk pada artikel Humphry (2010; 2011) dan Humphry dan Andrich (2008).

Menebak (*Guessing)*

Menebak dapat terjadi bila soal menggunakan format pillihan ganda. Bila seseorang menebak maka estimasi kemampuan orang dan tingkat kesulitan soal menjadi kurang akurat, karena tidak menggambarkan hal yang sebenarnya, yaitu estimasi kemampuan orang menjadi lebih tinggi dan soal menjadi lebih mudah (Rogers, 1999; Waller, 1973).

Pada model 3 PL menebak atau parameter c (*guessing*) diestimasi bersama dengan parameter lain (Birnbaum, 1968). Ciri dari model 3 PL adalah seseorang dengan estimasi kemampuan yang sama dengan tingkat kesulitan soal tidak mempunyai probabilitas 50% memperoleh skor benar pada soal tersebut. Soal dengan tingkat kesulitan sama tetapi dengan estimate parameter menebak yang berbeda mempunyai probabilitas yang berbeda dalam memperoleh skor benar (Embretson & Reise, 2000; Lord, 1980). Menurut Choppin(1985) hal ini tidak logis. Lord (1968) mengemukakan hal yang menjadi masalah dengan model 3 PL adalah adanya asumsi bahwa semua orang menebak dengan derajat yang sama pada semua soal padahal kenyataannya belum tentu demikian.

Model 3 PL juga mempunyai keterbatasan dalam mengestimasi parameter, yaitu memerlukan jumlah subjek/sampel dan jumlah soal yang besar. Namun bahkan dengan besarnya data, parameter menebak seringkali tidak dapat diestimasi (Rogers, 1999).

Model Rasch tidak memandang menebak sebagai parameter soal. Dalam model Rasch faktor yang berperan dalam estimasi hanya kemampuan orang dan tingkat kesulitan soal (Andrich, Marais, & Humphry, 2012). Oleh karena itu, pada prinsipnya menebak seharusnya tidak mempengaruhi respon subjek. Bila ya, maka berarti ada faktor lain yang mempengaruhi estimasi kemampuan orang dan tingkat kesulitan soal. Waller (1973) mengemukakan bila ada unsur menebak maka akan menghasilkan estimasi yang kurang atau tidak konsisten dengan model. Pengecekan apakah menebak terjadi dapat dilakukan dengan cara yang dikemukakan Andrich, Marais dan Humphry (2012).

Berdasarkan model Rasch menebak merupakan hal yang tidak dikehendaki. Oleh karena itu, harus dihindari. Menurut Wright (1997) salah satu cara untuk mengurangi atau menghindari munculnya perilaku menebak adalah dengan menyajikan soal yang sesuai dengan kemampuan peserta tes.

Kritik terhadap Model Rasch

Model Rasch telah dikritik sejak tahun 1970-an, terutama karena kesederhanaan model tersebut, khususnya karena tidak mengikutsertakan parameter daya beda dalam model. Dengan hanya satu parameter yang diperhitungkan, secara umum kesesuaian (*fit*) data dengan model

Rasch tidak sebaik kesesuaian dengan model yang lebih kompleks.

Bock (1997) mengemukakan bahwa kenyataan di lapangan menunjukkan hampir tidak mungkin menjumpai soal-soal dengan daya beda yang sama. Selain itu, informasi mengenai daya beda soal diperlukan dalam menyusun suatu tes untuk menjamin reliabilitas tes yang tinggi dan distribusi skor yang diinginkan. Divgi (1986) bahkan menyimpulkan bahwa model Rasch tidak cocok digunakan untuk tes dengan format pilihan ganda. Berdasarkan studinya ia menemukan bahwa model dengan lebih banyak parameter *fit* lebih baik dibandingkan model Rasch.

Embretson dan Reise (2000), meskipun mengakui kelebihan model Rasch, tidak merekomendasikan model Rasch digunakan untuk semua situasi. Menurut mereka, untuk alat ukur psikologi adanya daya beda soal yang bervariasi tidak dapat dihindarkan. Oleh karena itu, mereka merekomendasikan untuk menggunakan model yang lebih kompleks daripada model Rasch. Mereka beragumen bahwa hal ini akan mencegah didropnya soal-soal yang penting yang mungkin akan menyebabkan berubahnya konstruk yang akan diukur. Meskipun untuk hal ini dapat diperdebatkan bahwa dasar dalam membuang soals seharusnya tidak hanya berdasarkan kriteria statistik.

Paradigma Rasch dan Paradigma Model Statistik

Uraian sebelumnya menunjukkan bahwa kritik terhadap model Rasch pada umumnya adalah karena model Rasch mempunyai kemungkinan lebih kecil untuk *fit* dengan data. Dengan kata lain, model Rasch tidak dapat menerangkan data dengan baik. Hal ini menunjukkan adanya pandangan bahwa model pengukuran yang baik adalah bila *fit* atau dapat menerangkan data. Data dalam hal ini dipandang sebagai hal yang selalu benar. Menurut Andrich (2004) pandangan demikian merupakan paradigma tradisional model statistik, yaitu fungsi dari suatu model adalah untuk menerangkan data. Oleh karena itu, model yang dipilih adalah model yang *fit* dengan data.

Sebaliknya, menurut paradigma Rasch, model pengukuran berfungsi sebagai kerangka acuan dalam penyusunan alat ukur (Andrich, 2004). Oleh karena itu, suatu model tidak dipilih untuk menerangkan data, tetapi model pengukuran dipilih untuk memberi arah sekaligus menunjukkan apakah data memenuhi kriteria model pengukuran. Bila data hasil tes tidak *fit* atau tidak memenuhi kriteria model pengukuran, maka perlu dilakukan pengecekan dan penjelasan ketidaksesuaian atau tidak terpenuhinya kriteria. Dengan demikian model pengukuran juga berfungsi sebagai alat diagnostik.

Hal ini seperti yang dilakukan Georg Rasch pada awal tahun 1950-an. Ketika ia menemukan ketidakkonsistenan antara model yang dikembangkannya dengan data tes inteligensi, ia tidak mengubah model, namun melakukan pengecekan pada data dan kemudian mengusulkan perubahan pada konstruksi soal sehingga menghasilkan data yang *fit* atau konsisten dengan model (Andrich, 2005).

Andrich (2004) mengemukakan bahwa kritik terhadap model Rasch bersumber pada adanya perbedaan paradigma tersebut. Mereka yang mengritik model Rasch berpandangan bahwa suatu model berfungsi menerangkan data sementara berdasarkan paradigma Rasch, suatu model pengukuran berfungsi sebagai kerangka acuan dalam penyusunan alat ukur.

Adanya paradigma yang berbeda ini juga diakui oleh Hambleton (1994; 2000) dan Hambleton dan Cook (1977). Hambleton, yang menganut pandangan bahwa fungsi model pengukuran adalah untuk menerangkan data, mengemukakan bahwa dalam memilih suatu model perlu dipertimbangkan apakah data tes memenuhi asumsi model atau tidak. Bila tidak, maka perlu mempertimbangkan model lain yang lebih sesuai dengan data.

Pembahasan

Implikasi Menggunakan Model Rasch dan Paradigmanya

Menggunakan model Rasch dan paradigmanya berarti menggunakan model Rasch sebagai kerangka acuan dalam penyusunan alat ukur. Sebagai implikasi, properti model Rasch digunakan sebagai kriteria dalam mengevaluasi hasil tes. Bila hasil tes atau data tidak memenuhi kriteria, maka hal yang perlu dilakukan adalah memeriksa atau mengecek data, dan bukan mencari model lain

untuk menerangkan data. Hal ini sesuai dengan paradigma Rasch bahwa data tidak selalu benar. Model digunakan untuk mengecek apakah data yang dihasilkan sesuai dengan kriteria model. Tidak terpenuhinya kriteria model memberi arah pada perbaikan yang perlu dilakukan.

Analisis untuk melihat kesesuaian antara model dan hasil tes sudah umum dilakukan. Pada umumnya statistik dari analisis *fit* yang dijadikan dasar untuk menentukan *fit* tidaknya suatu soal. Sebagai contoh, menggunakan program Quest (Adams & Khoo, 1996) soal dikatakan *fit* bila *fitmeansquare* berkisar antara 0.7–1.3; dengan program RUMM (Andrich, Sheridan, & Luo, 2010) menggunakan *residual fit statistic* dengan rentang -2.5 – +2.5. Umumnya analisis selesai setelah dilakukan pemilihan soal. Soal yang memenuhi kriteria statistik tersebut digunakan atau dimasukkan dalam bank soal, sedangkan soal-soal yang tidak memenuhi kriteria dibuang. Analisis untuk mengetahui penyebab tidak *fit*-nya (*misfit*) soal belum umum dilakukan. Demikian pula analisis *DIF* juga belum dijadikan analisis yang rutin.

Analisis menggunakan model Rasch hanya untuk memperoleh statistik *fit* kurang optimal, karena kurang memanfaatkan model Rasch sebagai kerangka acuan dan alat diagnostik dalam penyusunan alat ukur. Pemanfaatan model Rasch akan lebih optimal, antara lain dengan melakukan analisis *fit* lebih mendalam, analisis unidimensi dan independen respon, dan analisis *DIF.*

Analisis *Fit*

Informasi mengenai kesesuaian data dan model tidak hanya dapat diperoleh dari statistik *fit* tetapi juga dari hasil yang berupa grafik, dalam hal ini *Item Characteristic Curve (ICC).* Kedua informasi ini perlu diperhatikan agar diperoleh informasi yang lengkap. Statistik memang lebih praktis karena memberikan informasi mengenai *fit* dalam bentuk angka, namun statistik *fit* dipengaruhi oleh jumlah subjek data analisis. Dengan jumlah subjek yang besar, tes statistik menjadi lebih sensitif dalam mendeteksi *misfit,* sehingga dapat terjadi suatu soal secara statistik menunjukkan *misfit* namun *ICC* menunjukkan *misfit* yang terjadi sangat kecil atau tidak berarti. Selain itu, statistik *fit* bersifat umum, tidak memberi gambaran yang lebih spesifik, misalnya apakah *misfit* yang terjadi pada semua kelompok ataukah hanya pada kelompok tertentu misalnya kelompok dengan kemampuan tinggi.

Berdasarkan model Rasch, suatu soal dikatakan *misfit* tidak hanya bila daya bedanya rendah, tetapi juga bila daya beda suatu soal terlalu tinggi dibandingkan daya beda yang diprediksi oleh model, seperti ditunjukkan oleh *ICC.* Hal ini berbeda dengan model 2 PL atau 3 PL dan teori tes klasik yang menginginkan daya beda yang setinggi mungkin. Seperti telah dikemukakan sebelumnya, terlalu tingginya daya beda dapat mengindikasikan bias yang merugikan kelompok bawah. Selain itu, daya beda soal yang tinggi juga dapat mengindikasikan dependensi antarsoal. Dependensi antarsoal ini selanjutnya dapat dicek dengan melihat korelasi residual antarsoal seperti yang akan dibahas pada analisis independen respon.

Selain itu, *ICC* juga dapat menunjukkan indikasi menebak, yakni bila proporsi jawaban benar subjek pada kelompok bawah lebih tinggi dari proporsi yang diprediksi oleh model. Metode untuk mendeteksi menebak secara lebih rinci dapat dilihat pada Andrich, Marais dan Humphry (2012).

Analisis Unidimensi dan Independen Respon

Model Rasch merupakan model pengukuran unidimensi yang mensyaratkan respon yang independen secara statistik (*statistical independence)*. Properti tersebut tidak khas properti model Rasch, karena juga merupakan properti model IRT (Andrich, 1988; Hambleton, Swaminathan & Rogers, 1991). Disebut model unidimensi karena hanya satu dimensi dari individu yang menjadi fokus, yaitu *ability* pada suatu dimensi (β). Bila tes tidak unidimensi berarti ada faktor lain dari individu selain β yang mempengauhi responnya terhadap soal tes. Marais and Andrich (2008) menyebutnya sebagai *trait dependence*.

Independen respon secara statistik berarti probabilitas dari suatu *outcome* tidak tergantung pada *outcome* lain. Dalam kaitannya dengan respon seseorang terhadap lebih dari satu soal berarti respon seseorang terhadap satu soal tidak tergantung pada responnya terhadap soal lain (Andrich, 1988). Independen secara statisik tidak

terpenuhi bila respon terhadap suatu item dipengaruhi oleh responnya terhadap item sebelumnya. Marais and Andrich (2008) menyebut hal ini sebagai *response dependence*. Dalam literatur *trait dependence* dan *response dependence* pada umumnya tidak dibedakan, keduanya dikategorikan sebagai *local independence* (Marais & Andrich, 2008).

Pengecekan adanya pelanggaran *local independence* dapat dilakukan dengan menghitung korelasi residual antaritem. Dependensi dicurigai bila korelasi residual antaritem relatif tinggi (Tennant & Gonaghan, 2007; Zenisky, et al., 2002). Korelasi residual yang tinggi menunjukkan adanya dependensi antaritem yang tidak dapat diterangkan oleh parameter *ability* orang dan item. Hal ini berarti selain variabel yang diukur ada dimensi atau faktor lain yang mempengaruhi respon individu terhadap item tertentu. Relatif tingginya korelasi residual dapat mengindikasikan *trait dependence* atau *response dependence.*

Analisis *Differential Item Functioning (DIF)*

Seperti telah dikemukakan sebelumnya, model Rasch mempunyai properti perbandingan objektif dalam suatu kerangka acuan. Implikasi dari properti tersebut adalah setiap subset respon harus menghasilkan parameter item yang sama. Oleh karena itu, perlu dilakukan pengecekan apakah estimasi dari kelompok-kelompok pada suatu kerangka acuan adalah sama. Sebagai contoh bila perempuan dan laki-laki termasuk dalam suatu kerangka acuan yang sama maka perlu dicek apakah estimasi parameter kelompok laki-laki dan kelompok perempuan sama. Bila terdapat perbedaan yang signifikan maka berarti perbandingan antarkelompok jenis kelamin tidak objektif atau disebut *differential item functioning (DIF).*

Ilustrasi Hasil Aplikasi Model Rasch Sebagai Kerangka Acuan

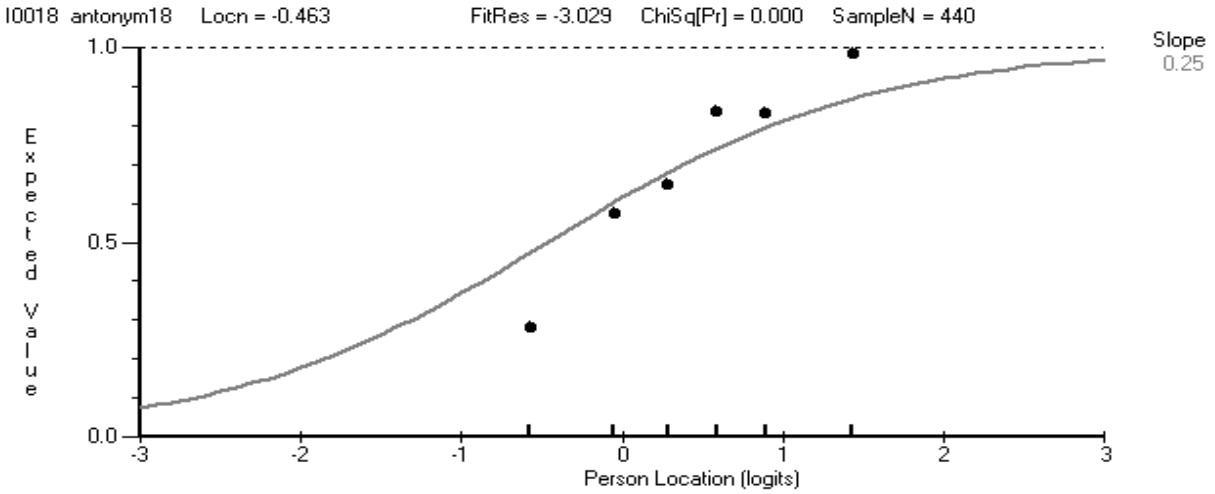
Untuk memberi gambaran penggunaan model Rasch sebagai kerangka acuan penyusunan alat ukur, disajikan ilustrasi hasil analisis data dengan menggunakan model Rasch. Data yang digunakan adalah data Tes Bakat Skolastik (TBS) dari Pusat Penilaian Pendidikan, Balitbang Kemdikbud. Tes Bakat Skolastik terdiri atas tiga subtes yaitu: Verbal, Kuantitatif, dan Penalaran. Oleh karena sifatnya hanya ilustrasi, maka hanya dipilih hasil analisis yang relevan.

Soal dengan Daya Beda yang Sangat Tinggi

Seperti telah dikemukakan sebelumnya, dalam menentukan *fit* atau kesesuaian dengan model Rasch, perlu melihat tidak hanya statistik *fit* residual, tetapi juga *ICC* sehingga adanya masalah pada soal dapat teridentifikasi. Sebagai contoh disajikan soal nomor 18 yang secara statistik menunjukkan *fit* residual -3.029. Dari statistik tersebut gambaran mengenai apa yang terjadi tidak begitu jelas. Bila melihat *ICC* soal 18 seperti pada Gambar 3, dapat dilihat bahwa soal 18 menunjukkan daya beda atau diskriminasi yang lebih tinggi daripada yang diharapkan oleh model. Dapat dilihat bahwa performa (*mean*) kelompok paling bawah (kelompok dengan *ability* rendah) diwakili oleh titik hitam paling bawah, berada di bawah kurva, sementara *mean* kelompok paling atas berada di atas kurva. Idealnya titik-titik (*mean* kelompok) tersebut berada atau mendekati kurva. Berdasarkan model Rasch daya beda yang sangat tinggi tidak dikehendaki karena mengindikasikan bias yang merugikan kelompok bawah. Selain itu, daya beda soal yang sangat tinggi juga dapat mengindikasikan dependensi antarsoal. Hal ini berbeda dengan model 2 PL dan 3 PL yang menghendaki soal dengan daya beda yang setinggi-tingginya. Pengecekan lebih lanjut telah dilakukan untuk melihat kemungkinan bias dan dependensi antarsoal, namun dalam hal ini tidak ditemukan bukti adanya dependensi antaritem. Namun, soal 18 menunjukkan DIF terjadi antara kelompok master dan doktoral seperti didiskusikan pada hasil analisis *DIF*.

Soal dengan Indikasi Adanya Perilaku Menebak

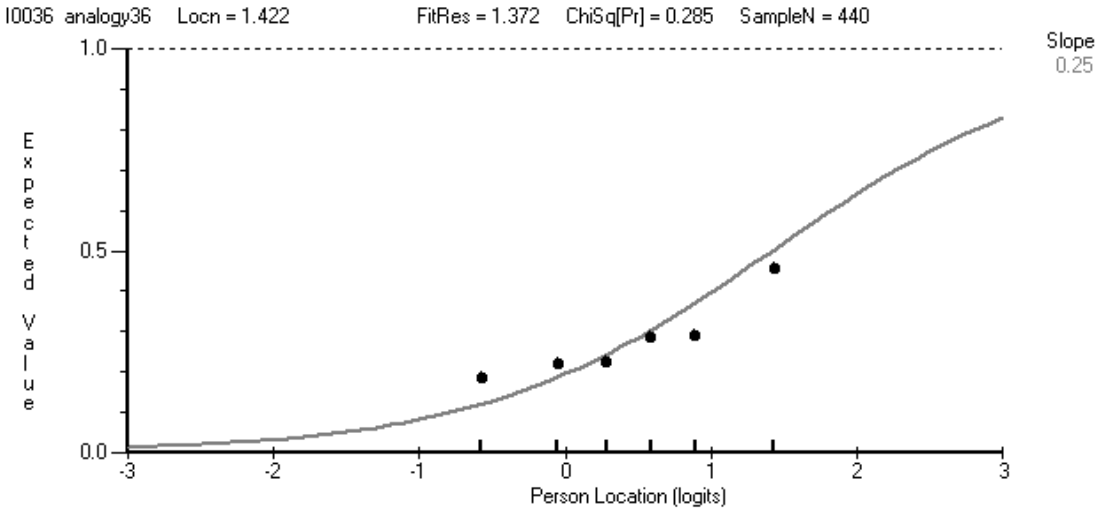
ICC juga dapat menunjukkan adanya perilaku menebak. Hal ini dapat dilihat dari lebih tingginya *mean* (performa) kelompok dari yang diharapkan oleh model dan lebih rendahnya *mean* kelompok atas daripada yang diharapkan model. Hal ini seperti ditunjukkan oleh ICC soal 36 pada Gambar 4. *ICC* hanya menunjukkan indikasi menebak, pengecekan lebih lanjut perlu dilakukan untuk memperoleh bukti atau konfirmasi adanya



Gambar 3 ICC Soal 18 yang Menunjukkan Daya Beda Tinggi

menebak. Pengecekan yang dilakukan mengacu pada Andrich, Marais dan Humphry (2012), menunjukkan bukti secara statistik yang mendukung hipotesis adanya perilaku menebak pada soal 36.

Untuk memahami mengapa soal 36 terdapat indikasi menebak, dilakukan telaah terhadap isi soal 36 (Gambar 5). Soal 36 merupakan soal Analogi. Subjek diminta memilih jawaban, sehingga pasangan kata pertama mempunyai hubungan yang sama dengan pasangan kedua. Hubungan



Gambar 4 *ICC* Soal 36 yang mengindikasikan adanya perilaku menebak

---

Petunjuk: Pilihlah satu dari lima pilihan jawaban itu untuk melengkapi pasangan yang terletak di belakang tanda sama dengan (=) itu, sehingga pasangan itu mempunyai hubungan **yang sama** atau **serupa** dengan pasangan kata yang terdapat di depan tanda sama dengan (=).

36. ULAT : bumi = ELANG : ….
a. udara
b. angkasa
c. langit *
d. awan
e. angin

Gambar 5 Isi soal no. 36

pada soal 36 adalah tempat di mana binatang tertentu banyak ditemukan atau dilihat. Pasangan kata pertama adalah ulat di bumi, sedangkan pasangan kedua ditanyakan elang di mana. Jawaban yang benar atau kunci adalah langit, c. Masalah pada soal ini adalah pilihan jawaban a dan b juga dapat menjadi jawaban yang benar. Langit, udara, angkasa dapat digunakan secara bergantian meskipun dapat diperdebatkan bahwa pada konteks soal 13, langit, pilihan jawaban merupakan yang paling tepat. Adanya kemungkinan jawaban lain yang benar tampaknya membuat soal ini menjadi sulit sehingga peserta tes cenderung menebak.

Soal yang Menunjukkan Dependensi Respon

Adanya dependensi respon dapat dicek dengan menghitung korelasi residual antarsoal. Pada analisis TBS ditemukan adanya korelasi residual yang tinggi antara dua soal, yaitu soal 56 dan 58, sebesar 0.683. Konfirmasi dependensi antarsoal yang dilakukan dengan prosedur Andrich (2010) menunjukkan bukti bahwa kedua soal tersebut secara signifikan dependen. Mencermati substansi kedua soal tersebut, dependensi antarsoal yang terjadi dapat dimengerti karena kedua soal tersebut adalah soal deret angka dengan menggunakan pola yang sama, yaitu bilangan prima. Oleh karena kedua soal ini dependen, baik secara statistik dan substansi, maka hanya salah satu soal yang dapat dimasukkan ke dalam set tes karena kedua soal memberi informasi yang sama. Isi kedua soal tersebut disajikan pada Gambar 6.

Soal Menunjukkan *Differential Item Functioning (DIF)*

DIF dapat ditunjukkan oleh statistik maupun *ICC.* Gambar 7 menunjukkan ICC soal 18 untuk dua kelompok, yaitu calon mahasiswa doktoral dan calon mahasiswa master. Pada soal ini 18, performa calon mahasiswa kelompok calon mahasiswa doktoral lebih tinggi daripada kelompok calon mahasiswa master calon mahasiswa. Dengan kata lain soal 18 lebih sulit untuk kelompok calon mahasiswa master daripada kelompok doktoral. Hal ini menunjukan soal 18 tidak berfungsi sama untuk tingkat pendidikan yang berbeda. Pada bagian terdahulu telah ditunjukkan bahwa soal 18 merupakan soal dengan daya beda yang tinggi. Dengan demikian, soal 18 menunjukkan daya beda tinggi dan sekaligus *DIF*. Hal ini menunjukkan bahwa soal 18 tidak dapat digunakan untuk membandingkan kedua kelompok tersebut. Soal tersebut dapat digunakan untuk perbandingan hanya pada masing-masing kelompok.

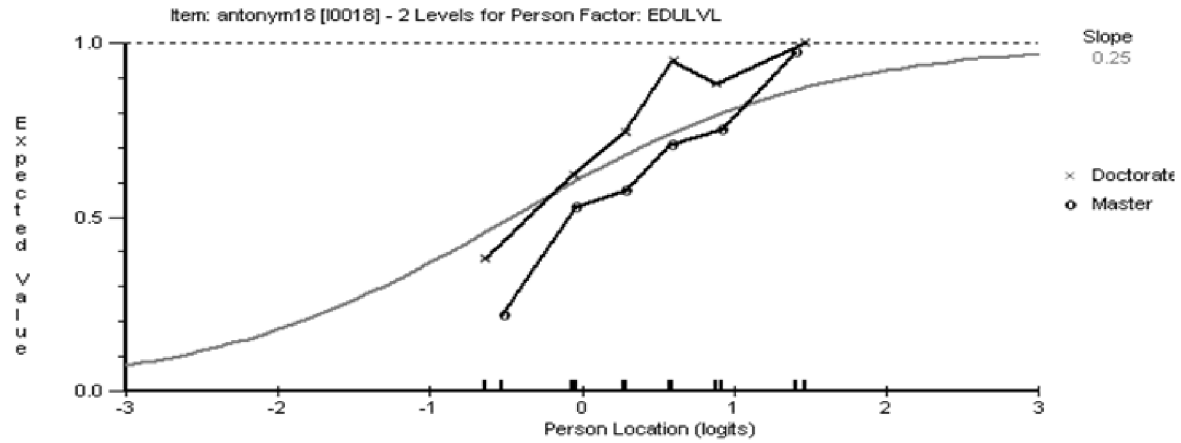
Mencermati isi soal 18 (Gambar 8), soal 18 adalah soal antonim, yang menanyakan kata dengan arti berlawanan. Kata reduksi merupakan kata serapan bahasa asing. Diharapkan untuk tingkat pascasarjana kata tersebut bukan kata yang asing lagi, namun analisis menunjukkan kata ini tidak berfungsi sama, secara umum soal 18 lebih sulit untuk kelompok calon mahasiswa master daripada kelompok calon mahasiswa doktoral.

Petunjuk: Pilihlah salah satu jawaban untuk menyelesaikan **deretan angka** itu, sesuai dengan prinsip yang mendasari..

56. 2 4 7 12 19 ...
a. 27
b. 28
c. 30*
d. 31
e. 32

58. 5 7 10 15 22 ...
a. 35
b. 33*
c. 31
d. 30
e. 29

Gambar 6 Isi soal no 56 dan 58



Gambar 7 Soal 18 yang menunjukkan DIF untuk tingkat pendidikan

---

Petunjuk: Pilihlah satu kemungkinan jawaban yang mempunyai arti yang **berlawanan** dengan kata yang dicetak dengan huruf kapital.

18. REDUKSI
    a. laba
    b. pembagian
    c. permintaan
    d. penambahan*
    e. keuntungan

Gambar 8 Isi soal 18

Simpulan dan Saran

Simpulan

Simpulan dari hasil kajian sebagai berikut. Pertama, ada perbedaan antara model Rasch dan model 2 PL dan model 3 PL, baik secara matematis (karakteristik) maupun filosofis (paradigma), meskipun semua termasuk dalam kelompok teori tes modern. Model Rasch mempunyai kekhususan dalam menghasilkan perbandingan yang objektif. Hal ini karena komponen kecukupan statistik (*statistical sufficiency*) dan pemisahan parameter orang dan soal pada model Rasch. Dengan kecukupan statistik pada model, total skor seseorang merupakan satu-satunya informasi yang diperlukan untuk mengestimasi *ability* orang tersebut dan total skor soal merupakan satu-satunya informasi yang diperlukan untuk mengestimasi tingkat kesulitan soal. Sebagai konsekuensi dari kecukupan statistik ini individu yang menyelesaikan soal yang sama dan memperoleh total skor yang sama akan memperoleh estimasi *ability* yang sama. Demikian pula soal dengan total skor yang sama akan memperoleh estimasi tingkat kesulitan soal yang sama. Pada pemisahan parameter orang dan soal berarti perbandingan tingkat kesulitan antara dua soal tidak tergantung pada kemampuan individu yang menjawab soal-soal tersebut dan perbandingan kemampuan antarindividu tidak tergantung pada tingkat kesulitan soal yang digunakan. Properti perbandingan yang objektif tidak dimiliki oleh model 2 PL dan 3 PL. Hal ini karena estimasi *ability* pada model 2 PL dan model 3 PL tidak hanya ditentukan oleh total skor individu, dan estimasi tingkat kesulitan soal tidak hanya ditentukan oleh total skor soal; pola respon juga mempengaruhi estimasi. Artinya estimasi *ability* individu tergantung pada soal mana ia menjawab benar. Dua individu yang memperoleh skor benar pada soal yang berbeda mungkin akan memperoleh estimasi yang berbeda, tergantung pada daya beda soal yang dijawab benar. Karakteristik model Rasch yang lain adalah daya beda dan menebak tidak diestimasi, tetapi dicoba dikontrol, karena dipandang sebagai gangguan pengukuran.

Kedua, sebagai implikasi karakteristik dan paradigma model Rasch, penggunaan model Rasch dalam analisis data untuk pengembangan instrumen adalah sebagai kerangka acuan penyusunan alat ukur. Tujuan analisis data dengan model Rasch adalah melihat kesesuaian antara data dengan model. Bila ada ketidak-sesuaian, maka dilakukan pengecekan terhadap data, bukan mencari model yang dapat mene-rangkan data. Oleh karena itu, model Rasch juga berfungsi sebagai alat diagnostik untuk mende-teksi masalah pada data. Bila ditemui adanya soal yang mempunyai daya beda sangat tinggi dilakukan analisis lebih lanjut dan dipelajari. Bila terbukti ada bias atau dependensi antarsoal dapat dilakukan perbaikan pada instrumen. Begitu pula soal yang menunjukkan indikasi ditebak, dapat dipelajari dan diperbaiki. Dengan demikian, model Rasch berfungsi sebagai kerangka acuan dalam penyusunan atau pengembangan alat ukur.

Saran

Sejalan dengan simpulan di atas dapat disarankan agar dalam analisis data dengan menggunakan model Rasch, analisis dilakukan dengan tujuan memperoleh informasi yang sebanyak-banyaknya mengenai tes. Analisis data hendaknya dilakukan menyeluruh dengan memperhatikan tidak hanya hasil berupa statitistik soal, tetapi juga ICC dan substansi soal. Dengan demikian, berbagai gangguan pengukuran yang mungkin timbul karena konstruksi tes yang kurang baik atau materi yang kurang sesuai dapat diidentifikasi dan usaha perbaikan soal tes dapat dilakukan.

Pustaka Acuan

Adams, R. J., & Khoo, S.-T. 1996. *Quest: The Interactive Test Analysis System (Version 2.1)*. Victoria, Australia: ACER.

Andrich, D. 1985. An Elaboration of Guttman Scaling with Rasch Model for Measurement. In N. Brandon-Tuma (Ed.), *Sociological Methodology* (pp. Chapter 2, 33-80). San Francisco: Jossey-Bass.

Andrich, D. 1988. *Rasch Models for Measurement*. Newbury Park: Sage.

Andrich, D. 2004. Controversy and the Rasch Model: A Characteristic of Incompatible Paradigm. *Medical Care, 42*(1), 7-16.

Andrich, D. 2005. Rasch, Georg. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (Vol. 3, pp. 299-306). Amsterdam: Academic Press.

Andrich, D., Sheridan, B. E., & Luo, G. 2010. *RUMM2030: A Window Program for Rasch Unidimensional Models for Measurement.* Perth, Australia: RUMM Laboratory.

Andrich, D., Marais, I., & Humphry, S. 2012. Using a Theorem by Andersen and the Dichotomous Rasch Model to Assess the Presence of Random Guessing in Multiple Choice Items. *Journal of Educational and Behavioral Statistics, 37*(3), 417-442. doi: 10.3102/1076998611411914

Birnbaum, A. 1968. Some Latent Trait Models and Their Use in Inferring Examinee's Ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.

Bock, R. D. 1997. A Brief History of Item Response Theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33.

Choppin, B. H. L. 1985. A Two-Parameter Latent Trait Model. *Evaluation in Education, 9*, 43-62.

Divgi, D. R. 1986. Does the Rasch Model Really Work for Multiple Choice Items? Not if You Look Closely. *Journal of Educational Measurement, 23*(4), 283-298.

Embretson, S., and Reise, S. P. 2000. *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.

Hambleton, R. K. 1994. Item Response Theory: A Broad Psychometric Framework For Measurement Advances. *Psicothema, 6*(3), 535-556.

Hambleton, R. K., & Cook, L. L. 1977. Latent Trait Models and Their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement, 14*(2), 75-96.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. 1991. *Fundamental of Item Response Theory*. Newbury Park, CA: Sage.

Humphry, S. M. 2010. Modeling the Effects of Person Group Factors on Discrimination. *Educational and Psychological Measurement, 70*(2), 215-231.

Humphry, S. M. 2011. The Role of the Unit in Physics and Psychometrics. *Measurement: Interdisciplinary Research & Perspective, 9*(1), 1-24. doi: 10.1080/15366367.2011.558442

Humphry, S. M., & Andrich, D. 2008. Understanding the Unit Implicit in The Rasch model. *Journal of Applied Measurement, 9*(3), 249-264.

Lord, F. M., & Novick, M. R. 1968. *Statistical Theories of Mental Test Scores*. Menlo Park, California: Addison-Wesley.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawlence Erlbaum.

Marais, I., & Andrich, D. 2008. Formalising Dimension and Response Violations of Local Independence in The Unidimensional Rasch Model. *Journal of Applied Measurement, 9*(3), 200-215.

Masters, G. N. 1988. Item Discrimination: When More is Worse. *Journal of Educational Measurement, 25*(1), 15-29.

Rogers, H. J. 1999. Guessing in Multiple Choice Tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 235-243). Amsterdam: Pergamon.

Smith, R. M., & Plackner, C. 2009. The Family Approach to Assessing Fit in The Rasch Measurement. *Journal of Applied Measurement, 10*(4), 423-437.

Tennant, A., & Gonaghan, P. 2007. The Rasch Measurement Model in Rheumatology: What is it and Why Use it? When Should it be Applied, and What Should One Look for in a Rasch Paper? *Arthritis & Rheumatism (Arthritis Care & Research), 57*(8), 1358-1362.

Waller, M. I. 1973. Removing the Effects of Random Guessing from Latent Trait Ability Estimates. Unpublished Ph.D Dissertation. The University of Chicago, Chicago.

Wright, B. D. 1997. A History of Social Science Measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45.

Wright, B. D., & Stone, M. K. 1979. *Best Test Design: Rasch Measurement*. Chicago: Mesa Press.

Yen, W. M. 1980. The Extent, Causes and Importance of Context Effects on Item Parameters for Two Latent Trait Models. *Journal of Educational Measurement, 17*(4), 297-311.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. 2002. Identification and Evaluation of Local Item Dependencies in The Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291-309.