DIFFERENTIAL ITEM FUNCTIONING
(KEBERBEDAAN FUNGSI BUTIR)

Oleh: Muh. Ilyas Ismail*

ABSTRACT: *This paper theoretically deals some aspects related to Differential Item Functioning (DIF), consisting of Test and Test Bias. The purpose of this is to describe the importance of understanding Different Item Functioning or to discuss its existence. Based on the theoretical connotation of DIF, it can be inferred that DIF may be used to make sure whether or not a bias item of a test exists. For example, a test item shows the existence of DIF when the tested from different group having the same ability do not have the same opportunity to answer the item correctly. Therefore, understanding the DIF, a teacher may avoid item bias on a test to exist.*

KEYWORDS: *Differential Item Functioning, test, test bias.*

DALAM kaitan ini, persoalan yang akan disoroti dan dikaji adalah dari aspek penggunaan tes. Hal ini berhubungan dengan tingkat kevalidan atau kesahihan tes yakni sejauh mana tes tersebut benar-benar mengukur aspek yang diukur artinya tidak terjadi bias tes. Aiken mendefinisikan validitas sebagai berikut "*Validity of a test has been defined as the extent to wich the test measures what it was designed to measures.* [1]

Pada prinsipnya, pengukuran bertujuan untuk mengetahui karakteristik suatu objek yang akan diukur. Bias tes telah menjadi tema utama dalam pengukuran pendidikan sejak tahun 1960-an. Kajian bias tes dihubungkan dengan keadilan dan persamaan hak bagi kelompok-kelompok dalam msyarakat, Perdebatan secara luas tentang bias tes telah mendorong usaha-usaha untuk memecahkan masalah tersebut.

Selain hal-hal tersebut di atas, butir tes yang baik harus terbebas dari bias. Tes yang baik tidak memihak pada kelompok tertentu atau golongan tertentu dari peserta tes. Tes yang baik akan memberikan hasil pengukuran yang sama terhadap peserta tes yang memiliki kemampuan sama meskipun berasal dari kelompok atau golongan yang berbeda. Bila tes memberi-

*Kandidat Doktor pada Program Pascasarjana Universitas Negeri Jakarta, Prodi PEP, ini adalah Dosen Tetap pada Fakultas Tarbiyah dan Keguruan UIN Alauddin Makassar.

kan hasil yang berbeda maka tes tersebut bias, yang berarti perangkat tersebut tidak valid secara konstruktif. Sebuah tes yang validitasnya rendah berarti tes tersebut tidak mampu secara akurat mengukur apa yang seharusnya diukur.

Cara awal untuk mengetahui ada tidaknya bias item pada suatu item tes adalah dengan melakukan analisis *Differential Item Functioning (DIF)* atau dikenal dengan keberbedaan fungsi butir. Analisis ini dilakukan untuk mengidentifikasikan butir tes yang memiliki perbedaan fungsi untuk kelompok siswa yang berbeda, misalnya kelompok siswa laki-laki dan kelompok siswa perempuan. Sebuah butir tes menunjukkan *Differential Item Functioning (DIF)* atau dikenal dengan keberbedaan fungsi butir jika tes yang mempunyai kemampuan yang sama, tetapi berasal dari kelompok yang berbeda, tidak mempunyai peluang yang sama untuk menjawab benar.

Oleh karena itu, maka dalam tulisan ini penulis akan menelaah secara teoretis tentang beberapa aspek yang berhubungan dengan keberbedaan fungsi butir (DIF), yang mencakup tentang Tes dan Bias Tes, Pengertian keberbedaan fungsi butir (DIF), dan Kesimpulan.

PENGERTIAN TES

Manakala mendiskusikan istilah tes, maka ia tidak lepas dari sejarahnya. Salah satu yang membahas itu yang dikutip oleh Allen dan Yen[2] adalah yang ditulis oleh Philip H. Du Bois pada 1970. Du Bois mengkategorisasikan penggunaan tes ke dalam tiga cakupan, yakni tes pegawai, tes sekolah, dan tes individu. Tes pegawai mulai digunakan di Cina pada dinasti Chan pada 1115 SM ketika pemerintah menguji kompetensi para pegawainya.[3] Tes itu selanjutnya digantikan oleh bukti akademik formal pada 1905 yang pada masa itu di Inggris dan Amerika baru dimulainya penggunaan tes untuk menyeleksi pegawai pemerintah.

Pada abad ke-12, tes dikembangkan di sekolah-sekolah di Eropa, khususnya tes lisan. Selanjutnya, mengikuti bentuk tes lisan, bentuk tes tulisan berkembang hingga kini. Tes individu mulai dikembangkan di Inggris oleh S.F. Galton. Alat ukur tes untuk menyekor keahlian motorik dan sensori dikembangkan pada masa itu.

Istilah tes menurut Cronbach dimaknai sebagai suatu prosedur yang sistematik untuk membandingkan perilaku di antara dua peserta tes atau lebih.[4] Hal itu didukung oleh pendapat Popham ketika mendefinisikan tes pada acuan norms, yakni untuk menentukan kedudukan individu peserta tes dari para peserta tes lainnya berkenaan dengan performansinya pada

tes yang diberikan.[5] Nitko mendefinisikan tes sebagai sebuah instrumen atau prosedur sistematik untuk mengamati dan mendeskripsikan karakteristik peserta tes dengan menggunakan skala numerik atau skema klasifikasi.[6] Tes juga dipahami sebagai istilah yang menunjuk pada sajian seperangkat pertanyaan yang direspons oleh peserta tes. Respon itu melahirkan hasil ukur yang berupa skor sebagai representasi karakteristik peserta tes,[7] dan berupa kategori yang merepresentasikan perilaku yang diperoleh melalui penyampelan.[8] Karakteristik yang dimaksud ini berupa abilitas. Secara lebih luas, tes juga dipahami sebagai sebuah istilah yang menunjuk pada segala sesuatu, mulai dari penyelenggaraan hingga interpretasi skor.[9] Menurut Lord, istilah 'tes' yang melekati kata 'pendidikan' atau 'psikologi' yang membentuk frasa baru 'tes pendidikan' atau 'tes psikologi' dipahami sebagai alat untuk memeroleh sampel perilaku yang berupa abilitas.[10] Abilitas ini merupakan salah satu variabel yang berkenaan dengan psikologi, di samping intelegensi, kepribadian, sikap, minat, motivasi, dan nilai.[11]

Hal senada adalah yang dikemukakan oleh Crocker dan Algina yang mendefinisikan tes sebagai prosedur baku untuk memeroleh sampel perilaku dari ranah yang telah ditentukan.[12] Pendapat itu selaras dengan gagasan Popham yang mendefinisikan tes pada acuan kriterion, yakni untuk menentukan kedudukan peserta tes berkenaan dengan ranah perilaku atau tingkat profisiensi yang ditentukan.[13]

Dengan demikian, tes dipahami sebagai sebuah alat ukur atau prosedur sistematik yang terdiri atas sejumlah pertanyaan atau pernyataan sebagai butir-butir tes dengan aturan pengskorannya yang memerlukan respons peserta tes yang digunakan sebagai sampel perilaku atau karakteristik peserta tes melalui deskripsi dan perbandingan di antara para peserta tes dengan menggunakan Skala numerik atau skema klasifikasi berdasarkan atas ranah yang ditentukan sebelumnya dan yang selanjutnya menghasilkan hasil ukur yang berupa sekor atau kategori.

PENGERTIAN BIAS TES

Ada beberapa sebab yang dapat dijadikan sumber ketimpangan skor, diantara sumber ketimpangan skor tersebut adalah terletak pada peserta uji tes dan perangkat uji tes beserta butir-butirnya.[14] Bila ketimpangan skor tersebut terjadi pada kelompok peserta tes maka hal tersebut dinamakan sebagai ketidakwajaran, sebaliknya bila ketimpangan terjadi pada butir uji tes maka hal itu disebut sebagai bias soal.

Ketimpangan skor karena penggunaan soal yang bias atau ketimpangan skor karena, pengskoran yang tidak dapat mengkompensasi

pengguna soal yang bias, kesemuanya bertolak dari adanya soal yang bias itu. Jadi, kita berpegang saja kepada bias soal atau soal yang bias itu, dan yang terpenting sekarang adalah apa yang dimaksud dengan bias soal itu?

Ada berbagai definisi atau pengertian bias dalam teori pengukuran dan penilaian. Setiap ahli memiliki pengertian yang berbeda-beda. Menurut Zumbo bias soal adalah ketika peserta tes dari dua kelompok yang mempunyai kemampuan sama, tetapi salah satu kelompok dapat menjawab benar yang lebih sedikit dari kelompok lainnya.[15] Ironson dalam Berk mengatakan bahwa suatu soal tidak termasuk bias bila individu yang mengikuti tes dari kelompok berbeda yang memiliki kemampuan sama, mempunyai kemungkinan yang sama untuk menjawab soal dengan benar.[16]

Naga memberi pengertian terhadap bias soal sebagai ketimpangan skor yang disebabkan oleh butir uji tes, sementara peserta dan kelompok peserta adalah wajar. Menurutnya ketimpangan yang dinamakan bias soal terdapat pada peserta secara berkelompok atau sub kelompok, sedangkan ketimpangan skor yang terdapat pada peserta secara individu dinamakan ketidakwajaran skor.[17]

Menurut Hambleton, pengertian dari soal yang bias adalah adanya perbedaan skor perolehan yang disebabkan oleh adanya unsur yang menguntungkan atau merugikan bagi peserta.[18]

Menurut Popham, mengatakan bahwa bila suatu proses pengukuran yang menggunakan soal menghasilkan dampak yang berbeda pada kelompok peserta tes yang mempunyai kemampuan sama, maka hal itu disebut bias.[19] Sedangkan Nitko menjelaskan bahwa soal yang bias bila rata-rata perolehan skor tes satu kelompok lebih rendah dengan kelompok lainnya.[20]

Adapun faktor-faktor yang menyebabkan terjadinya bias soal dalam pelaksanaan tes adalah pengaruh perbedaan ras, jenis kelamin, wilayah, budaya dan etnis[21] dan pada tahun 1972, Anggof dalam Holland,[22] memperkenalkan suatu metode untuk mempelajari adanya bias soal yang disebabkan oleh perbedaan budaya, begitu juga halnya dengan Cardall dan Coffman yang meneliti adanya bias soal yang disebabkan oleh perbedaan ras.

Dari uraian di atas dapat disimpulkan bahwa bias soal adalah kemungkinan perbedaan perolehan skor tes dua kelompok yang mempunyai kemampuan sama. Perbedaan skor tes tersebut disebabkan oleh faktor-faktor perbedaan ras, jenis kelamin, latar belakang budaya, dan perbedaan etnis.

PENGERTIAN KEBERBEDAAN FUNGSI BUTIR (DIF)

Istilah Bias butir dan istilah keberbedaan fungsi butir atau *DIF (Differential Item Functioning)* sering digunakan oleh para pakar pengukuran untuk merujuk pada konsep yang sama. Istilah "bias item" maknanya lebih luas daripada istilah *Differential Item Functioning* (*DIF)* yang semata-mata merupakan hasil temuan dari olah statistik, sementara bias telah melibatkan analisis lanjutan secara kualitatif dari hasil temuan olah statistik tadi.

Suatu item dikatakan bias apabila dua kelompok yang memiliki kemampuan sama memperoleh hasil yang berbeda pada item tersebut. Secara matematis bias item dapat dinyatakan dalam bentuk probabilitas. Artinya, orang yang mempunyai kemampuan sama tetapi tidak memiliki peluang sama untuk memperoleh jawaban benar. Kemudian Angoff, lebih lanjut menjelaskan "*An item is biassed if equal able (or proficient) individuals, from difference groups, do not have equal probabilities of answering the item correctly*.[23]

Apabila suatu item relatif lebih sulit untuk kelompok yang memiliki budaya dan latar belakang pengalaman tertentu berarti butir tersebut bias. Bias butir dalam suatu pengukuran mengindikasikan adanya kesalahan sistematik dalam pengukuran tersebut.

Differential Item Functioning (*DIF)* atau tidak, diperlukan indeks *Differential* Bias item memiliki dua karakter, yaitu (1) arah *(direction)* dan besaran *(magnitude).* Besaran bias dapat diestimasi secara statistik. Ada berbagai metode untuk mengestimasi besaran tersebut, antara lain: metode pendekatan klasik, metode *chi-square,* pendekatan IRT, analisis faktor *konfirmatori* (AFK), dan Model persamaan struktutal (MPS).

Bias butir dapat terjadi sebanyak jenis pengelompokan yang diinginkan, oleh penulis (peneliti). Namun pengelompokkan yang sering dilakukan oleh para peneliti adalah bias karena budaya, dan gender. Sebuah butir disebut bias budaya apabila perbedaan kelompok yang akan diteliti atau diperbandingkan ditetapkan berdasarkan aspek budaya (etnis, ras, dan bahasa yang digunakan).

Selanjutnya, ada dua faktor yang mempengaruhi timbulnya bias butir. Secara garis besar bias butir disebabkan oleh (1) butir itu sendiri yang dalam kajian ini disebut sebagai faktor internal, dan (2) faktor di luar butir yang dalam kajian ini disebut faktor eksternal. Ketika kajian bias butir difokuskan pada faktor internal berarti fokus deteksi bias butir adalah karakteristik butir. Apabila kajian bias butir difokuskan pada faktor eksternal maka fokus deteksi bias butir adalah penempuh tes.

Bias butir karena faktor internal terjadi apabila kajian difokuskan pada komponen butir, misalnya, bentuk butir, materi butir, kalimat dan kata yang digunakan, gambar, petunjuk *(clue),* dan obyek atau stimulus yang digunakan dalam butir. Dalam penulisan butir tes ada dua bentuk butir yang lazim digunakan, yaitu bentuk pilihan ganda dan bentuk uraian. Beberapa peneliti menemukan butir dalam bentuk uraian lebih adil gender ketika digunakan untuk mengukur prestasi IPA.

Secara konseptual, *Differential Item Functioning* (*DIF*) atau keberbedaan fungsi butir dikatakan muncul pada sebuah butir soal, jika peserta tes yang mempunyai kemampuan yang sama pada konstruks yang diukur oleh tes, tetapi berasal dari kelompok berbeda, mempunyai peluang berbeda dalam menjawab benar item soal tersebut.[24] Untuk menentukan apakah suatu butir soal terindikasi *Item Functioning* (*DIF),* yaitu indeks yang menunjukkan seberapa kuat indikasi *Differential Item Functioning* (*DIF)* ada pada butir itu. Jika tingkat indikasi *Differential Item Functioning (DIF)* tersebut secara praktik dianggap signifikan, dapat dengan mengujinya memakai uji statistik tertentu atau hanya dengan melihat indeksnya saja, maka Butir soal yang bersangkutan dikatakan terdeteksi sebagai butir *Differential Item Functioning (DIF).*

Dalam konteks *item response theory,* terjadi atau tidak terjadinya *Differential Item Functioning* (*DIF)* pada sebuah butir soal terletak kepada fungsi respons butir *(item response function)* untuk butir soal tersebut pada kelompok yang dipersoalkan. Kurva yang menggambarkan fungsi respons butir disebut kurva respons butir atau kurva karakteristik butir atau *item characteristic curve (ICC).*

Terdapat dua jenis *Differential Item Functioning* (*DIF),* yaitu *Differential Item Functioning* (*DIF)* uniform (konsisten) dan *Differential Item Functioning* (*DIF)* tidak uniform (tidak konsisten). *Differential Item Functioning* (*DIF)* uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan *Differential Item Functioning* (*DIF)* tidak uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan. Jika dikaitkan dengan pengertian interaksi, yang popular pada uji statistik analisis variansi, *Differential Item Functioning* (*DIF)* uniform terjadi jika tidak terdapat interaksi antara tingkat kemampuan peserta tes dan keanggotaan kelompok dan, *Differential Item Functioning* (*DIF)* tidak uniform terjadi jika terdapat interaksi antara tingkat kemampuan peserta tes dan keanggotaan kelompok.

Differential Item Functioning (*DIF)* uniform terjadi jika kurva karakteristik butir untuk suatu item soal berbeda untuk kelompok yang berbeda

dan kedua kurva tersebut tidak saling berpotongan. Sebaliknya, *Differential Item Functioning* (*DIF)* tidak uniform terjadi jika kurva karakteristik item untuk suatu item soal berbeda untuk kelompok yang berbeda, namun kedua kurva tersebut berpotongan.

SIMPULAN

Berdasar pada paparan di atas, maka penulis menarik beberapa kesimpulan sebagai berikut:

1. Tes dipahami sebagai sebuah alat ukur atau prosedur sistematik yang terdiri atas sejumlah pertanyaan atau pernyataan sebagai butir-butir tes dengan aturan pengskorannya yang memerlukan respon peserta tes yang digunakan sebagai sampel perilaku atau karakteristik peserta tes melalui deskripsi dan perbandingan di antara para peserta tes dengan menggunakan Skala numerik atau skema klasifikasi berdasarkan atas ranah yang ditentukan sebelumnya dan yang selanjutnya menghasilkan hasil ukur yang berupa skor atau kategori.
2. Bias soal adalah kemungkinan perbedaan perolehan skor tes dua kelompok yang mempunyai kemampuan sama. Perbedaan skor tes tersebut disebabkan oleh faktor-faktor perbedaan ras, jenis kelamin, latar belakang budaya, dan perbedaan etnis.
3. D*ifferential item functioning* atau keberbedaan fungsi butir adalah probabilitas menjawab benar butir soal dari dua kelompok yang berbeda tetapi mempunyai kemampuan sama, atau *Differential Item Functioning* (*DIF*) dikatakan muncul pada sebuah butir soal, jika peserta tes yang mempunyai kemampuan yang sama pada konstruks yang diukur oleh tes, tetapi berasal dari kelompok berbeda, mempunyai peluang berbeda dalam menjawab benar item soal tersebut.

CATATAN AKHIR

1. Lewis R. Aiken, *Psychological Testing and Assessment*, Boston: Allyn and Bacon, Inc, 1988, h. 103.
2. Mary J. Allen, and Wendy M. Yen, *Introduction to Measurement Theory*, Monterey, California: Cole Publishing company, 1979, h. 2–3.
3. Ronald J. Cohen and Mark E. Swerdlik, *Psychological Testing and* Assessment: *An Introduction to Tests and Measurement,* California: Mayfield Publishing Company, 1999, h. 44.
4. Lee J. Cronbach, *Essentials of Psychological Testing,* New York: Harper & Brothers, 1949, h. 11.
5. W. James Popham, *Classroom Assessment, What Teacher Need to Know* Boston: Allyn and Bacon, 1994, h. 26.
6. Anthony Nitko, *Educational Test and Measurement an Introduction*, New York: Harcourt Brace Javanovich, Inc, 1983, h. 6.

7. William A. Mehrens and Irvin J. Lehmann. *Using Standardized Tests in Education*, Edisi ke-4. White Plains, New York: Longman Inc., 1987, h. 7.
8. Robert J. Gregory, *Psychological Testing: History, Principles, ang Applications*, Edisi ke-3 Needham Heights, Massachusetts: Allyn & Bacon Inc., 2000, h. 30.
9. Ronal J. Cohen, and Mark E. Swerdlik, *op.cit*, h. 2.
10. Frederic M. Lord, *Applications of Item Response Theory to Practical Testing Problems*, New Jersey: Lawrence Erlbaum Associates, 1980, h. 3.
11. Ronald J. Cohen and Mark E. Swerdlik, *op.cit*, h. 7.
12. Linda Crocker and James Algina, *Introduction to Classical and Madenr Test Theory,* Florida: Holt, Rinehart and Winston, 1986, h. 4.
13. W. James Popham, *op.cit.*, h. 27.
14. Dali S. Naga, *Pengantar Teori Skor,* Jakarta: Guna Darma, 1992, h. 436.
15. Bruno D. Zumbo, *A Handbook on the Theory and Methods of Differential Item Functioning,* Canada: Directorate of Human Resources, h. 12.
16. Ronald A. Berk, *Handbook; of Method For detecting Test Bias,* USA:The John Hopkins University Press, 1982, h. 117.
17. Dali S. Naga, *op.cit.,* h. 440.
18. Ronald K. Hambleton, Swaminathan, Jane Rogers, *Fundamentals of Item Response Theory,* California: Sage Publications, Inc. 1991, h. 282.
19. W. James Popham, *op.cit.,* h. 63.
20. Anthony J. Nitko, *op.cit.,* h. 43.
21. Ronald A. Berk, *op.cit.,* h. 1.
22. Paul W. Holland & Howard Wainer (ed), *Differential Item Fungtioning* New Jersey: Lawrence Erlbaum Associates Publisher, 1993, h. 5
23. *Ibid.*,
24. Charles L. Hulin, Pritsz Drasgow and Charles K. Parsons. *Rein Response Theory, Applications to Psychological Measurement,* Illinois: Dow Jones-Irwin Homewood, 1983.

DAFTAR PUSTAKA

Aiken, Lewis R., Psychological Testing and Assessment. Boston: Allyn and Bacon, Inc, 1988.

Allen, Mary J. dan Wendy M. Yen, *Introduction to Measurement Theory*. Monterey, California:Cole Publishing company, 1979.

Berk, Ronald A., *Handboo; of Method For detecting Test Bias,* USA:The John Hopkins University Press, 1982.

Cohen, Ronald J. dan Mark E. Swerdlik, *Psychological Testing and* Assessment: *An Introduction to Tests and Measurement,* California: Mayfield Publishing Company, 1999.

Crocker, Linda dan James Algina. *Introduction to Classical and Madenr Test Theory,* Florida: Holt, Rinehart and Winston, 1986.

Cronbach, Lee J., *Essentials of Psychological Testing,* New York: Harper & Brothers, 1949.

Gregory, Robert J., *Psychological Testing: History, Principles, ang Applications*, Edisi ke-3 Needham Heights, Massachusetts: Allyn & Bacon Inc., 2000.

Hambleton, Ronald K. & Hariharan Swaminathan, *Item Response Theory, Principles and Applications,* Boston: Kluwer Academic Publisher, 1984.

Hambleton, Ronald K. & Jane Rogers Swaminathan, *Fundamentals of Item Response Theory,* California: Sage Publications, Inc. 1991.

Holland, Paul W. & Howard Wainer (ed), *Differential Item Fungtioning,* New Jersey: Lawrence Erlbaum Associates Publisher, 1993.

Hulin, Charles L. Pritsz Drasgow and Charles K. Parsons, *Rein Response Theory, Applications to Psychological Measurement,* Illinois: Dow Jones-Irwin Homewood, 1983.

Lord, Frederic M., *Applications of Item Response Theory to Practical Testing Problems*, New Jersey: Lawrence Erlbaum Associates, 1980.

Mehrens, William A. dan Irvin J. Lehmann, *Using Standardized Tests in Education*, Edisi ke-4. White Plains, New York: LongmanInc., 1987.

Naga, Dali S., *Pengantar Teori Skor,* Jakarta: guna Darma, 1992.

Nitko, Anthony J., *Educational Test and Measurement an Introduction*, New York: Harcourt Brace Javanovich, Inc, 1983.

Popham, W. James, *Classroom Assessment, What Teacher Need to Know* Boston: Allyn and Bacon, 1994.

Zumbo, Bruno D., *A Handbook on the Theory and Methods of Differential Item Functioning*, Canada: Directorate of Human Resources.