EFEK KEBERGANTUNGAN BUTIR LOKAL PADA VALIDITAS BUTIR MAUPUN TES DALAM TEORI RESPON BUTIR (STUDI PADA UJICOBA SOAL OLIMPIADE MATEMATIKA)

Abadyo

Jurusan Matematika FMIPA Universitas Negeri Malang.

Abstrak

Salah satu asumsi dalam teori respon butir (TRB) yaitu respon-respon yang diberikan oleh peserta ujian pada butir-butir tes adalah bebas satu sama lain (*Locally Independence*). Akan tetapi, penelitian-penelitian sebelum ini telah menunjukkan bahwa banyak tes saat ini yang berisi kebergantungan butir lokal (*Local Item Dependence*). Apabila kebergantungan butir lokal ini tidak diperhitungkan dalam mengestimasi parameter-paramater butir, tes, maupun kemampuan peserta tes akan menyebabkan kesesatan dalam estimasi itu. Dalam kajian ini, penulis (a) mereviu metode-metode untuk mendeteksi kebergantungan butir lokal (KBL), (b) mendiskusikan penggunaan *testlets* untuk memperhitungkan adanya KBL dalam kontek himpunan-pimpunan butir yang bergantungan, dan (c) mengevaluasi hasil-hasil estimasi kemampuan peserta tes dan reliabilitas skor tes. Hasil kajian ini memberi peringatan bahwa kehadiran KBL akan memberi pengaruh yang kuat pada estimasi kemampuan peserta tes. Secara praktis, efek-efek dari kehadiran KBL didiskusikan pada tes-tes berdasar bagian (sub-tes) dengan menggunakan teori respon butir (TRB).

***Kata kunci**: kebergantungan butir local, validitas, reliabilitas, sub-tes.*

PENDAHULUAN

Ahli-ahli pengukuran secara rutin mengasumsikan respon-respon yang diberikan oleh seorang peserta tes atau responden pada butir-butir tes dan atau kuesioner adalah bebas satu sama lain. Akan tetapi, belakangan ini banyak hasil riset yang menunjukkan bahwa banyak tes yang memuat ketergantungan butir-butirnya, dan karena tidak diperhitungkan adanya ketergantungan ini maka akan membawa kesesatan estimasi bagi parameter-parameter butir, tes, maupun kemampuan peserta tes (Abadyo, 2008). Sejumlah peneliti yang telah membicarakan ketergantungan lokal (y.i: Andrich, 1985; Rosenbaum, 1988; Steinberg & Thissen, 1996; Thissen & Steinberg, 1988; Thissen, Steinberg & Mooney, 1989; Wilson, 1988; Wilson & Adams, 1995; Yen, 1984, 1993), mereka telah menunjukkan bahwa *testlet* dapat dipandang sebagai model dari konjungsi antara butir-butir yang memiliki ketergantungan lokal.

Unit yang paling dasar dalam penyusunan tes adalah butir tes. Butir-butir tes ini diperlukan untuk (a) menjangkau kecukupan domain isi atau konstrak tes, dan (b) mengadakan estimasi keandalan dari kecakapan pengambil tes. Dalam teori tes klasik telah lama diketahui bahwa salah satu cara untuk meningkatkan reliabilitas skor tes adalah menambah banyaknya butir dalam tes itu. Akan tetapi, apabila hal ini dikerjakan hanya dengan menduplikasi butir-butir yang sama akan tidak menyelesaikan persoalan validitas dan reliabilitas pengukuran. Jadi, para penyusun tes harus mengembangkan butir-butir yang menyediakan informasi tunggal yang terkait dengan kemampuan, ketrampilan, maupun pengetahuan pengambil tes. Butir-butir yang menyediakan informasi berlebihan tidak diperlukan. Butir-butir yang tidak memberikan kontribusi tunggal pada suatu asesmen tidak meningkatkan representasi konstrak dan menambah faktor-faktor konstrak yang tidak relevan yang mungkin terkait dengan suatu butir. Dengan pertimbangan seperti ini, apa yang sekarang diketahui sebagai **kebergantungan butir lokal** (KBL) perlu diperhitungkan dalam pengembangan dan penyekoran tes.

Konsep KBL paling baik dipahami melalui kerangka kerja teori respon butir (TRB). Model-model TRB yang paling populer mengkhususkan pada karakteristik laten tunggal (*unidimensional*) untuk menghitung semua kebergantungan antara butir-butir sebagaimana semua

perbedaan kemampuan di antara para pengambil tes secara statistik. Karakteristik laten ini dinotasikan theta (θ), yang membedakan butir-butir ditinjau dari tingkat kesulitannya, dan membedakan para pengambil tes berdasar kemampuannya. Peluang pengambil tes memberikan respon tertentu pada suatu butir adalah fungsi dari lokasi pengambil tes pada θ dan satu atau lebih parameter-parameter butir itu (tergantung pada model TRB yang dipilih). Fungsi ini mengambarkan hubungan antara butir dan θ. Karena model TRB adalah probabilistik, maka kebebasan butir bersyarat kepada θ harus diasumsikan di antara respon-respon sembarang pasangan butir-butir itu. Kebebasan bersyarat ini disebut kebebasan butir lokal (*Locally Independence*) (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968). Ketika KBL nampak pada suatu tes, maka terjadi ketidakakuratan estimasi parameter-parameter butir dan kemampuan pengambil tes (Fennessy, 1995; Sireci, Thissen, & Wainer, 1991; Thissen, Stienberg, & Mooney, 1989). Tambahan lagi, KBL mengintrodusir adanya tambahan dimensi (dan umumnya tidak diharapkan) ke dalam tes itu dengan mengorbankan konstrak yang sedang diperhatikan (Wainer & Thissen, 1996).

Teori-teori tes klasik juga menaruh perhatian pada estimasi-estimasi yang dihasilkan tidak akurat ketika kebergantungan antar butir tidak diperhitungkan secara cermat. Sebagai contoh, Kelley (1924), Guilford (1936), Thorndike (1951), Anastasi (1961), dan yang lain memperingatkan bahwa butir-butir yang terkait dengan stimulus bersama atau skenario bersama (yaitu, seperangkat butir terkait dengan penggalan bacaan, tabel, gambar, peta ,dsb.) semuanya harus diletakkan pada sisi yang sama dari *half-test* ketika menghitung reliabelitas *split-half*. Sebaliknya, apabila butir-butir itu diletakkan berhadap-hadapan maka akan terjadi penggelembungan hasil estimasi reliabelitas, karena butir-butir ini saling bergantung dan kebergantungan ini akan membuat penggelembungan korelasi yang palsu antara ke dua *half-tests* itu. Karena koefisien alpha menggambarkan semua kemungkinan *split-half* itu, maka koefisien ini juga akan membengkak bila kita tidak memperhitungkan kebergantungan antar butir itu. Oleh sebab itu, persoalan ketidakcermatan penghitungan yang terkait dengan KBL tidak terbatas pada TRB.

Walaupun ketergantungan lokal tidak diharapkan, ada alasan yang baik untuk memasukkan butir-butir yang saling bergatung pada suatu asesmen. Banyak persoalan nyata membutuhkan penyelesaian yang terkait pada persoalan atau penyelesaian soal tunggal dalam cara tahap demi tahap. Jadi, termasuk butir-butir bergantungan pada konteks di suatu tes mungkin meningkatkan validitas konstrak. Contoh-contoh mengenai konstrak-konstrak yang bertautan, butir-butir yang bergantungan termasuk butir-butir yang menuntut responden menyelesaikan persoalannya dan selanjutnya memberikan penjelasan bagaimana mereka memperoleh jawaban itu atau penggunaan butir berganda untuk mengukur secara komprehensif pada bagian bacaan, skenario, atau grafik. Oleh sebab itu, yang menjadi tantangan bagi pengembang tes adalah tidak mengeliminasi ketergantungan butir, tetapi bagaimana membuat model yang pantas sedemikian hingga KBL tidak terjadi. Untungnya ada beberapa metode untuk mendeteksi KBL, dan pemodelan yang tepat untuk konstruk-konstruk yang bertautan dengan KBL dalam model TRB.

Dalam makalah ini, kita mengkaji pendekatan-pendekatan yang berbeda untuk mendeteksi KBL pada butir-butir soal ujicoba olimpiade matematika. Tujuan khusus dari kajian ini adalah untuk mengintesvigasi (a) tingkat KBL yang ada dalam soal ujicoba olimpiade matematika dalam bentuk pilihan ganda, (b) dampak dari kebergantungan butir ini pada estimasi reliabilitas, dan (c) pengunaan penyekoran berbasis-*testlet* dalam meminimalkan konsekuensi negatif dari kebergantungan butir ini. Kajian dari tingkat terjadinya KBL dalam data pilihan ganda akan memberikan kesempatan untuk mengeksplorasi metode-metode penyekoran yang bisa meminimalkan efek-efek bias. Ketika penyekoran secara dikotomus digunakan pada soal dalam bentuk pilihan ganda, maka dapat dilihat seberapa serius bias yang terjadi pada statistik butir, tes, dan kemampuan pengambil tes yang dapat menyediakan informasi tentang kualitas psikometrik dari suatu tes.

Jika diperoleh kebergantungan dalam data ketika digunakan himpunan-himpunan butir dalam konteks-kebergantungan, maka salah satu metode penyekoran butir-butir itu adalah menggunakan model TRB politomus dan *testlet* (Thissen, et al., 1989; Thissen, Billeaud, McLeod, & Nelson, 1997; Yen, 1993). Suatu *testlet* adalah unit penyekoran dalam suatu tes yang lebih kecil

dari pada suatu tes, terdiri dari butir-butir yang mungkin atau mungkin tidak mempunyai kebergantungan lokal (Wainer & Kiely, 1987). Sebagai contoh, penggalan bacaan pada seksi *READING COMPREHENSION* dalam *Toefl* dan butir-butir yang dikaitkannya dapat disusun menjadi satu *testlet*. Dalam penggunaan model TRB politomus untuk menyekor *testlet*, data itu dapat dianalisis dan sementara itu kebebasan lokal di antara *testlet* yang berbeda tetap dapat dijaga.

Berbicara tentang akurasi estimasi reliabilitas, estimasai yang paling akurat adalah pada butir-butir yang bebas lokal, karena butir-butir yang bergantung lokal cenderung menggelembungkan estimasi reliabilitas (Sireci et al., 1991). Ketika nampak butir-butir berbeda terkait pada kebergantungan suatu penggalan, pengelompokan secara bersama butir-butir itu ke dalam suatu *testlet* merupakan model struktur tes yang lebih tepat. Dengan menggunakan strategi ini, kebebasan butir lokal dapat dipertahankan untuk semua *testlet*, karena *testlet* dimodelkan sebagai unit (y.i, sebagai butir politomus). Jadi, himpunan butir-butir yang bergantung lokal sebagai model *testlet* dari struktur tes berbasis-*testlet* akan memenuhi asumsi kebebasan lokal TRB.

Salah satu kekhawatiran penggunaan model-model TRB politomus adalah hilangnya informasi penting yang termuat di masing-masing butir. Dengan menjumlahkan skor-skor butir dalam suatu *testlet* untuk menghitung skornya *testlet* itu, informasi yang terkait dengan jawaban benar pada butir-butir tertentu dari para peserta tes menjadi hilang. Sebagai contoh, jika ada 10 testlet yang masing-masing terdiri dari 5 butir disekor secara dikotomus menggunakan model TRB tiga-parameter, maka ada 3 x 5 x 10 = 150 parameter butir yang akan diestimasi. Sebaliknya, jika tes itu dikalibrasi menggunakan model politomus bagi struktur *testlet* (misalnya dihitung dengan menggunakan model respon berjenjang, Samejima (1969)), maka hanya ada satu parameter pembeda dan 5 parameter *threshold* akan diestimasi untuk masing-masing testlet (total ada 6 x 10 = 60 parameter). Jadi, beberapa informasi pengukuran mungkin akan hilang ketika butir-butir itu diringkas ke dalam *testlet*. Ketika kebergantungan butir tidak tampak, pembentukan *testlet* dan penyekoran secara politomus tidak akan memperbaiki estimasi parameter-paramater butir, tes, maupun kemampuan peserta tes. Oleh sebab itu, derajat keberadaan KBL pada suatu tes harus diketahui dengan pasti sebelum menentukan bagaimana sebaiknya model tes yang akan dipakai.

METODE-METODE ASESMEN KBL

Beberapa metode yang berbeda untuk mengases kebergantungan butir lokal dalam data dikotomus telah dikembangkan. Yen(1984) mengusulkan statistik Q_3 sebagai indeks dari KBL. Q_3 adalah korelasi dari residu bagi pasangan butir-butir setelah pemartisian estimasi *trait*. Untuk menghitung Q_3, estimasi kemampuan ($\hat{\theta}_a$) dihitung untuk setiap peserta ujian dan digunakan untuk mengestimasi performansi harapan dari peserta ujian pada setiap butir (y.i., E_{ja}, dimana *j* menyatakan butir dan *a* menyatakan peserta ujian). Residu (dinyatakan d_{ja}) dihitung dengan mengambil deviasi antara performansi peserta ujian terobservasi dan performansi harapan peserta ujian pada suatu butir. Jadi, untuk butir-butir *j* dan *j'*, Q_3 adalah korelasi dari skor-skor deviasi seluruh peserta ujian (y.i., $Q_{3jj'} = r_{(dj,\, dj')}$).

Kemampuan peserta ujian digunakan baik untuk penghitungan skor-skor harapan ($E_{ja} = \hat{\theta}_a$) maupun skor-skor terobservasi, duplikasi ini cenderung menghasilkan nilai-nilai Q_3 yang secara marjinal negatif (Kingston & Dorans, 1982). Apabila tiada kebergantungan lokal, nilai harapan dari Q_3 adalah $-1/(n-1)$, dimana *n* adalah banyaknya butir dalam tes itu. Statistik Q_3 ini telah digunakan secara sukses oleh Yen(1993), Fennessy(1995), dan Chen & Thissen (1997).

Indeks lain yang diusulkan untuk mengidentifikasi KBL di dalam praktek adalah statistik G^2, yang berdistribusi χ^2 dengan derajat bebas 1 (Bishop, Fienberg, & Holland, 1975; Chen & Thissen, 1997). Statistik G^2 adalah uji rasio kemungkinan:

$$G^2 = -2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln\left(\frac{E_{ij}}{O_{ij}}\right)$$

Statistik G^2 ini telah dibandingkan dengan statistik Q_3-nya Yen; keduanya dapat mendeteksi adanya kebergantungan butir lokal dengan kuasa tertentu dan Q_3 nampaknya memiliki performansi lebih baik dari pada G^2 (Chen & Thissen, 1997).

Korelasi bersyarat antar-butir juga telah diusulkan sebagai ukuran dari KBL (Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Huynh & Ferrara, 1994). Dalam metode ini, peserta ujian dibagi ke dalam (delapan sampai sepuluh) kelompok berdasar pada skor tes total, dan korelasi antar-butir dihitung dalam masing-masing interval skor tes. Korelasi antar-butir dalam suatu *testlet* dapat dirata-rata sepanjang tiap-tiap tingkatan skor dan masing-masing butir untuk mendapatkan ukuran statistik dari KBL untuk masing-masing *testlet*. Ukuran dari KBL dalam *testlet* ini dapat dibandingkan dengan statistik yang dihitung pada seluruh *testlet*. Jika rata-rata korelasi dalam *testle*t lebih tinggi dari korelasi antar *testlet*, estimasi reliabilitas yang diturunkan dari penyekoran dikotomus butir-butir itu secara positif akan menjadi bias. Lee dan Frisbie (1999) juga menghitung rata-rata korelasi dalam-dan antara-*testlet* dengan pendekatan teori perumuman untuk mengases reliabilitas tes yang tersusun atas *testlet*. Ketika penyekoran testlet digunakan pada himpunan butir-butir mereka, perbedaan antara hasil perhitungan koefisien reliabilitas dan permuman adalah kecil. Berdasarkan keadaan itu dapat disimpulkan bahwa penyekoran *testlet* merupakan penyekoran yang tepat digunakan bila dibandingkan dengan penyekoran butir dikotomus.

Sireci et al.(1991), Wainer (1995), dan Wainer & Thissen (1996) juga menunjukkan bahwa kehadiran KBL pada suatu tes dapat diketahui secara pasti dengan membanding-kan dua estimasi reliabilitas yang terpisah. Estimasi pertama mengasumsikan semua butir adalah bebas lokal dan mengabaikan struktur *testlet*. Model estimasi kedua inheren dengan struktur *testlet*, yang melibatkan pembentukan *testlet* untuk semua himpunan butir dalam kontek-kebergantungan. Jika estimasi reliabilitas berdasar-*testlet* secara substansi lebih rendah daripada estimasi berdasar-butir, ini menunjukkan adanya kehadiran KBL.

Dalam makalah ini, penulis akan mendiskusikan penggunaan dua metode untuk mendeteksi keberadaan KBL. Pertama, kita modelkan himpunan-himpunan butir dalam kontek-kebergantungan menggunakan *testlet* dan membandingkan hasil estimasi reliabilitas yang diperoleh ketika tes dipertimbangkan hanya terdiri dari butir-butir yang bebas lokal. Kedua, kita menghitung statistik Q_3 di antara butir-butir itu.

METODOLOGI

Data

Data berasal dari pengadministrasian soal ujicoba olimpiade matematika SD di DIY 2008 yang dilaksanakan pada 25 Oktober 2008, jam 08.00 sampai dengan 10.00. Adapun tempatnya di:

1. SD Sampangan , Banguntapan Bantul.
2. SD Padokan 2 , Kasihan, Bantul.
3. SD Bantul Timur.
4. SD Muhammadiyah, Senggotan, Kasihan, Bantul.

Jumlah seluruh subyek 100 siswa SD kelas VI. Jumlah seluruh butir ada 25 terdiri dari 20 pilihan ganda dan 5 isian singkat. Yang dianalisis adalah soal-soal pilihan ganda dengan 4 obsi.

Analisis Data

Estimasi reliabilitas marjinal TRB dan koefisien a dihitung untuk data baik yang diskor secara dikotomus maupun politomus. Dua strategi yang digunakan untuk menghitung estimasi reliabilitas ini. Strategi pertama didasarkan pada penyekoran tradisional dimana semua butir diperlakukan secara diskrit dan diskor secara dikotomus. Estimasi lain didasarkan pada penyekoran secara politomus *testlet*. Penyekoran berdasar-*testlet* ini, skor seorang peserta tes pada suatu *testlet* dihitung dengan menjumlahkan semua jawaban benar butir-butir dalam *testlet* itu. Perbandingan estimasi reliabilitas ditetapkan oleh dua rancangan penyekoran yang menyediakan pengukuran derajat KBL terkait pada butir-butir yang mengukur penggalan bersama. Sebagai contoh, jika koefisien reliabilitas berdasar-*testlet* lebih rendah dari koefisien reliabilitas berdasarkan penyekoran dikotomus, koefisien terakhir mungkin *overestimate* (Sireci et al., 1991, Thissen et al., 1989). Akan tetapi, seperti yang diperoleh Sireci et al.(1991), penurunan dalam reliabilitas diharapkan lebih sedikit butir-butir *testlet* yang dibentuk dari butir-butir diskrit. Oleh karena itu, untuk tujuan perbandingan itu, *testlet* juga dibentuk secara acak (y,i., penggabungan butir-butir bersama dari

pengalan berbeda) untuk mengukur penurunan reliabilitas terkait dengan proses pembentukan *testlet*.

Analisis Q_3

Data skor dikotomus dikalibrasi menggunakan model TRB logistik 3-parameter. Statistik Q_3 digunakan untuk mengases kebergantungan dalam *testlet* "sejati" (y.i., *testlet* berdasar-penggalan) dan *testlet* " palsu"(y.i., yang dibentuk *testlet* secara acak) untuk tes pilihan ganda (Yen, 1984). Statistik Q_3 dihitung dari *testlet* " palsu" yang menyediakan basis evaluasi besarnya KBL yang diperoleh dalam analisis lain, sebagai perkiraan pengelompokan bersama butir-butir secara acak yang tidak akan menampakkan KBL. Matrik Q_3 untuk setiap tes dihitung menggunakan program IRTNEW (Chen, 1998). Statistik ringkasannya kemudian dibandingkan. Nilai Q_3 dan statistik ringkasannya diperiksa untuk hubungan pola pengurutan efek-efek, himpunan butir, dan jenis-jenis penggalan.

Estimasi Kemampuan

Himpunan data yang diskor berdasar-*testlet* digunakan mengkalibrasi dengan menggunakan MULTILOG (Thissen, 1991). Pemilihan model TRB politomus tidak sulit dilakukan sebab dua model TRB politomus yang dibunakan bersama, model respon berjenjang (Samejima, 1969) dan model kredit parsial umum (Muraki, 1992), menyediakan hasilhasil yang sangat mirip ketika digunakan untuk menganalisis data dengan respon dalam kategori ganda (Thissen et al., 1997).

HASIL

Analisis Reliabilitas

Ringkasan dari hasil analisis reliabilitas a disajikan dalam Tabel 1. Tiga himpunan estimasi disediakan untuk masing-masing bentuk tes: estimasi reliabilitas a tradisional berdasarkan penyekoran semua butir secara dikotomus, kalkulasi reliabilitas berdasar-*testlet* "sejati" menggunakan skor *testlet* untuk semua butir berdasar-penggalan (y.i., kontek-kebergantungan), dan kalkulasi reliabilitas berdasar-*testlet* "palsu" dengan menjumlahkan bersama butir-butir yang dikelompokkan secara acak untuk membentuk *testlet*. Dari 20 butir dijadikan 8 penggalan.

Tabel 1. Koefisien Reliabilitas a

Format Tes	Butir Dikotomus		*Testlet*"sejati"	*Testlet*"palsu"	#Butir dlm penyekoran *testlet*
	#Butir	a			
P . G 1	20	0,85	0,79	0,85	8
P . G 2	20	0,87	0,82	0,87	8

Dari hasil ini, tidak ada perbedaan antara estimasi reliabilitas yang dihitung dari data skor dikotomus dan yang dihitung dari *testlet* "palsu". Sebaliknya, estimasi bagi data skor dikotomus cenderung menjadi besar dari pada kontek-kebergatungan *testlet*. Hasil ini mengindikasikan adanya KBL dalam data itu. Penggunaan rumus Spearman-Brown merupakan salah satu jalan untuk mengestimasi reliabilitas tes yang dapat dibandingkan untuk menentukan besarnya *overestimate* dari reliabilitas dalam kasus dikotomus (Sireci et al.,1991; Wainer, 1995). Tabel 2 menyoroti bias dalam estimasi reliabilitas tes yang diskor secara asli dikotomus.

Tabel 2. Statistik Spearman-Brown (dari Estimasi Koefisien Reliabilitas a)

Format Tes	Perpanjangan *Testlet*"sejati"	Perpanjangan *Testlet*"palsu"
P . G 1	1,51	1,00
P . G 2	1,47	1,00

Hasil dalam Tabel2 menunjukkan bahwa estimasi reliabilitas berdasar pada penyekoran dikotomus dari penggalan-penggalan digelembungkan oleh KBL.

Sebagai tambahan dari estimasi koefisien reliabilitas a, estimasi reliabilitas marjinal berdasar-TRB dihitung dengan menggunakan model logistik 3-parameter untuk butir-butir yang diskor

secara dikotomus, dan model respon berjenjang untuk *testlet* yang diskor secara politomus. Estimasi reliabilitas marjinal ini disajikan dalam Tabel 3. Panjang tes diperpanjang untuk mencapai tingkat estimsai reliabilitas dari data dikotomus disajikan dalam Tabel 4. Hasilnya menunjukkan hal yang sama seperti reliabilitas a dan menguatkan indikasi munculnya KBL.

Tabel 3. Reliabilitas Marjinal TRB

Format Tes	Butir Dikotomus		*Testlet*"sejati"	*Testlet*"palsu"	#Butir dlm penyekoran *testlet*
	#Butir	a			
P . G 1	20	0,87	0,81	0,85	8
P . G 2	20	0,88	0,83	0,87	8

Tabel 4. Statistik Spearman-Brown (dari Estimasi Marjinal TRB)

Format Tes	Perpanjangan *Testlet*"sejati"	Perpanjangan *Testlet*"palsu"
P . G 1	1,57	1,18
P . G 2	1,77	1,10

Reliabilitas marjinal mempunya kecenderungan sedikit lebih tinggi dari koefisien a, tapi umumnya sekitar 0,02 (Wainer & Thissen, 1996).

Analisis Q_3

Matrik Q_3 diperoleh dari setiap format tes. Dengan menggunakan matrik ini nilai rata-rata Q_3 untuk setiap format dihitung dengan merata-rata nilai Q_3 untuk pasangan butir-butir yang terletak dalam *testlet* yang sama. Tabel 5 menunjukkan rata-rata ini untuk *testlet*"sejati" dan *testlet*"palsu".

Tabel 3. Reliabilitas Marjinal TRB

Tes	*Testlet*"palsu"		*Testlet*"sejati"	
	Format 1	Format 2	Format 1	Format 2
Pilihan Ganda (harapan Q_3 : - 0,019)	- 0,026	- 0,018	- 0,024	- 0,032

Rata-rata nilai Q_3 untuk *testlet*"palsu" mendekati nilai harapan, sementara itu nilai rata-rata untuk *testlet*"sejati" lebih tinggi.

DISKUSI & SIMPULAN

Beberapa temuan yang menarik terkait dengan KBL muncul dalam kajian ini. Sejumlah strategi praktis dan mudah diimplementasikan untuk mendeteksi adanya KBL telah ada, walaupun interpretasi dari statistik-statistik ini masih meninggalkan problematik. Pembandingan estimasi reliabilitas melalui penyekoran *testlet* dan *non-testlet* dalam kontek-kebergantungan butir merupakan salah satu cara untuk mendeteksi adanya KBL. Statistik Q_3 lebih berguna untuk mengidentifikasi pasangan butir-butir khusus yang bergantung lokal. Statistik ini adalah deskriptif. Besaran mereka seringkali nampak sangat kecil (nilai yang terbesar disebutkan dalam literatur sekitar 0,10), hal ini menambah kesulitan dalam menginterpretasikan maknanya secara praktis.

Metode-metode untuk segera mendapatkan efek-efek KBL secara praktis adalah bermanfaat bagi investigasi lebih lanjut yang terkait dengan tes dimana penggalan dan himpunan butir digunakan. Kebergantungan butir dapat berdampak serius pada statistik dalam desain tes itu dan skor yang dilaporkan ke peserta tes. Salah satu bidang riset masa depan adalah mendesain tes-lapangan dari berbagai versi himpunan butir dalam kontek-kebergantungan yang dapat mengungkap KBL dan bagaimana KBL ini harus dimodelkan. Arah lain yang potensial bagi riset masa depan adalah investigasi efek penyekoran dan validitas prediktif.

REFERENSI

Abadyo. (2008). Model TRB politomus untuk mengungkap pola pikir multidimensi. Jogyakarta: Makalah dipresentasikan pada seminar pengukuran dan pengujian di PPS UNY tanggal 12 Desember 2008.

Anastasi, A. (1961). Psychological testing (2nd ed.). New York: Macmillan.

Chen, W. (1998). IRTNEW: A computer program for the detection of local item dependence.Chapel Hill, N.C.: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22 (3), 265-289.

Ferrara, S., Huynh, H., & Bagli, H. (1997). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. Applied Measurement in Education,12, 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. Journal of EducationalMeasurement, 36, 119-140.

Guilford, J. P. (1936). Psychometric methods (1st ed.). New York: McGraw-Hill.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, NJ: Sage.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. Journal of EducationalMeasurement, 31, 125-141.

Kingston, N. M., & Dorans, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (ETS Research Report 82-12). Princeton, NJ:Educational Testing Service.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. Applied Measurement in Education, 12, 237-255.

Lord, F. M., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. Applied Psychological Measurement, 18(13),245-56.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores.Psychometrika, Monograph Supplement, No. 17.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Thissen, D. (1991). MULTILOG 6.3 [Computer program]. Mooresville, IN: Scientific

Software.Thissen, D., Billeaud, K., McLeod, L., & Nelson, L. (1997, August). A brief introduction toitem response theory for items scored in more than two categories. Paper presented at the NationalAssessment Governing Board Achievement Levels Workshop, Boulder, CO.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiplecategorical response models. Journal of Educational Measurement, 26, 247-260.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), Educational measurement (pp.560-620). Washington, D.C.: American Council on Education.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991Law School Admissions Test as an example. Applied Measurement in Education, 8 (2), 157-186.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized-adaptive testing: A case for testlets. Journal of Educational Measurement, 24(3), 185-201.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, 15, 22-29.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item

dependence. Journal of Educational Measurement, 30(3), 187-213.