KAJIAN METODE DETEKSI *DIFFERENTIAL ITEM FUNCTION* (*DIF*) BUTIR SOAL UJIAN NASIONAL DENGAN TEORI TES KLASIK*)

THE VARIOUS METHODS OF DETECTING THE EXISTENCE OF DIFFERENTIAL ITEM FUNCTION (DIF) ITEM NATIONAL EXAM WITH CLASSICAL TEST TEORY

Sudaryono
STMIK Raharja Tangerang, Jl. Jend. Sudirman No. 40 Cikokol -Tangerang
Email: sudaryono2@yahoo.com

Abstract: *The general objective of this study is intended to explain the various methods of detecting the existence of Differential Item Function (DIF) in items of national exam with classical test theory. While the specific purpose of writing the article is intended to explain: 1) the various methods that can be used to detect the presence of DIF in items of national exam based on classical test theory, and 2) the advantages and disadvantages of each method used and find out which method is most sensitive in detecting the presence of DIF items the national exam. Problems of this study are: 1) what methods can be used to detect the presence of DIF in items based on the national exam classical test theory? 2) which method is most sensitive in detecting the presence of DIF in items such national exam based on classical test theory. The methodology used is to review literature from books, journals and study the results of research that has been done. There are many ways to detect item bias and bias test the scores achieved by the theory of classical scores, namely: single group validity, differential validity, item discrimination procedure, the delta plot method, methods of standardization, Scheuneman chi-squared approach, Camilli chi-square approach, Mantel-Haenszel method, a standard procedure which has been developed by Dorans and Kulick, and item bias estimation method with Confirmatory factor Analysis.*

Keywords: *differential item function, item national exam, classical test theory, the delta plot, and estimate the DIF*

Abstrak: *Tujuan umum kajian ini dimaksudkan untuk menjelaskan berbagai metode pendeteksian keberadaan Differential Item Function (DIF) pada butir-butir soal ujian nasional dengan teori tes klasik. Tujuan khusus penulisan ini dimaksudkan untuk menjelaskan: 1) berbagai metode yang dapat digunakan untuk mendeteksi keberadaan DIF pada butir soal ujian nasional berdasarkan teori tes klasik (classical test theory); dan 2) kelebihan dan kekurangan masing-masing metode yang digunakan dan mengetahui metode mana yang paling sensitif dalam mendeteksi keberadaan DIF butir soal ujian nasional. Permasalahan kajian ini adalah: 1) metode apa saja yang dapat digunakan untuk mendeteksi keberadaan DIF pada butir soal ujian nasional berdasarkan teori tes klasik?; 2) metode mana yang paling sensitif dalam mendeteksi keberadaan DIF pada butir soal ujian nasional tersebut berdasarkan teori tes klasik. Metodologi yang digunakan adalah melakukan kajian pustaka dari buku-buku, jurnal-jurnal dan telaah hasil-hasil penelitian yang telah dilakukan. Ada banyak cara untuk mendeteksi butir bias dan uji tes bias pada skor yang dicapai melalui teori skor klasik, yaitu: korelasi kelompok tunggal, korelasi diferensial, prosedur diskriminasi butir, metode plot delta, metode Standarisasi, metode Chi-square Scheuneman, metode Chi-square Camilli, metode Mantel-Haenszel, prosedur standar yang telah dikembangkan oleh Dorans dan Kulick, dan metode estimasi bias butir dengan Analisis Faktor Konfirmatori.*

Kata kunci: *differential item function, butir soal ujian, teori tes klasik, plot delta, dan estimasi bias butir*

Pendahuluan

Kegiatan menganalisis butir soal merupakan suatu kegiatan yang harus dilakukan guru untuk meningkatkan mutu soal yang telah ditulis. Kegiatan ini merupakan proses pengumpulan, peringkasan, dan penggunaan informasi dari jawaban siswa untuk membuat keputusan tentang setiap penilaian (Nitko, 1996). Tujuan penelaahan soal adalah untuk mengkaji

*)Diterima tanggal 29 Pebruari 2012 - dikembalikan tanggal 1 Mei 2012 - disetujui tanggal 1 Juni 2012

dan menelaah setiap butir soal agar diperoleh soal yang bermutu sebelum soal digunakan. Tujuan analisis butir soal juga untuk membantu meningkatkan tes melalui revisi soal yang tidak efektif, serta untuk mengetahui informasi diagnostik pada siswa (Hun Li & Stout, 1996).

Soal yang bermutu adalah soal yang dapat memberikan informasi setepat-tepatnya sesuai dengan tujuan, di antaranya dapat menentukan peserta didik mana yang sudah atau belum menguasai materi yang diajarkan. Dalam melaksanakan analisis, soal dapat dianalisis secara kualitatif, dalam kaitan dengan isi dan bentuknya; dan kuantitatif dalam kaitan dengan ciri-ciri statistiknya atau prosedur peningkatan secara *judgement* dan prosedur peningkatan secara empirik.

Analisis kualitatif mencakup pertimbangan validitas isi dan konstruk, analisis kuantitatif mencakup pengukuran kesulitan butir soal dan diskriminasi soal, termasuk validitas soal dan reliabilitasnya. Tujuan utama analisis butir soal dalam sebuah tes yang dibuat guru atau dinas pendidikan, yaitu untuk mengidentifikasi kekurangan-kekurangan dalam tes atau dalam pembelajaran. Selain itu, hasil analisis butir soal dapat digunakan untuk menelaah dan menganalisis berbagai aspek yang berhubungan dengan umpan balik terhadap kesulitan belajar siswa. Berdasarkan tujuan tersebut, maka kegiatan analisis butir soal memiliki banyak manfaat, yaitu: 1) dapat membantu para pengguna tes dalam evaluasi atas tes yang digunakan; 2) sangat relevan bagi penyusunan tes secara nasional dan lokal seperti tes yang disiapkan guru untuk siswa di kelas; 3) mendukung penulisan butir soal yang efektif; 4) secara materi dapat memperbaiki tes di kelas; dan 5) meningkatkan validitas soal dan reliabilitas soal (Anastasi & Urbina, 1997).

Di samping itu, manfaat lainnya adalah: 1) menentukan apakah suatu fungsi butir soal sesuai dengan yang diharapkan; 2) memberi masukan pada siswa tentang kemampuan dan sebagai dasar untuk bahan diskusi di kelas; 3) memberikan masukan pada guru tentang kesulitan siswa; 4) memberikan masukan pada aspek tertentu untuk mengembangkan kurikulum; 5) merevisi materi yang dinilai atau diukur. Berbagai uraian di atas menunjukkan bahwa analisis butir soal adalah: 1) untuk menentukan soal-soal yang cacat atau tidak berfungsi penggunaannya; dan 2) untuk meningkatkan kualitas butir soal melalui tiga komponen analisis, yaitu tingkat kesukaran, daya pembeda, dan pengecoh soal, serta meningkatkan pembelajaran melalui ambiguitas soal dan keterampilan tertentu yang menyebabkan peserta didik sulit dalam merespon butir soal.

Beberapa interprestasi yang dapat ditampilkan terkait dengan data analisis butir adalah pertama, data analisis butir tidak analog dengan validitas butir. Tes-tes psikologi harus memperhitungkan validitas butir, seperti *construct validity*. Namun, untuk tes hasil belajar, meneliti konsistensi internal butir tampak lebih penting dibandingkan menganalisis validitasnya. Hal ini karena tes hasil belajar lebih menyandarkan diri pada validitas isi. Jadi kriteria internal menjadi lebih penting untuk diperhitungkan dan kriteria internal mendasarkan diri pada skor total tes.

Kedua, indeks daya beda butir tidak selalu suatu ukuran kualitas butir. Artinya rendahnya indeks daya beda butir bukan ukuran rendahnya kualitas butir tersebut. Ada beberapa alasan mengapa indeks daya beda butir bisa bernilai rendah: 1) semakin sukar atau semakin mudah suatu butir, semakin rendah indeks daya bedanya, tetapi guru sering membutuhkan item-item yang sukar atau mudah agar representatif terhadap karakteristik materi dan tujuan belajar siswa; dan 2) tujuan item yang berhubungan dengan tes keseluruhan akan mempengaruhi besarnya indeks daya beda butir. Hal ini karena skor total merupakan kriteria internal yang digunakan. Skor total merupakan gabungan skor keseluruhan butir, baik yang sukar maupun yang mudah, dari berbagai pokok bahasan dengan segala keragaman karakteristiknya dan dari keragaman jenjang tes.

Dalam melakukan pengukuran diperlukan perangkat tes yang valid dan reliabel, sehingga dapat memperoleh hasil pengukuran yang sesuai dengan apa yang hendak diukur. Untuk mengetahui kualitas suatu alat ukur perlu dilakukan uji psikometrik terhadap alat ukur tersebut. Para ahli psikometrika telah menetapkan kriteria bagi suatu alat ukur psikologis untuk dapat dinyatakan sebagai alat ukur yang baik dan mampu memberikan informasi yang tidak menyesatkan (Azwar, 1986). Butir-butir dalam perangkat tes yang dipengaruhi oleh faktor-faktor lain selain yang hendak diukur dinamakan bias butir. Istilah bias item dan istilah *Differential Item Functioning* (*DIF*) sering digunakan oleh pakar pengukuran untuk merujuk pada konsep yang sama. Istilah bias item maknanya lebih luas daripada istilah

DIF yang merupakan hasil temuan dari pengolahan statistik. Oleh karena itu, yang menjadi permasalahan tulisan ini adalah: 1) metode apa saja yang dapat digunakan untuk mendeteksi keberadaan *DIF* pada butir soal ujian nasional berdasarkan teori tes klasik?; dan 2) metode mana yang paling sensitif dalam mendeteksi keberadaan *DIF* pada butir soal ujian nasional tersebut berdasarkan teori tes klasik.

Berdasarkan rumusan masalah tersebut di atas, tujuan penulisan ini dimaksudkan untuk menjelaskan: 1) berbagai metode yang dapat digunakan untuk mendeteksi keberadaan *DIF* pada butir soal ujian nasional berdasarkan teori tes klasik (*classical test theory*); dan 2) kelebihan dan kekurangan masing-masing metode yang digunakan dan mengetahui metode mana yang paling sensitif dalam mendeteksi keberadaan *DIF* butir soal ujian nasional.

Kajian Literatur dan Pembahasan

Konsep *Differential Item Functioning* (*DIF*)

Bias butir merupakan salah satu ancaman terhadap validitas pengukuran karena skor tercemar oleh sesuatu yang tidak direncanakan untuk diukur. Apabila suatu butir 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 sistemik dalam pengukuran tersebut. Bias butir memiliki dua karakter, yaitu arah dan besaran. Besaran bias dapat diestimasi secara statistik. Suatu item dikatakan bias apabila dua kelompok yang memiliki kemampuan sama memperoleh hasil yang berbeda pada butir soal tersebut. Secara matematis bias butir dapat dinyatakan dalam bentuk probabilitas (Rahayu W, 2008). Artinya orang yang mempunyai kemampuan sama, tetapi tidak memiliki peluang yang sama untuk memperoleh jawaban benar. Apabila suatu butir relatif lebih sulit untuk kelompok yang memiliki budaya dan latar belakang pengalaman tertentu, maka berarti butir tersebut bias. Bias butir dalam suatu pengukuran mengindikasikan adanya kesalahan sistematik dalam pengukuran tersebut. Prosedur dalam mendeteksi bias butir yang digunakan akan menentukan apakah butir soal yang diberikan akan memberikan informasi yang valid.

Tampak di sini bahwa bias butir atau butir yang bias itu muncul karena: 1) butir ujites mengukur ciri peserta yang seharusnya tidak diukurnya; dan 2) butir tes ikut mengukur ciri yang seharusnya tidak diukurnya itu, sehingga skor butir di antara kelompok atau subkelompok peserta ujites yang seharusnya tidak berbeda, kini menjadi berbeda. Di Amerika Serikat peristiwa bias butir ini menjadi masalah yang cukup besar. Mereka berkata bahwa bias butir itu merugikan etnik Negro dan menguntungkan etnik kulit putih. Di pihak lain, kaum feminis juga berkata bahwa bias butir itu merugikan kaum wanita dan menguntungkan kaum pria (Naga, 1992).

Berhadapan dengan tuduhan tersebut, bias butir di sana mendapat perhatian yang serius. Karena itu, ada ahli yang tidak ingin menggunakan istilah bias butir. Mereka menamakannya *Differential Item Functioning* (*DIF*), yakni pemfungsian yang berbeda dari butir uji tes. Suatu butir menunjukkan *DIF* kalau responsi butir tidak berfungsi sama pada sub-kelompok peserta yang berbeda. Sebaliknya, suatu butir tidak menunjukkan *DIF* kalau karakteristik butir berfungsi sama pada subkelompok peserta yang berbeda.

Kalau butir uji tes itu berfungsi untuk mengukur ciri X, maka butir itu menunjukkan *DIF* dengan catatan butir uji tes itu tidak mengukur X secara sama pada subkelompok peserta yang berbeda. Dan butir uji tes itu tidak menunjukkan *DIF* kalau karakteristik butir itu mengukur X secara sama pada subkelompok peserta yang berbeda (Naga, 1992). Suatu butir soal disebut bias, apabila butir soal tersebut memperbesar kemungkinan sekelompok orang untuk menjawab benar atau menjawab salah.

Informasi yang diperoleh dari tes yang mengandung bias butir soal akan merugikan atau menguntungkan sekelompok peserta, karena mereka dapat memperoleh skor yang lebih tinggi atau rendah dari skor yang seharusnya mereka peroleh. Sebagai contoh, jika suatu butir soal secara sistematik lebih menguntungkan kelompok peserta wanita, maka butir soal tersebut mengandung bias yang positif terhadap wanita (bias gender), begitu pula sebaliknya.

Selain bias gender, ada pengelompokan lain seperti bias budaya, bias bahasa, dan bias etnik. Bias butir soal secara statistika dapat diestimasi arah dan besarannya, sehingga dapat dilakukan koreksi secara Statistika atau Matematika. DIF dapat diidentifikasi dan diukur dengan berbagai metode, salah satunya adalah melihat perbedaan probabilitas menjawab benar dari dua kelompok yang diteliti. Dengan kata lain, *DIF* adalah perbedaan probabilitas menjawab

benar butir soal dari dua kelompok yang berbeda setelah mengontrol tingkat kemampuan (Crocker & Algina, 1986). Bias butir dapat terjadi sebanyak jenis pengelompokan yang diinginkan oleh peneliti. Namun, pengelompokan yang sering dilakukan oleh peneliti adalah bias karena budaya atau gender. Butir disebut bias budaya apabila perbedaan kelompok yang akan diteliti atau diperbandingkan ditetapkan berdasarkan aspek budaya, ras, dan bahasa yang digunakan.

Selanjutnya ada dua faktor yang mempengaruhi timbulnya bias butir, yang secara umum bias butir disebabkan oleh: 1) item itu sendiri yang dalam penelitian ini disebut sebagai faktor internal; dan 2) faktor di luar butir yang dalam penelitian ini disebut faktor eksternal. Ketika kajian bias butir difokuskan pada faktor internal berarti fokus deteksi bias butir dalam karakteristik butir. Apabila kajian bias butir difokuskan pada faktor eksternal, maka fokus deteksi bias butir yaitu peserta tes. Bias butir karena faktor internal terjadi apabila kajian difokuskan pada komponen butir, misalnya bentuk butir, materi butir tes, kalimat dan kata yang digunakan, gambar, petunjuk, dan obyek atau stimulus yang digunakan dalam butir tes.

Secara konseptual, *DIF* dikatakan muncul pada sebuah butir soal, jika peserta yang mempunyai kemampuan yang sama pada konstruks yang diukur oleh tes, tetapi dari kelompok yang berbeda, mempunyai peluang yang berbeda dalam menjawab benar soal tersebut (Hulin & Parson, 1983). Konstruk yang sama, misalnya mengukur hanya satu kemampuan atau *unidimensional* dan kelompok yang berbeda, contohnya kelompok laki-laki dan kelompok perempuan.

Selanjutnya, Hambleton (1991) mengemukakan bahwa suatu butir menunjukkan *DIF* jika peserta tes memiliki kemampuan sama berada dalam kelompok yang berbeda, tidak mempunyai probabilitas sama untuk menjawab betul. Jadi, suatu butir mengandung *DIF* bila dua kelompok peserta tes yang memiliki kemampuan sama memiliki probabilitas menjawab betul yang tidak sama pada butir tersebut.

Lebih lanjut, Hambleton (1991) mengemukakan bahwa suatu butir menunjukkan *DIF* jika peserta tes memiliki kemampuan sama berada dalam kelompok yang berbeda, tidak mempunyai probabilitas sama untuk menjawab betul. Jadi, suatu butir mengandung *DIF* bila dua kelompok peserta tes yang memiliki kemampuan sama memiliki probabilitas menjawab betul yang tidak sama pada butir tersebut.

Untuk menentukan apakah suatu butir terindikasi suatu *DIF* atau tidak, diperlukan indeks *DIF*, yaitu indeks yang menunjukkan sekuat indikasi *DIF* ada pada butir soal itu. Jika tingkat indikasi *DIF* tersebut secara praktik signifikan, dapat dengan mengujinya memakai uji statistik tertentu atau hanya dengan indeksnya saja, maka butir soal yang bersangkutan dikatakan terdeteksi sebagai butir yang bias. Dalam konteks teori responsi butir, terjadi atau tidaknya *DIF* pada sebuah butir soal terletak pada fungsi respons butir (*Item Response Function*) untuk butir tersebut pada kelompok yang dipersoalkan. Kurva yang menggambarkan fungsi respons disebut kurva respons butir atau kurva karakteristik (*Item Characteristic Curve ICC*).

Untuk melakukan pendeteksian keberadaan *DIF* pada butir tes, sebuah populasi dibagi menjadi dua kelompok, yaitu kelompok vokal dan kelompok referensi. Kelompok vokal merupakan kelompok yang diselidiki apakah ada butir yang mengandung *DIF* pada kelompok tersebut. Kelompok referensi merupakan kelompok pembanding. Kedua kelompok diambil dari populasi dan mengerjakan butir pada perangkat tes yang sama pula. Perangkat tes yang sama memiliki validitas dan reliabilitas yang sama.

Tipe *Differential Item Functioning* (*DIF*)

Hambleton (1991) juga mengemukakan definisi *DIF* secara operasional dihubungkan dengan kurva karakteristik butir, yaitu suatu butir menunjukkan *DIF* apabila kurva karakteristik butir pada subkelompok berbeda tidak berhimpit, dan sebaliknya suatu butir tidak menunjukkan *DIF* apabila kurva karakteristik butir dari subkelompok yang berbeda ternyata berhimpit. Sebagaimana yang dikemukakan oleh Lord (1980) suatu butir menunjukkan *DIF* apabila dua kurva karakteristik butir dari dua kelompok berbeda. Penentuan apakah suatu butir soal terindikasi *DIF* atau tidak memerlukan indeks *DIF*, yaitu indeks yang menunjukkan seberapa kuat indikasi *DIF* ada pada butir itu.

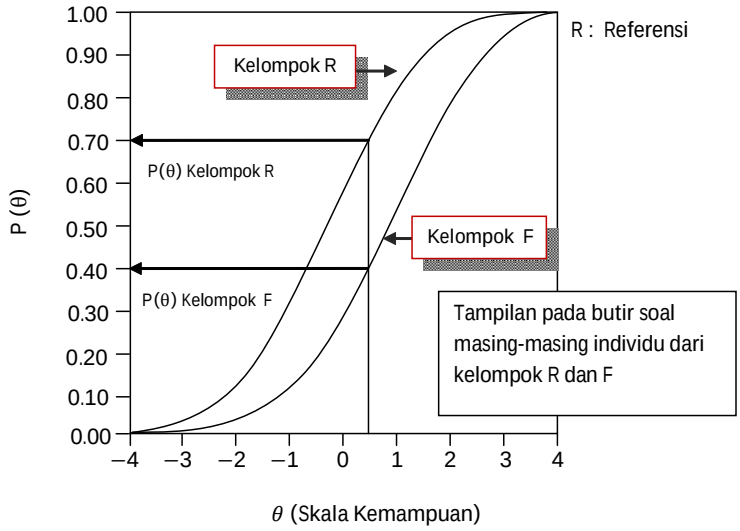
Terdapat dua jenis *DIF*, yaitu *DIF uniform* (konsisten) dan *DIF non uniform* (tidak konsisten). *DIF uniform* muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan *DIF non uniform* muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan. Berdasarkan pembahasan di atas,

dapat ditarik suatu kesimpulan bahwa suatu butir menunjukkan tidak DIF apabila kurva karakteristik butir dari dua kelompok peserta tes yang memiliki kemampuan sama menunjukkan berhimpit. Pengertian berhimpit adalah dua kelompok memiliki pola garis yang sama dan sejajar dengan kemampuan siswa yang menjawab butir tes. Umumnya terdapat dua jenis *DIF,* sebagai berikut.

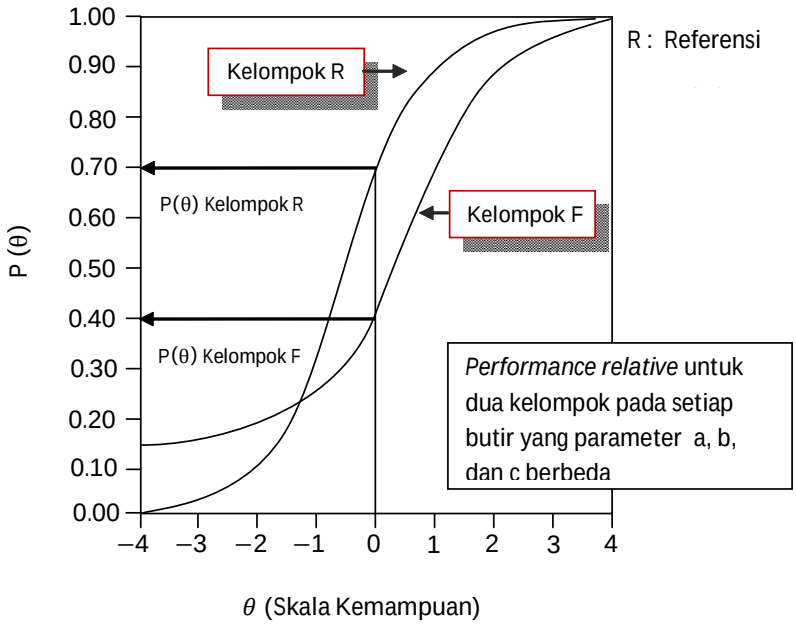
Pertama, *DIF uniform* (konsisten) dan DIF tidak *uniform* (tidak konsisten). *DIF* uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sebagaimana pada Gambar 1.

Pada Gambar 1 tersebut *DIF Uniform* terjadi pada saat *ICCs* dari dua kelompok adalah berbeda dan tidak berpotongan. Hal ini menjelaskan bahwa untuk satu kelompok memiliki rentang kemampuan yang tidak sama. Hal ini terjadi ketika dua *ICCs* memiliki parameter daya beda yang sama.

Kedua, *DIF* tidak *uniform* muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap kemampuan, sebagaimana pada Gambar 2.



Gambar 1. Kemungkinan Pola Jawaban yang Benar untuk Kelompok R dan F



Gambar 2. Kemungkinan Pola Jawaban yang Benar untuk Kelompok R dan F

Pada Gambar 2 tersebut *ICCs* untuk dua kelompok adalah berbeda, tetapi berpotongan pada satu titik pada skala kemampuan tertentu. Jika dikaitkan dengan pengertian interaksi, pada uji statistik analisis varian, *DIF uniform* terjadi jika tidak terdapat interaksi antara tingkat kemampuan peserta dan keanggotaan kelompok dan *DIF* tidak *uniform* terjadi jika terdapat interaksi antara tingkat kemampuan peserta tes dan keanggotaan kelompok (Hambleton, 1991). DIF *uniform* terjadi jika kurva karakteristik butir untuk suatu butir soal berbeda kelompok yang berbeda dan kedua kurva tersebut tidak saling berpotongan. Sebaliknya, tidak *uniform* terjadi jika kurva karakteristik butir untuk suatu butir soal berbeda untuk kelompok yang berbeda, namun kedua kurva tersebut berpotongan.

Pendeteksian Klasik Keberadaan DIF

Telah dikemukakan di atas bahwa bias butir terjadi karena dua hal. Pertama, skor dari butir itu dipengaruhi oleh sumber variasi yang terletak di luar sumber variasi yang dimaksud untuk diukur oleh butir uji tes tersebut. Kedua, pengaruh sumber variasi tersebut memberikan keuntungan yang tidak adil pada suatu subpopulasi uji tes terhadap subpopulasi uji tes lainnya yang sama-sama menggunakan butir uji tes itu. Biasanya uji tes bias merupakan jumlah dari butir bias yang terdapat di dalam uji tes itu. Dalam hal ini, dapat saja terjadi bahwa sejumlah butir bias di dalam uji tes itu saling mengkompensasi kebiasaan mereka. Butir bias atau uji tes bias berkaitan dengan cara penskoran butir atau penskoran perangkat uji tes. Pendeteksian bias dilakukan pada skor yang diperoleh melalui skor klasik, sehingga pendeteksian itu dinamakan pendeteksian klasik terhadap bias.

Ada banyak cara untuk mendeteksi butir bias dan uji tes bias pada skor yang dicapai melalui teori skor klasik. Beberapa di antaranya yang akan dibahas adalah korelasi kelompok tunggal (*single group validity*), korelasi diferensial (*differential validity*), prosedur diskriminasi butir (*item discrimination procedure*), metode plot delta (*delta plot method*), metode Standarisasi, metode *Chi-square* Scheuneman (*Scheuneman chi-squared approach*), metode *Chi-square* Camilli (*Camilli chi-square approach*), metode Mantel-Haenszel, prosedur standar yang telah dikembangkan oleh Dorans dan Kulick, dan metode estimasi bias butir dengan Analisis Faktor Konfirmatori.

Korelasi Kelompok Tunggal dan Korelasi Diferensial

Ada kalanya populasi peserta uji tes terdiri atas lebih dari satu macam subpopulasi, masing-masing dengan ciri yang berbeda-beda. Di Amerika Serikat, hubungan di antara uji tes dengan subpopulasi semacam ini sering memperoleh sorotan yang tajam di dalam masyarakat manakala subpopulasi itu adalah golongan kulit putih berhadapan dengan golongan minoritas atau kelamin pria berhadapan dengan kelamin wanita. Bahkan hal ini berkaitan dengan kegiatan gerakan hak warganegara yang terdapat di Negara itu.

Masyarakat dan gerakan itu menuntut agar butir uji tes dan bahkan seluruh perangkat ujites yang dikerjakan oleh peserta uji tes tidak sampai bias terhadap golongan atau kelamin tertentu dalam pengertian memberi keuntungan yang tidak adil kepada salah satu golongan atau jenis kelamin. Untuk itu, mereka berusaha mendeteksi kemungkinan adanya butir bias atau perangkat uji tes bias di dalam pengujian. Mereka baru merasa puas apabila pengujian itu bebas dari butir atau perangkat uji tes yang bias (Zumdo, 1999).

Kadangkala subpopulasi di dalam populasi peserta uji tes memiliki banyak ciri yang spesifik bagi setiap subpopulasi. Selain itu, ciri spesifik dari setiap subpopulasi yang diteliti juga harus mendapat perhatian yang serius dalam pengukuran, terutama dalam bidang pendidikan. Ciri tersebut sama di dalam subpopulasi tetapi berbeda di antara subpopulasi peserta uji tes. Di sini cukup memperhatikan dua macam ciri. Pertama adalah ciri yang mau diukur oleh uji tes yang dimiliki, yang disebut ciri kemampuan atau kinerja (*performance*) peserta tes.

Kedua, yaitu ciri lainnya di luar ciri yang akan diukur oleh uji tes dan spesifik bagi setiap subpopulasi peserta. Ciri ini berbeda di antara subpopulasi. Pendeteksian bias yang berbentuk korelasi kelompok tunggal ini berusaha mendeteksi koefisien korelasi di antara uji tes dengan ciri eksternal dari salah satu subpopulasi itu. Korelasi demikian tidak terdapat pada subpopulasi lainnya. Misalnya dengan memperhatikan dua subpopulasi, yaitu masing-masing subpopulasi 1 dan subpopulasi 2.

Berikut adalah penjelasan singkat mengenai koefisien korelasi dan jenis-jenis korelasi yang banyak digunakan dalam pengukuran pendidikan. Analisis korelasi merupakan salah satu teknik statistik yang

sering digunakan untuk mencari hubungan antara dua variabel. Korelasi diartikan sebagai hubungan. Analisis korelasi bertujuan untuk mengetahui pola dan keeratan hubungan antara dua atau lebih variabel. Dua variabel yang hendak diselidiki korelasinya biasanya dilambangkan dengan X dan Y. Perlu diingat bahwa uji korelasi tidak membedakan adanya variabel dependen dan variabel independen. Arah korelasi menunjukkan pola gerakan variabel Y terhadap gerakan variabel X. Terdapat dua arah korelasi, yaitu *positive correlation, negative correlation,* dan *nihil correlation.*

Jika kenaikan nilai X diikuti oleh kenaikan nilai Y dan sebaliknya terjadi penurunan nilai X yang juga diikuti oleh penurunan nilai Y, atau dengan kata lain perubahan pada satu variabel diikuti oleh perubahan variabel yang secara teratur dengan arah gerakan yang sama, maka hubungan ini disebut sebagai *positive correlation.* Jika kenaikan nilai X justru diiringi dengan penurunan nilai Y dan sebaliknya penurunan nilai X dibarengi dengan kenaikan nilai Y, atau dengan kata lain perubahan pada satu variabel diikuti oleh perubahan variabel yang lain secara teratur dengan arah gerakan yang berlawanan, maka hubungan seperti ini disebut sebagai *negative correlation.*

Selain arah korelasi, permasalahan yang juga penting, yaitu seberapa besar tingkat keeratan hubungan antara dua variabel. Misalnya, jika ada yang mengatakan hubungan antara merokok dengan narkoba sangat erat, maka akan muncul pertanyaan seberapa erat hubungan tersebut? Untuk menentukan keeratan hubungan tentu akan lebih mudah kalau dinyatakan dalam koefisien korelasi. Koefisien korelasi merupakan ukuran besar kecilnya atau kuat tidaknya hubungan antara variabel-variabel apabila bentuk hubungan tersebut linier. Koefisien korelasi sering dilambangkan dengan huruf (r). Koefisien korelasi dinyatakan dengan bilangan, bergerak antara 0 sampai +1 atau 0 sampai -1. Nilai korelasi mendekati +1 atau -1 berarti terdapat hubungan yang kuat, sebaliknya korelasi yang mendekati nilai 0 berarti terdapat hubungan yang lemah. Apabila korelasi sama dengan 0, maka berarti antara kedua variabel tidak terdapat hubungan sama sekali. Apabila korelasi +1 atau -1, maka berarti terdapat hubungan yang sempurna antara kedua variabel.

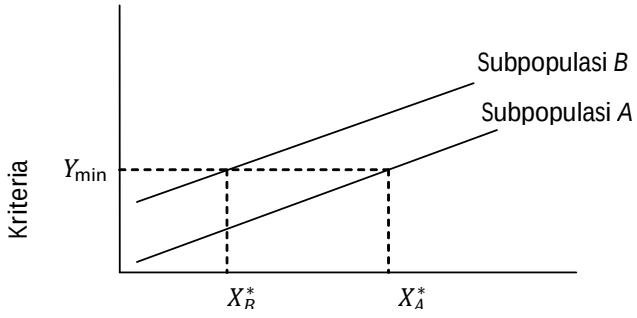
Notasi positif (+) atau negatif (-) menunjukkan arah hubungan antara kedua variabel. Notasi positif (+) berarti hubungan antara kedua variabel searah (*positive correlation*), jika variabel satu naik maka variabel yang lain juga naik. Notasi negatif (-) berarti kedua variabel berhubungan terbalik (*negative correlation*), artinya kenaikan satu variabel akan diikuti dengan penurunan variabel lainnya. Arah dan nilai koefisien dapat dirangkum sebagai berikut: 1) jika nilai $r > 0$, maka artinya telah terjadi hubungan yang linier positif (*positive correlation*), yaitu makin besar nilai variabel X makin besar pula nilai variabel Y, atau makin kecil nilai variabel X makin kecil pula nilai variabel Y yang akan diprediksi; 2) jika nilai $r < 0$, maka artinya telah terjadi hubungan yang linier negatif (*negative correlation*), yaitu makin besar nilai variabel X makin kecil nilai variabel Y, atau makin kecil nilai variabel X maka makin besar pula nilai variabel Y; 3) jika nilai $r=0$, maka artinya tidak ada hubungan sama sekali antara variabel X dan variabel Y; dan 4) jika nilai $r=1$ atau $r=-1$, maka dapat dikatakan telah terjadi hubungan linier sempurna, berupa garis lurus, sedangkan untuk r yang makin mengarah ke angka 0 (nol), maka garis makin tidak lurus.

Hal yang harus dijelaskan di sini adalah bahwa analisis korelasi hanya mengukur ko-variasi. Pengukuran ini bersifat numerik dan menunjukkan suatu korelasi yang terdapat antara dua atau lebih variabel. Pengukuran ini tidak menunjukkan adanya hubungan sebab-akibat, tetapi ini adalah suatu hal yang harus digarisbawahi. Dua variabel yang sudah terbukti mempunyai hubungan atau korelasi tidak berarti mempunyai hubungan sebab-akibat, tetapi hubungan sebab-akibat pasti menunjukkan bahwa kedua variabel mempunyai hubungan. Terdapat tiga jenis pembagian korelasi, yaitu pertama: korelasi positif dan korelasi negatif yang telah diuraikan di atas. Kedua, korelasi sederhana, parsial, dan ganda. Ketiga, korelasi linier dan nonlinier.

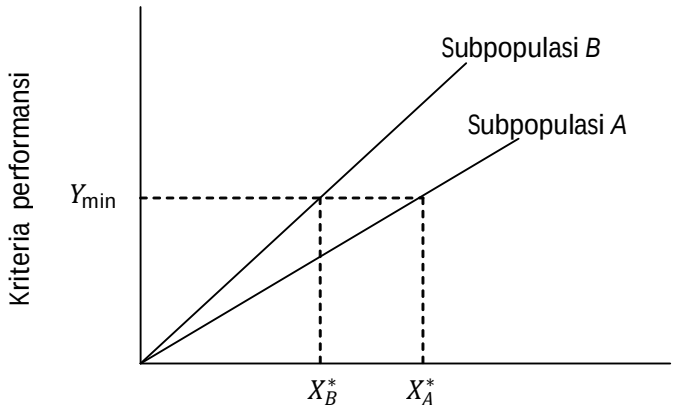
Uji hubungan melalui teknik statistik korelasi dapat dilakukan terhadap bermacam data, baik data yang berskala interval, ordinal maupun nominal. Korelasi yang dipergunakan untuk uji hubungan antarsesama data interval adalah korelasi produk *moment* dari Pearson (*Pearson product moment correlation*). Jika yang dikorelasikan adalah antara data yang berskala ordinal, maka teknik korelasi yang digunakan adalah korelasi tata jenjang (*rank-order correlation*). Sebaliknya jika yang dikorelasikan adalah antara data berskala interval dengan yang berskala nominal, maka teknik korelasi yang digunakan adalah korelasi *point-biserial* (*point-biserial correlation*).

Salah satu *tool* yang paling banyak digunakan dalam penelitian adalah analisis regresi. Analisis regresi menjadi sangat terkenal dan banyak digunakan karena ada beberapa yang istimewa di dalam analisis regresi, di antaranya di dalam analisis regresi sudah termasuk analisis korelasi antara variabel independen (X) yang juga sering disebut faktor-faktor penyebab, dengan variabel dependen (Y).

Di dalam model regresi ini, patokan kinerja Y pada peserta diregresikan secara linier terhadap skor ujites X yang menjadi prediktor. Di sini kita memperhatikan dua subpopulasi uji tes masing-masing subpopulasi 1 dan subpopulasi 2. Regresi di antara Y dan X untuk subpopulasi 1 dan subpopulasi 2 adalah $Y = A_1 + B_1 X$ dan $Y = A_2 + B_2 X$. Bias itu terjadi karena pada kedua subpopulasi itu terdapat kekeliruan baku yang berbeda, di mana titik potong yang berbeda dari garis regresi sumbu Y disebabkan oleh koefisien A yang berbeda. Salah satu yang khas dari analisis regresi yaitu adanya persamaan yang dihasilkan.



Gambar 3. Bias Butir yang Disebabkan oleh Perbedaan Intersep



Gambar 4. Bias Butir yang Disebabkan oleh Perbedaan Slope

Karena digunakan untuk memprediksi, variabel bebas juga sering disebut sebagai variabel prediktor. Yang selalu melekat dalam analisis regresi yaitu analisis korelasi, karena kalau variabel independen (X) berpengaruh nyata terhadap variabel dependen (Y) atau disebut berkorelasi kuat, maka sudah otomatis segala perubahan pada nilai X tersebut akan sangat berpengaruh pada nilai Y.

Prosedur Diskriminasi Butir

Prosedur ini menggunakan korelasi butir-butir atau korelasi biserial untuk mendeteksi keberadaan bias atau butir uji tes. Dalam keadaan tidak bias, koefisien korelasi biserial butir untuk setiap subpopulasi, yaitu sama atau paling tidak sama secara statistika. Kalau sampai terjadi bahwa koefisien korelasi biserial butir terdapat pada suatu subpopulasi dan tidak terdapat pada subpopulasi lainnya, maka di dalam bahan butir uji tes itu ada hal yang dimiliki oleh subpopulasi itu, tetapi tidak dimiliki oleh subpopulasi lainnya. Dengan demikian, butir ujites itu bias terhadap subpopulasi itu dibandingkan dengan terhadap subpopulasi lainnya (Naga, 1992).

Dalam pendeteksian klasik estimasi bias butir dapat dilakukan dengan menghitung daya beda butir. Deteksi bias butir menggunakan prosedur yang sama dengan delta tingkat kesulitan, hanya saja data yang digunakan untuk membuat plot adalah data daya beda dari masing-masing kelompok yang akan diteliti. Butir yang lebih diskriminatif pada salah satu kelompok mengindikasikan butir tersebut mengandung bias butir. Estimasi bias butir menggunakan parameter tingkat kesulitan butir diukur berdasarkan persentase menjawab benar dan daya beda butir diukur dengan korelasi *point biserial.*

Hal ini menimbulkan permasalahan, karena: 1) karakteristik orang dan karakteristik butir dianalisis secara terpisah; 2) indeks tingkat kesukaran butir bergantung pada kelompok peserta tes (*dependent group*); dan 3) skor yang diperoleh bergantung pada tes yang berarti bahwa skor seseorang bergantung pada tes yang berarti bahwa skor seseorang bergantung pada tes yang dikerjakan. Selain itu, skor yang diperoleh dari tes yang berbeda tidak dapat diperbandingkan, karena tidak menggunakan skala yang sama dan tidak ada hubungan fungsional.

Sudah ada sejumlah penelitian tentang bias butir yang menggunakan prosedur diskriminasi butir ini. Penelitian tersebut menemukan bahwa makin tinggi

indeks diskriminasi suatu butir terhadap suatu subpopulasi, makin bias butir itu dalam pengertian bahwa butir itu memberikan skor yang lebih menguntungkan subpopulasi itu daripada subpopulasi lainnya. Prosedur pendeteksian bias ini dilakukan dengan jalan menghitung koefisien korelasi biserial dari setiap butir uji tes terhadap setiap subpopulasi peserta (Ridho dan Azwar, 2005). Dengan membandingkan nilai koefisien korelasi biserial itu, dapat dipastikan butir uji tes mana yang bias dan mana yang tidak bias. Metode tersebut mempunyai kemampuan untuk menstransformasi nilai-nilai yang dihasilkan dari analisis butir secara klasik. Namun, dewasa ini prosedur ini mendapat kritik dari ahli pengukuran dalam pendidikan, terutama yang menyangkut validitas dan reliabilitas butir soal.

Metode Plot Delta

Pembahasan tentang taraf sukar butir pada analisis butir, kita menemukan penentuan taraf sukar butir melalui skala delta. Dengan menggunakan distribusi normal baku, proporsi jawaban benar pada ujites dipadankan kepada kumulasi distribusi dan selanjutnya melalui nilai z pada distribusi normal baku itu. Delta adalah ukuran taraf sukar butir, ditentukan skala delta dengan persamaan: $\Delta = 13 + 4z$ Dalam uji tes, setiap butir memiliki taraf kesukaran butir. Kalau taraf kesukaran butir itu dilihat dari setiap subpopulasi peserta uji tes, maka kita menemukan sejumlah taraf kesukaran butir, masing-masing terkait dengan setiap subpopulasi peserta uji tes. Kalau butir uji tes itu tidak bias terhadap salah satu subpopulasi, maka taraf kesukaran butir pada semua subpopulasi adalah sama (Naga, 1992).

Populasi dibagi dalam dua subpopulasi, masing-masing subpopulasi 1 dan subpopulasi 2 (misalnya pria dan wanita). Selanjutnya kita memperhatikan butir uji tes ke-i. Untuk butir ke-1, taraf sukar butir adalah Δi_1 dan Δi_2 butir adalah bias jika $\Delta i_1 \neq \Delta i_2$ dan tidak bias jika $\Delta i_1 = \Delta i_2$. Jika taraf kesukaran butir dalam skala delta pada kedua subpopulasi peserta uji tes ini kita plot ke dalam sumbu koordinat, maka plot itu akan sama jauh dari kedua sumbu koordinat itu. Plot itu akan terletak pada garis yang membentuk sudut arah 45 derajat dan melewati titik asal (Shepard, 1982).

Dari penjelasan hubungan sumbu koordinat, maka dapat ditentukan hubungan linier taraf sukar butir antara Δi_1 dan Δi_2 Rerata taraf sukar butir pada subpopulasi 1 dan 2 adalah $\mu\Delta_1$ dan $\mu\Delta_2$. Sedangkan kekeliruan baku taraf sukar butir pada subpopulasi 1 dan subpopulasi 2 adalah σ_1 dan σ_2. Koefisien korelasi di antara taraf sukar butir pada subpopulasi 1 dan subpopulasi 2 adalah ρ. Hubungan linier di antara dua taraf sukar butir apabila tidak ada bias dapat ditentukan dengan persamaan: $\Delta_2 = k\Delta_1 + d$. Penyimpangan dari garis linier di antara dua taraf sukar adalah bias. Diperlukan suatu ketentuan untuk memutuskan apakah suatu butir bias atau tidak bias terhadap kriteria. Semua butir uji tes yang plot deltanya terletak pada garis tersebut dianggap tidak bias, sedangkan butir uji tes yang plotnya terletak di luar garis itu mungkin bias. Kalau plot delta ini melibatkan sampel peserta uji tes, maka penyimpangan dari garis itu dapat diuji secara statistika dengan menghitung jarak plot ke garis itu.

Pendekatan *Chi-square* Camilli

Pada prinsipnya pendekatan *Chi-square* Camilli sama dengan pendekatan *Chi-square* Scheuneman. Pendekatan *Chi-square* Scheuneman hanya memperhatikan proporsi jawaban betul. Oleh karena itu, pendekatan *Chi-square* Scheuneman dikenal juga sebagai *Chi-square* jawaban benar (*correct* atau *true*). Pendekatan *Chi-square* Camilli, selain memperhatikan proporsi jawaban betul, juga memperhatikan proporsi jawaban salah. Namun apabila responden menjawab soal betul semua, maka butir tidak dapat dianalisis.

Oleh karena itu, pendekatan *Chi-square* Camilli dikenal sebagai *Chi-square* penuh atau lengkap (Camilli & Shepard, 1994). Semua rumus pada pendekatan *Chi-square* Scheuneman digunakan di sini. Perbedaan hanya terletak pada perhitungan akhir, yakni pada *Chi-square*. Statistik *Chi-square* Camilli adalah χ^2 betul = $\chi 2Pk_1 + \chi 2Pk_2$ dan χ^2 salah = $\chi 2Qk_1 + \chi 2Qk_2$, sehingga *Chi-square* Camilli menjadi $\chi^2 C = \sum \chi^2$ betul + $\sum \chi^2$ salah.

$$vC = (SP - 1)K$$

Langkah yang paling sulit pada pendeteksian bias butir dengan menggunakan kedua pendekatan tersebut terletak pada penentuan interval skor. Penentuan interval skor ini dilakukan dengan menggunakan metode penyetaraan butir dan penyamaan skala dengan metode gandeng melingkar (Shepard and Everill, 1981).

Pedoman untuk membentuk interval skor tidak selalu dapat menghasilkan satu macam interval skor

yang sama sekalipun kita telah berusaha mengikuti pedoman itu secara cermat. Penentuan batas yang berbeda pada interval skor akan menghasilkan perangkat interval skor yang berbeda dan hal ini akan menghasilkan yang berbeda. Kelemahan yang lain, yaitu semua peserta di dalam interval yang sama dianggap memiliki kemampuan yang sama sekalipun di antaranya ada yang berbeda kemampuannya.

Pendekatan *Chi-Square* Scheuneman. Pada uji statistik dengan menggunakan metode *Chi-square* Scheuneman, terlebih dahulu perangkat tes terlebih dahulu ditentukan karakteristiknya menggunakan teori tes klasik. Butir soal yang tingkat kesulitannya kurang dari 0,2 berdasarkan teori tes klasik dan butir yang tidak cocok dengan model logistik satu parameter atau model *Rasch* juga tidak diikut-sertakan dalam analisis selanjutnya (Scheuneman dan Bleintein, 1989).

Selanjutnya, dilakukan estimasi parameter butir perangkat tes secara terpisah pada kelompok laki-laki dan kelompok perempuan dan estimasi varians-kovarians dari parameter butir. Kemudian ditentukan matriks untuk menghitung nilai *Chi-square* masing-masing kelompok. Langkah-langkah yang dilakukan pada pendeteksian *DIF* dengan metode *Chi-square* Scheuneman, sebagai berikut: 1) populasi responden atau peserta didik dibagi ke dalam subpopulasi yang diduga terkena bias butir, yaitu ke dalam subpopulasi 1 (peserta didik pria) dan subpopulasi 2 (peserta didik wanita); 2) skor responden dibagi ke dalam interval-interval atau selang-selang (k interval). Ada k interval skor pada subpopulasi 1(pria) dan ada k interval skor pada subpopulasi 2 (wanita); dan 3) butir tidak bias jika proporsi jawaban betul pada setiap interval adalah sama untuk dua subpopulasi itu (Retnawati, 2005).

Sedangkan langkah-langkah yang dilakukan dalam pemeriksaan bias butir soal tes adalah sebagai berikut: 1) menentukan butir mana yang akan diperiksa bias atau tidak bias, misalkan butir ke-8; 2) pada butir ke-8 tersebut diurutkan skor responden dari kecil ke besar, dan dengan memperhatikan salah satu skor, misalkan skor 12; 3) memperhatikan semua responden dengan skor 12 dan mereka dipecahkan ke dalam dua subpopulasi yang diduga terkena bias butir; 4) menghitung proporsi jawaban betul pada setiap populasi: a) subpopulasi 1: frekuensi betul dan salah, dan b) subpopulasi 2: frekuensi betul dan salah; 5) skor lainnya dibagi ke dalam interval sehingga seluruhnya (termasuk skor 12) menjadi 3 sampai 5 interval. Menurut Scheuneman, setiap interval mengandung 10 sampai 20 skor; 6) karena Scheuneman menggunakan distribusi probabilitas *Chi-square,* jadi setiap sel harapan jangan kurang dari 5 skor (syarat pendekatan ke distribusi probabilitas *Chi-square*); 7) memperhatikan statistik setiap interval skor pada setiap subpopulasi, misalnya, interval skor ke-k.

Subpopulasi	Interval	Banyaknya responden	Banyaknya jawaban betul
1 2	k_1 k_2	m_{k1} m_{k2}	A_{k1} A_{k2}

8) Statistik Jawaban: Proporsi jawaban betul P dan jawaban salah Q;

Subpop 1	$P_{k1} = A_{k1}/m_{k1}$	$Q_{k1} = 1 - P_{k1}$
Subpop 2	$P_{k2} = A_{k2}/m_{k2}$	$Q_{k1} = 1 - P_{k1}$
Gabungan subpop	$P_{kt} = \frac{A_{k1} + A_{k2}}{m_{k1} + m_{k2}}$	$Q\_kt = 1 - P\_kt$

9) Harapan matematik jawaban betul dan salah

Subpop 1	$E_{Pk1} = P_{kt}\ m_{k1}$ $E_{Qk1} = Q_{kt}\ m_{k1}$
Subpop 2	$E_{Pk2} = P_{kt}\ m_{k2}$ $E_{Qk2} = Q_{kt}\ m_{k2}$

10) Statistik *Chi-square* tiap interval

$$\chi^2_{P_{k1}} = \frac{\left(A_{k1} - E_{P_{k1}}\right)^2}{E_{P_{k1}}} \quad \chi^2_{P_{k2}} = \frac{\left(A_{k2} - E_{P_{k2}}\right)^2}{E_{P_{k2}}}$$

11) Chi-square Scheuneman pada K interval

$$\chi^2_s = \sum_{k=1}^{K} \chi^2_{P_{k1}} + \sum_{k=1}^{K} \chi^2_{P_{k2}}$$

$V\_s = (SP - 1)(K - 1)$

SP = banyaknya subpopulasi

K = banyaknya interval

Misalkan, suatu data dibagi ke dalam dua subpopulasi berupa subpopulasi 1 kelompok pria dan subpopulasi 2 kelompok wanita. Skor 12 dijadikan satu interval sebagai k=3. Selanjutnya skor 1 sampai 9 menjadi k-1, skor 10 sampai 11 menjadi k=2, skor 13 sampai 14 menjadi format statistik.

Statistik	Interval skor k				Jumlah
	1	2	3	4	
Skor	1–9	10–11	12	13-14	

Isi tiap interval (harapan) agar tidak kurang dari 5 atau menurut Scheuneman di antara 10 sampai 20.

Untuk memahami metode pendeteksian bias butir dengan metode *Chi-square* Scheuneman berikut diberikan satu contoh perhitungannya. Suatu data dibagi ke dalam dua subpopulasi berupa subpopulasi 1 dan subpopulasi 2 (misal pria dan wanita). Dalam pendeteksian bias butir dengan menggunakan metode ini sampel yang akan menjadi obyek penelitian harus dibedakan menjadi dua kelompok, yaitu kelompok fokus dan kelompok referensi. Hal ini dilakukan untuk mengetahui seberapa besar nilai sensitivitas dari metode pendeteksian yang digunakan.

Metode Mantel-Haenszel

Prosedur Mantel-Haenszel (M-H) dikembangkan pertama kali oleh Mantel dan Haenszel pada tahun 1959, dan digunakan untuk mendeteksi DIF oleh Holland dan Thayer pada tahun 1988 yang sampai sekarang ini digunakan untuk menganalisis keberadaan DIF yang seragam (*uniform*). Prosedur MH bermanfaat untuk mengestimasi dampak dari ukuran sampel terhadap analisis keberadaan *DIF*. Selain itu digunakan untuk menguji hipotesis nol yang tidak mengandung *DIF* (Haenszel and Sato, 1995).

Penggunaan metode Mantel-Haenszel berdasarkan asumsi-asumsi sebagai berikut: 1) hanya mengukur satu dimensi (unidimensi); 2) kemampuan peserta dinyatakan dengan skor total yang diperoleh pesert tes dari seluruh butir soal dengan menganggap setiap soal memiliki bobot yang sama; 3) level dari kemampuan peserta tes dapat digolongkan dalam M kelompok yang berurutan; dan 4) setiap peserta tes dapat dikelompokkan kepada satu dan hanya satu kelompok, yaitu kelompok acuan atau kelompok fokus (Budiyono, 2005).

Menurut ETS (*Educational Testing Service*) suatu butir soal dikatakan mengandung DIF setelah dideteksi dengan metode ini, yaitu nilai mutlak Δ lebih besar atau sama dengan 1,5. Jika nilai *MH Chi-Square Statistic* lebih besar dari 1 butir-butir soal yang dianalisis mempengaruhi kelompok referensi, sedangkan bila nilai *MH Chi-Square Statistic* kurang dari 1, maka menunjukkan bahwa butir-butir yang dianalisis cenderung mempengaruhi kelompok fokal. Statistik *Chi-square* M-H yang digunakan adalah untuk menguji hipotesis statistik dengan nilai $\alpha_{MH} = 1$, di mana distribusi data bersifat distribusi normal dengan jumlah peserta tes yang besar.

Metode Standarisasi

Dalam perhitungan standarisasi dilakukan perhitungan regresi nonparametrik butir untuk masing-masing kelompok. Perbedaan empiris uji regresi butir merupakan indikasi ada bias butir (Dorans & Holland, 1993). Apabila kelompok yang ingin diteliti disebut f, kelompok yang menjadi acuan disebut r, *I* merupakan skor butir, dan M merupakan variabel yang dipasangkan, maka definisi bias butir dengan metode standarisasi adalah E_f (I/M) = E_r (I/M). Sedangkan E_f (I/M) adalah uji regresi butir empiris pada kelompok yang ingin diteliti dan E_r (I/M) adalah uji regresi butir empiris pada kelompok acuan. Apabila D_m adalah bias butir dengan metode standarisasi, maka perhitungan $D_m = E_f$ (I/M) - E_r (I/M).

Dorans dan Schmith (1989) telah melakukan penelitian menggunakan metode standarisasi untuk mengidentifikasi bias butir. Metode ini didasarkan pada data dalam bentuk fungsi respon butir di mana probabilitas menjawab benar diestimasi berdasarkan proporsi jawaban benar butir pada setiap tingkat kemampuan. Estimasi probabilitas sukses pada setiap tingkat skor ditetapkan berdasarkan kelompok acuan. Kelompok acuan adalah kelompok yang ditetapkan acuan kelompok vokal. Kelompok vokal adalah kelompok yang diminati peneliti dan biasanya adalah kelompok yang memiliki skor rendah.

Dorans dan Hollands (1993) menyatakan metode standarisasi dan Mantel-Haenszel memiliki kemiripan prosedur, yaitu: 1) keduanya merupakan metode nonparametrik; 2) tidak menuntut model respon *likelihood*; dan 3) keduanya menunjukkan kelebihan yang sama, yaitu efisien secara statistik

dan mudah dalam perhitungannya. Lord (1980) mengkritik analisis bias butir dengan metode plot delta, Mantel-Haenszel dan standarisasi. Analisis butir dengan metode Mantel-Haenszel berasumsi bahwa semua butir memiliki tingkat kesulitan yang sama. Dalam metode plot delta dan metode standarisasi menggunakan parameter tingkat kesulitan butir, yaitu dengan cara menghitung proporsi jawaban benar (*proportion correct*).

Analisis Faktor Konfirmatori

Analisis faktor merupakan suatu perangkat teknik untuk memproses data yang memuat pengujian hipotesis dan teknik untuk mereduksi data (Sappaile, 2006). Analisis faktor konfirmatori digunakan untuk mengkonfirmasikan sejumlah faktor yang mendasari pemikiran penelitian (Kaluge, 1988). Untuk mendeteksi secara akurat sumber bias perlu diteliti kontribusi dan interaksi berbagai variabel yang diperkirakan menjadi sumber bias. Estimasi bias butir dengan menggunakan IRT tidak dapat mendeteksi berbagai sumber bias secara simultan. Prosedur yang dapat mereduksi kontribusi dan interaksi antar varibel sumber bias adalah analisis faktor. Harga parameter pada analisis faktor dapat ditransformasikan menjadi parameter IRT (Wardani, 2009). Muatan faktor digunakan sebagai parameter kualitas butir, baik daya beda butir maupun tingkat kesulitan butir. Estimasi *DIF* dilakukan dengan membandingkan parameter daya beda dan parameter tingkat kesulitan butir dari dua kelompok.

Simpulan dan Saran

Simpulan

Kegiatan analisis butir soal memiliki banyak manfaat, yaitu: 1) dapat membantu para pengguna tes dalam evaluasi atas tes yang digunakan; 2) sangat relevan bagi penyusunan tes informal dan lokal seperti tes yang disiapkan guru untuk siswa di kelas; 3) mendukung penulisan butir soal yang efektif; 4) secara materi dapat memperbaiki tes di kelas; dan 5) meningkatkan validitas soal dan reliabilitas soal; 6) menentukan apakah suatu fungsi butir soal sesuai dengan yang diharapkan; 7) memberi masukan pada siswa tentang kemampuan dan sebagai dasar untuk bahan diskusi di kelas; 8) memberikan masukan pada guru tentang kesulitan siswa; 9) memberikan masukan pada aspek tertentu untuk mengembangkan kurikulum; dan 10) merevisi materi yang dinilai atau diukur.

Butir-butir dalam perangkat tes yang dipengaruhi faktor-faktor lain selain yang hendak diukur dinamakan bias butir. Istilah bias item dan istilah *Differential Item Functioning* (*DIF*) sering digunakan oleh pakar pengukuran untuk merujuk pada konsep yang sama. Istilah bias item maknanya lebih luas daripada istilah *DIF* yang merupakan hasil temuan dari pengolahan statistik. Ada banyak cara untuk mendeteksi butir bias dan uji tes bias pada skor yang dicapai melalui teori skor klasik. Beberapa diantaranya yang akan dibahas adalah korelasi kelompok tunggal (*single group validity*), korelasi diferensial (*differential validity*), prosedur diskriminasi butir (*item discrimination procedure*), metode plot delta (*delta plot method*), metode Standarisasi, metode *Chi-square* Scheuneman (*Scheuneman chi-squared approach*), metode *Chi-square* Camilli (*Camilli chi-square approach*), metode Mantel-Haenszel, prosedur standar yang telah dikembangkan oleh Dorans dan Kulick, dan metode estimasi bias butir dengan Analisis Faktor Konfirmatori.

Saran

Saran bagi pengembang tes dan peneliti yang akan meneliti bias butir atau *DIF* agar memasukkan bias butir sebagai salah satu kriteria mutu dalam memilih butir tes dengan memperhitungkan variabel internal dan variabel eksternal. Selain itu, dalam mendeteksi bias butir atau *DIF* menggunakan metode yang dapat memperhitungkan variabel eksternal untuk mengestimasi besar dan arah bias butir yang dideteksi.

Saran bagi guru dan dosen dalam mengembangkan soal-soal ujian akhir sekolah maupun ujian akhir semester hendaknya memperhatikan analisis butir soal dengan memperhatikan kaidah-kaidah pengukuran yang ada. Selain itu, untuk menghindari adanya *DIF* dalam butir soal, sehingga butir soal yang telah dibuat harus diujicoba agar dapat diketahui kualitas butir soal dan terhindar dari *DIF.* Pada akhirnya butir soal yang telah dibuat memiliki kualitas yang memadahi sebagai langkah awal dalam pengembangan butir soal menjadi bank soal yang bebas dari *DIF*.

Pustaka Acuan

Anthony J. Nitko. 1996. *Educational Assessment of Students.* New Jersey: Prentice-Hall International.

Anastasi, A dan S. Urbina. 1997. *Psychological Testing.* New Jersey: Prentice Hall, Inc.

Azwar, S. 1986. *Dasar-Dasar Psikometri.* Yogyakarta: Pustaka Pelajar.

Budiyono. 2005. Perbandingan Metode Mantel-Haenszel, SIBTEST, Regresi Logistik dan Perbedaan Peluang dalam Mendeteksi Keberadaan DIF. *Disertasi*, Yogyakarta: Universitas Negeri Yogyakarta.

Crocker, L dan James A. 1986. *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart, and Winston.

Dorans, N.J dan Schmitt, A.P. 1989. *The Methods for Dimensionality Assessment and DIF Detection.* Paper Presented at The Annual Meeting of The National Council on Measurement in Education, San Fransisco.

Dorants, N.J. dan Holland, P.W. 1993. *DIF Detection and Description: Mantel-Haenszel and Standardization.* Lawrence Erlbaum Associates, Inc.Publishers.

Hambleton, Ronald K, H. Swaminathan, dan Rogers, H. J. 1991. *Fundamentals of Item Response Theory.* California: Sage Publications.

Haenszel S. Kim, dan Sato A. Kohen. 1995. A Comparison of Lord's Chi Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Function, *Journal of Aplied Measurement in Education* 8 (1995): pp. 291-312.

Hsin-Hun Li dan William Stout. 1996. A New Procedure for Detection of Crossing DIF, *Journal Psychometrica* 61 (1996): pp. 647-677.

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

Kaluge, Lauren. 1988. *Analisis Faktor sebagai Eksploratori Variabel Laten*, Surabaya: FIP IKIP.

Naga, Dali S. 1992. *Pengantar Teori Skor pada Pengukuran Pendidikan.* Jakarta: Universitas Gunadarma, Besbtas.

Ridho, A. dan Saifuddin Azwar. 2005. Keberfungsian Item Tes UAN Matematika SMA di Propinsi DIY Tahun Pelajaran 2003/2004, Makalah disampaikan pada Seminar Nasional: *Hasil Penelitian tentang Evaluasi Hasil Belajar serta Pengelolaannya*, Pascasarjana UNY didukung oleh Direktorat P2TK & KPT dan HEPI, Yogyakarta, 14-15 Mei 2005.

Rahayu, W. 2008. Pengaruh Metode *Linking* terhadap Banyak Butir *False Positive* pada Pendeteksian *DIF* Berdasarkan Teori Responsi Butir. *Disertasi,* Jakarta: Universitas Negeri Jakarta.

Retnawati, H. 2005. Keberfungsian Butir Diferensial pada Perangkat Tes Seleksi Masuk SLTP Mata Pelajaran Matematika. Makalah disampaikan pada Seminar Nasional: *Hasil Penelitian tentang Evaluasi Hasil Belajar serta Pengelolaannya*, Pascasarjana UNY didukung oleh Direktorat P2TK & KPT dan HEPI, Yogyakarta, 14-15 Mei 2005.

Sappaile, B. I. 2006. Dimensi dan Reliabilitas Suatu Instrumen dengan Menggunakan Rotasi Varimax pada Analisis Faktor Eksploratori. *Jurnal Pendidikan dan Kebudayaan*, Tahun 12 No.060. pp. 351-362.

Scheuneman, J.D dan Bleintein, 1989. A Consumer's Guide to Statistics for Identifying Diferential Item Functioning. *Applied Measurement in Education* 7 (1989): p.255.

Shepard, L.A, Cammili, G. dan Everill. 1981. Comparison of Prosedures for Detecting Test Item Bias with Both External and Internal Ability Criteria. *Journal of Education Statistics* 6 (1981). p. 319.

Shepard, L.A, 1982. *Detecting of Bias.* Dalam R.A Berk (ed). Handbook of Methods for Detecting Item Bias. Baltimore: Johns Hopkins University Press. p. 23.

Zumdo, B. D. 1999. *A Handbook on the Theory and Methods of DIF: Logistic Regression and Modeling as a Unitary Framework for Binay and Likert-Type Item Scores.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Wardani, N. Y. 2009. Perbedaan Sensitivitas Metode Analisis Faktor Konfirmatori (AFK) dan Model Persamaan Struktural, *Jurnal Pendidikan dan Kebudayaan* 15 (2009): p. 445.