

IDENTIFIKASI *DIFFERENTIAL ITEM FUNCTIONING* PADA TES PENGERTIAN MEKANIK (C4) DENGAN PENDEKATAN RASCH MODEL

Nurul F Prahastuti

UIN Sunan Kalijaga Yogyakarta

nurulfajriyahprahastuti@gmail.com

Abstrak

Tes Pengertian Mekanik (C4), digunakan dalam assesmen untuk mengukur bakat mekanik. Tes Pengertian Mekanik memiliki sifat sebagai *high stakes exam*, membawa konsekuensi besar pada kehidupan subjek yang mengerjakannya. Tes Pengertian Mekanik perlu diuji kembali untuk memastikan bahwa item-itemnya masih dapat dijustifikasi penggunaannya atau perlu diperbaiki. Selain diharapkan masih memiliki kualitas aitem yang baik, penting pula untuk memastikan bahwa Tes Pengertian Mekanik (C4) tidak bias terhadap satu gender. Suatu aitem yang terindikasi DIF akan menguntungkan sekelompok subjek dan merugikan kelompok lainnya. Penelitian ini bertujuan untuk menguji kualitas aitem dan mengidentifikasi Differential Item Functioning (DIF) pada Tes Pengertian Mekanik (Mechanical Reasoning Test) dengan menggunakan Rasch Model. Pendekatan Rasch Model dipilih karena memiliki keunggulan dalam analisis data dibandingkan pendekatan model klasik. Analisis data dilakukan dengan bantuan software Winstep 3.73. Data terkumpul sebanyak 528 dengan jumlah $L = 184$ dan $P = 344$. Responden berada pada rentang usia 16 hingga 20 tahun. Hasil penelitian menunjukkan bahwa Tes Pengertian Mekanik merupakan tes yang sulit, karena tingkat kesulitan tes jauh lebih tinggi daripada kemampuan subjek. Nilai reliabilitas sebesar .73, sehingga dapat dikatakan reliabel. Seluruh soal Tes Pengertian Mekanik fit dan soal menunjukkan kemampuan subjek yang sebenarnya. Terdapat 24 aitem terindikasi DIF yang menguntungkan salah satu kelompok subjek, yakni kelompok subjek laki-laki.

Kata Kunci: Tes Pengertian Mekanik, Rasch Model

Abstract

Mechanical Reasoning Test (C4) was used in an assessment to measure the mechanical reasoning aptitude. Mechanical Reasoning Test has a quality of high-stakes exam, it has a major life consequence for the subject. Mechanical Reasoning Test needs to be re-tested to check whether its items' usage are still justified or if it needs to be changed. Also, besides having good items qualities, it is also important to make sure that Mechanical Reasoning Test (C4) are not biased for one gender. An item indicated to be DIF will be beneficiary for one group of subjects, while biased against another. The purpose of this research was to test the quality of items and to identify the Differential Item Functioning (DIF) of Mechanical Reasoning Test using Rasch Model. Rasch Model was chosen because it has advantages in data analysis compared to classical model. Data analysis was done using the software Winstep 3.73. The amount of data gathered were 528 divided into $M = 184$ and $F = 344$. Respondents' age range was 16 to 20 years old. The analysis showed that Mechanical Reasoning Test was a difficult test, because the difficulty level was much higher than subjects' capability. The reliability score was .73, as such it can be considered reliable. All the questions in Mechanical Reasoning Test were fit and can show subjects' true capability. There were 24 items indicated to be DIF and biased for one group of subjects.

Keywords: Mechanical Reasoning Test, Rasch Model

Pendahuluan

Tujuan pengukuran psikologi adalah melakukan diskriminasi atas suatu atribut psikologi yang akan diukur. Untuk mencapai fungsi diskriminatif tersebut alat ukur seharusnya objektif, dengan makna mampu memberikan perlakuan yang adil pada seluruh subjek yang mengerjakan. Meskipun subjek memiliki latar belakang yang berbeda, satu-satunya hal yang akan diukur adalah abilitas yang melandasi tujuan pengukuran.

Di dalam pengukuran psikologi, dikenal dua pendekatan yang dapat digunakan yaitu *Classical Test Theory (CTT)* dan *Item Response Theory (IRT)*. Pada pendekatan teori tes klasik, parameter aitem bergantung pada kelompok (*group dependant*), sementara pada pendekatan *item response theory* parameter aitem bersifat invariant dimana tidak tergantung pada subjek yang mengerjakan (Azwar, 1994). Terdapat tiga asumsi dasar dalam *Item Response Theory* yaitu *local independence*, unidimensionalitas, dan *invariant*.

Sifat invariant bermakna bahwa perbandingan antar individu tidak dipengaruhi oleh soal yang digunakan, dan perbandingan antar soal tidak terpengaruh oleh subjek yang mengerjakan. Dengan adanya sifat invariant, respon subjek terhadap suatu aitem dapat dipastikan hanya memberikan kemungkinan jawaban

benar saja atau salah saja (Engelhard, 2008). Tidak ada faktor lain yang mencemari respon jawaban subjek pada suatu aitem. Untuk dapat dikatakan sebagai instrumen pengukuran yang berkualitas, instrument tersebut harus dapat mencapai perbandingan yang invariant.

Independensi lokal berarti bahwa apabila abilitas subjek yang menentukan performansi dianggap konstan, maka respon subjek terhadap aitem lainnya akan independent secara statistic (Azwar, 2016; Hambleton, Swaminathan, & Rogers, 1991). Aitem yang satu dengan yang lain tidak berhubungan. Sementara itu, asumsi unidimensional berarti bahwa hanya terdapat satu abilitas atau kemampuan yang diukur melalui seperangkat aitem-aitem dalam sebuah tes (Azwar, 2016; Hambleton et al., 1991). Satu komponen dominan tersebut merupakan faktor yang bersifat permanen, sehingga tujuan ukur sangat jelas dan tidak ada faktor lain yang mencemari. Model Rasch merupakan salah satu bentuk analisis *Item Response Theory* yang menggunakan satu parameter, yaitu tingkat kesukaran aitem.

Suatu hasil analisis dengan menggunakan model Rasch menghasilkan perbandingan data terhadap kecocokan model (fit) dan data dianalisis pada level individual (subjek) (Engelhard, 2008). Dengan kata lain, analisis model Rasch

dilakukan terhadap per aitem dan per individu, serta melakukan uji fit data terhadap model. Selain dapat digunakan untuk melakukan analisis mengenai kualitas aitem, hasil analisis model Rasch juga mengeluarkan output hasil analisis *Differential Item Functioning (DIF)*. Identifikasi DIF dalam pendekatan model Rasch memiliki keunggulan dalam mendeteksi bias butir dibanding metode chi-square Scheuneman dan Mantel-Haenszel (Sudaryono, 2012).

Differential Item Functioning merupakan salah satu analisis yang bertujuan untuk mendeteksi potensi bias pengukuran (Anderson, Reilly, Gorrell, Schaumberg, & Anderson, 2016). DIF juga dapat diartikan sebagai peluang mendapatkan jawaban benar pada kelompok yang memiliki latar belakang berbeda, seperti ras, etnis, gender, atau sub-grup tertentu (Strobl, Kopf, & Zeileis, 2015). Dalam sebuah tes psikologi terdapat berbagai aitem yang memiliki sensitivitas berbeda untuk individu dengan latar belakang yang berbeda. Isu-isu yang populer pada suatu kelompok belum tentu memiliki popularitas yang sama pada kelompok lain. Sebagai contoh, perkembangan teknologi di kota besar lebih cepat berkembang daripada di daerah, sehingga soal yang menyinggung tema teknologi akan cenderung lebih menguntungkan kelompok subjek

perkotaan daripada subjek daerah. Proses identifikasi *Differential Item Functioning* pada dasarnya menggunakan perbandingan estimasi parameter aitem antara dua atau lebih kelompok subjek yang dispesifikasikan (Strobl et al., 2015).

Sebuah butir soal disebut mengandung bias gender apabila secara sistematis memberikan keuntungan pada kelompok laki-laki atau kelompok perempuan (Sudaryono, 2012). Untuk dapat terhindarkan dari terjadinya DIF, proses penyusunan aitem seharusnya menghindari stereotyping. Stereotyping yang ada dalam sebuah aitem akan mengundang masalah yang memancing berubahnya respon dari yang sebenarnya (Hambleton, 2006). Menurut Hambleton (2006), setidaknya ada tiga macam pertanyaan untuk mengidentifikasi apakah terjadi stereotyping. Pertama, apakah konten aitem (makanan tertentu, permainan, religi, aktivitas, penyakit, dan peristiwa sejarah) tidak familiar bagi kelompok lain? Kedua, apakah suatu aitem mengandung kata yang memiliki makna berbeda untuk suatu kelompok? Ketiga, apakah bentuk alternatif respon jawaban yang digunakan dapat memicu jawaban yang tidak sebenarnya? Beberapa kelompok subjek dapat memiliki kecenderungan untuk menghindari respon-respon ekstrim dalam sistem rating.

Analisis dengan menggunakan model Rasch dapat digunakan untuk tes dengan bentuk respon jawaban politomi maupun dikotomi. Dalam penelitian ini, analisis model Rasch digunakan untuk mengkaji kualitas Tes Pengertian Mekanik yang bertujuan untuk mengukur bakat mekanik. Tes Pengertian Mekanik memiliki format jawaban pilihan ganda, yaitu A, B, dan C. Tes Pengertian Mekanik termasuk dalam jenis tes dengan respon jawaban dikotomi.

Bakat mekanik adalah kemampuan memahami dan mengaplikasikan prinsip-prinsip mekanika untuk memprediksi performa dalam bidang yang berhubungan dengan mekanika di masa yang akan datang (Prahastuti, 2018). Untuk mengukur bakat mekanik, diperlukan suatu alat ukur yang secara spesifik mengukur bakat mekanik. Tes Pengertian Mekanik merupakan salah satu tes untuk pengukuran bakat mekanik. Tes tersebut merupakan hasil adaptasi dari *Mechanical Reasoning Test* (Bennett, Seashore, & Wesman, 1952). Tes Pengertian Mekanik telah lama digunakan di Indonesia, baik untuk keperluan asesmen bakat maupun seleksi pekerjaan yang berhubungan dengan bidang mekanika.

Tes Pengertian Mekanik berisi soal-soal yang berkaitan dengan prinsip mekanika. Untuk setiap butir soal, terdapat instruksi berupa kalimat pertanyaan, satu

gambar yang merupakan penerapan prinsip mekanika, dan tiga alternatif pilihan jawaban (Bennett et al., 1952). Penyajian Tes Pengertian Mekanik di Indonesia diberikan dalam batas waktu pengerjaan 30 menit, waktu instruksi 5 hingga 10 menit, serta terdapat booklet soal dan lembar jawaban yang terpisah (Prahastuti, 2018). Format penyajian tersebut disamakan dengan format penyajian versi asli *Mechanical Reasoning Test*. Tes Pengertian Mekanik tersusun atas beberapa tema yang berhubungan dengan prinsip mekanika maupun fisika.

Perbedaan antara laki-laki dan perempuan tampak nyata dalam berbagai ranah, termasuk dalam melakukan penalaran. Kelompok laki-laki akan lebih diuntungkan untuk tes yang bertujuan mengukur aspek spasial dan geometri, sementara kelompok perempuan lebih diuntungkan pada tes dengan tujuan mengukur kemampuan verbal (Sudaryono, 2012). Peneliti berasumsi bahwa dalam tes pengertian mekanik, terdapat potensi kelompok laki-laki lebih diuntungkan karena adanya DIF tersebut.

Kualitas hasil adaptasi dari *Mechanical Reasoning Test* telah diujikan di beberapa negara, diantaranya adalah India, Arab, Belanda, dan Indonesia (Dagenais, 1992; Nijenhuis, Evers, & Mur, 2000; Ojha, 1965; Prahastuti, 2018). Beberapa penelitian yang dilakukan

tersebut masih menggunakan pendekatan teori tes klasik dan masih membutuhkan eksplorasi menggunakan *item response theory*.

Metode Penelitian

Partisipan

Data yang digunakan dalam penelitian ini merupakan data yang diambil dari salah satu biro psikologi. Data terkumpul sebanyak 528, sebanyak 184 responden laki-laki dan 344 responden perempuan. Responden berada pada rentang usia 16 hingga 20 tahun. Responden merupakan siswa kelas XII dan mahasiswa semester awal. Seluruh partisipan merupakan warga negara Indonesia.

Instrumen Pengukuran

Instrumen pengukuran yang digunakan adalah Tes Pengertian Mekanik (C4) yang merupakan hasil translasi dari *Mechanical Reasoning Test*. Tes Pengertian Mekanik merupakan salah satu subtes *Differential Aptitude Test* yang merupakan baterai tes bakat. Tes Pengertian Mekanik memiliki 68 aitem dengan 3 pilihan jawaban yaitu A, B, dan C. Setiap jawaban benar mendapatkan skor 1 dan jawaban salah mendapatkan skor 0. Rentang skor yang dapat diperoleh subjek adalah 0 hingga 68.

Analisis Data

Analisis data dilakukan dengan menggunakan bantuan software Winstep 3.73. Analisis data dilakukan dengan pendekatan Rasch untuk mendapatkan informasi reliabilitas, *item map*, *item fit*, dan *differential item functioning*.

Hasil Penelitian

Hasil analisis person-item-map menunjukkan bahwa Tes Pengertian Mekanik merupakan tes dengan aitem-aitem yang cenderung sulit. Kemampuan siswa jauh berada di bawah tingkat kesukaran soal. Pada gambar 1 ditunjukkan bahwa nilai person-map di bawah item-map.

Nilai model RMSE sebesar 0.28 , dimana nilai tersebut lebih dari 0.05 . Dapat diartikan bahwa model penganalisaan fit atau sesuai. Nilai Indeks separation sebesar 1.64, dimana nilai ini lebih dari 1, maka dapat dikatakan bahwa kelompok subjek yang melaksanakan tes memiliki kemampuan penalaran mekanik yang heterogen (berkemampuan tinggi, sedang, rendah). Hasil analisis reliabilitas menunjukkan koefisien sebesar .73 , sehingga dapat disimpulkan bahwa Tes Pengertian Mekanik reliabel. Seluruh aitem fit karena tidak ditemukan nilai outfit MNSQ > 2

Identifikasi DIF dilihat berdasarkan hasil perolehan nilai probability > .05.

Berdasarkan hasil analisis yang telah dilakukan, diperoleh hasil 24 aitem terindikasi adanya DIF. Aitem-aitem yang teridentifikasi DIF yaitu aitem nomor 8, 17, 19, 20, 24, 30, 31, 35, 40, 44, 48, 49, 50, 51, 52, 53, 54, 57, 60, 62, 63, 64, 65, dan 68.

Diskusi

Penelitian ini bertujuan untuk mengidentifikasi kualitas butir Tes Pengertian Mekanik dan analisis DIF menggunakan Rasch Model. Hasil analisis menunjukkan bahwa Tes Pengertian Mekanik merupakan tes yang cenderung sulit, sejalan dengan hasil penelitian sebelumnya (Prahastuti, 2018). Tes Pengertian Mekanik merupakan tes yang aitem-aitemnya terdiri dari gambar dan kalimat pernyataan. Gambar-gambar yang ada dalam soal merepresentasikan suatu prinsip mekanika tertentu, namun banyaknya objek yang tertera di dalam gambar merupakan objek yang sudah lama dan jarang dipakai pada masa kini. Terdapat indikasi bahwa soal menjadi cenderung sulit karena objek sulit untuk dikenali, seperti bentuk mainan dan perkakas yang hanya ada pada masa lampau. Permasalahan subjek yang kesulitan mengenali objek dalam soal juga disebutkan dalam penelitian Ojha (1965) mengenai subtes Mechanical Reasoning Test untuk kelompok subjek India.

Nilai reliabilitas yang diperoleh dengan pendekatan model Rasch sebesar .73. Koefisien reliabilitas dengan pendekatan Rasch ini sedikit lebih tinggi daripada koefisien reliabilitas yang diperoleh dengan pendekatan klasik, yaitu .707 (Prahastuti, 2018). Semakin reliabel sebuah tes, maka semakin kecil error yang dihasilkan. Tes Pengertian Mekanik memiliki reliabilitas yang baik. Aitem-aitem dalam Tes Pengertian Mekanik memperoleh nilai Infit MNSQ > 2, sehingga dikatakan fit dengan model pengukuran. Dalam model Rasch yang menggunakan 1 parameter yaitu tingkat kesukaran, aitem-aitem Tes Pengertian Mekanik menjalankan fungsi ukurnya dengan baik.

Hasil penelitian menunjukkan 24 aitem terindikasi DIF. DIF ditemukan tidak hanya pada soal yang memiliki tingkat kesulitan tinggi, apabila dibandingkan dengan hasil penelitian Prahastuti (2018). Terdapat 3 aitem yang memiliki tingkat kesulitan rendah terindikasi DIF. Sejumlah 11 aitem yang memiliki tingkat kesukaran rendah ditemukan terindikasi DIF. Sementara itu, pada tingkat kesulitan tinggi terdapat 10 aitem terindikasi DIF. DIF terjadi pada soal yang memiliki tingkat kesulitan rendah, sedang, dan tinggi. Tingkat kesulitan aitem dalam penelitian Prahastuti (2018) menggunakan pendekatan teori tes klasik. Tingkat

kesulitan ditentukan dari perbandingan antara jumlah subjek yang menjawab dengan benar suatu aitem dibagi dengan banyaknya subjek yang mengerjakan aitem tersebut.

Kesimpulan

Berdasarkan analisis yang telah dilakukan, peneliti menyimpulkan bahwa Tes Pengertian Mekanik merupakan tes untuk mengukur bakat mekanik yang telah terstandar. Telah dilakukan adaptasi bahasa dan budaya oleh negara-negara yang menggunakan tes ini. Tes Pengertian Mekanik merupakan salah satu produk tes psikologi yang berusia tua namun masih relevan digunakan hingga saat ini. Tes Pengertian Mekanik memiliki reliabilitas yang cukup baik, menjalankan fungsi pengukurannya dengan baik, namun terdapat indikasi bahwa adanya bias gender yang diketahui melalui analisis *Differential Item Functioning*.

Daftar Pustaka

- Anderson, L. M., Reilly, E. E., Gorrell, S., Schaumberg, K., & Anderson, D. A. (2016). Gender-based differential item function for the difficulties in emotion regulation scale. *Personality and Individual Differences*, 92, 87–91.
- Azwar, S. (1994). Differential Item Functioning analysis, an issue on UMPTN item banking with IRT procedures. *Buletin Psikologi*, 1, 14–20.
- Azwar, S. (2016). *Dasar-Dasar Psikometrika*. Yogyakarta: Pustaka Pelajar.
- Bennett, G. K., Seashore, H., & Wesman, A. (1952). *Differential Aptitude Test Manual* (Second Edition). New York: The Psychological Corporation.
- Dagenais, F. (1992). Bennett Mechanical Comprehension Test: Normative data for a sample of Saudi Arabian technical trainees. *Perceptual and Motor Skills*, 74(1), 107–113.
- Engelhard, G. (2008). Historical Perspectives on Invariant Measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, 6(3), 155–189. <https://doi.org/10.1080/15366360.802197792>
- Goh, H. E., Marais, I., & Ireland, M. J. (2017). A Rasch Model Analysis of the Mindful Attention Awareness Scale. *Assessment*, 24(3), 387–398.
- <https://doi.org/10.1016/j.paid.2015.12.016>

- <https://doi.org/10.1177/1073191115607043>
- Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning: *Medical Care*, 44(Suppl 3), S182–S188. <https://doi.org/10.1097/01.mlr.0000245443.86671.c4>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.
- Nijenhuis, J. T., Evers, A., & Mur, J. P. (2000). Validity of the Differential Aptitude Test for the Assessment of Immigrant Children. *Educational Psychology*, 20(1), 99–115.
- Ojha, J. M. (1965). *Revision of Differential Aptitude Test for Higher Secondary Schools*. The Maharaja Sayajirao University of Baroda, Vadodara.
- Prahastuti, N. F. (2018). *Analisis Kualitas Tes Pengertian Mekanik (C4) berdasarkan Teori Tes Klasik*. Universitas Gadjah Mada, Yogyakarta.
- Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, 80(2), 289–316.
- <https://doi.org/10.1007/s11336-013-9388-3>
- Sudaryono. (2012). Sensitivitas metode pendeteksian differential item functioning (DIF). *Jurnal Evaluasi Pendidikan*, 3(1), 82–94.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Kimbark Ave: Mesa Press.