PERBANDINGAN KEEFEKTIFAN BENTUK TES URAIAN DAN *TESTLET* DENGAN PENERAPAN *GRADED RESPONSE MODEL* (*GRM*)

Purwo Susongko
Jurusan Matematika FKIP UPS Tegal
tirtodiporo@yahoo.com

Abstrak

Penelitian ini bertujuan untuk menemukan: (1) perbandingan nilai fungsi informasi *item* pada bentuk tes uraian dan *testlet* secara empirik dan simulasi, (2) pengaruh banyaknya *item* dan ukuran sampel terhadap perbandingan nilai fungsi informasi *item* pada bentuk tes uraian dan *testlet* secara simulasi, dan (3) keakuratan pemodelan *GRM* pada bentuk uraian dan *testlet.* Data empirik diambil dari respons siswa terhadap tes bentuk uraian dan bentuk *testlet* dari 772 siswa SMA kelas XI yang tersebar di lima SMA di Kabupaten Tegal. Bentuk tes uraian dan *testlet* bersama-sama diberikan pada siswa pada akhir semester I dan di awal semester II dengan waktu tenggang minimal 1 bulan. Data pada penelitian simulasi dibangkitkan dari parameter *item* hasil estimasi pada penelitian empirik dengan program *WinGen 2.* Hasil penelitian menunjukkan bahwa: (1) secara empirik dan simulasi, tes yang disajikan dalam bentuk uraian cenderung memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding dengan tes yang disajikan dalam bentuk *testlet*, (2) secara simulasi, ada pengaruh banyaknya *item* dan ukuran sampel terhadap perbandingan nilai fungsi informasi *item* pada bentuk tes uraian dan bentuk *testlet*, dan (3) pemodelan GRM pada bentuk tes uraian dan *testlet* paling akurat pada kondisi banyaknya *item* 20 dan ukuran sampel 2000.

Kata kunci: *keefektifan, bentuk tes, graded response model*

COMPARISON OF THE EFFECTIVENESS OF THE ESSAY TEST AND TESTLETS THROUGH THE GRADED RESPONSE MODEL (GRM) APPLICATION

Purwo Susongko
Jurusan Matematika FKIP UPS Tegal
tirtodiporo@yahoo.com

Abstract

The purposes of this study are to find: (1) the comparisons between the information function value of the test items in an essay test with that in testlets empirically and through simulations, (2) the effect of the number of items and the sample size on the information function value of the test items in an essay test and that in testlests through simulations, and (3) the accuracy of GRM modeling in an essay test and testlets. The empirical data were collected through students' responses to the essay test and the testlets administered to 772 Year XI students of five Senior High Schools in the Tegal Regeny. Both the essay test and the testlets were administered to the students at the end of semester I and at the beginning of semester II with a time interval of two months at the minimum. The data in the simulation study were generated from the item parameters of the estimation results of the empirical study by using the *WinGen 2* program. The findings of the study are: (1) empirically and through simulation, the essay test tends to have values of the item information function which are higher than those in the testlets, (2) through simulation, there is an effect of the number of items and the sample size on the comparison between the item information function value of the test items in an essay test and that in the testlets. (3) The simulation with a condition of 20 items and a sample size of 2000 testees has the smallest RMSD value in comparison with other conditions.
Key words: *effectiveness*, *essay test and testlets*, *graded response model*

Pendahuluan

Bentuk tes objektif dan bentuk tes uraian masing-masing memiliki kelebihan dan kelemahan. Bentuk tes uraian, memberikan kebebasan kepada setiap penempuh tes untuk mengekspresikan daya nalarnya, sehingga jawaban yang diberikan oleh setiap penempuh tes akan menunjukkan kemampuan berpikir secara kompleks. Namun demikian ada beberapa kelemahan bentuk tes uraian. Bentuk tes uraian dalam memberikan skor membutuhkan waktu yang lama dan relatif lebih sulit, sehingga bentuk uraian sulit digunakan untuk tes-tes yang berskala besar. Di samping itu, penskoran bentuk tes uraian bersifat subjektif dan harus dilakukan oleh ahli atau yang berwenang sehingga tidak dapat dilakukan komputerisasi dalam pensekoranya.

Berbeda dengan bentuk tes uraian, bentuk tes objektif lebih praktis dalam penskorannya. Pada bentuk tes objektif siapa pun yang memeriksa akan memberikan skor yang sama, sehingga kesalahan karena penskoran dapat menjadi kecil, apalagi bila digunakan komputer dalam penskoran. Namun demikian bentuk tes objektif mempunyai peluang menjawab benar dengan menebak cukup tinggi yang ditunjukkan oleh besarnya *blind guessing* maupun *pseudo-level chance*. Penskoran pada tes objektif bersifat dikotomus sehingga tidak optimal untuk mengetahui kemampuan penempuh tes.

Ada banyak keuntungan yang didapat bila bentuk tes uraian disusun dalam bentuk tes *testlet*. Selain keunggulan-keunggulan bentuk tes objektif pada umumnya, tes dalam bentuk *testlet* juga mempunyai sistem penskoran yang bersifat politomus. Berdasarkan kajian teoretik dan empirik yang telah dilakukan ternyata dari segi penskoran, *testlet* lebih praktis dibanding bentuk uraian karena penskoran dapat dilakukan secara objektif dan bersifat politomus.

Bentuk tes *testlet* selain mempunyai kelebihan-kelebihan juga mempunyai kelemahan. Ada kelemahan penskoran *testlet* secara politomus yaitu menggunakan skor total sehingga kehilangan informasi yang berisi pola yang tepat dari respons penempuh tes, namun dengan pendekatan model *GRM* informasi kemampuan penempuh tes akan lebih dapat dijelaskan.

Dari segi konstruksi tes, bentuk *testlet* lebih membutuhkan keterampilan yang jauh lebih kompleks dalam menyusunnya terutama berkaitan dengan pemilihan alternatif jawaban pada setiap *item*. Seperti telah diungkap sebelumnya, penentuan alternatif jawaban pada bentuk tes objektif menjadi sangat penting karena berkaitan dengan tingkat peluang penempuh tes menjawab benar dengan menebak. Meningkatnya peluang penempuh tes menjawab benar dengan menebak akan semakin menurunkan nilai fungsi informasi *item* sehingga memperbesar kesalahan pengukuran yang terjadi.

Bila bentuk *testlet* akan digunakan sebagai alternatif bentuk tes yang dapat menggantikan bentuk tes uraian, perlu dikaji keefektifannya secara psikometrik terhadap bentuk tes uraian. Telah banyak penelitian yang menerapkan penskoran model *IRT* politomus pada *testlet*, namun sejauh ini belum memberikan informasi keefektifan bentuk tes *testlet* bila dibanding dengan bentuk tes uraian.

Metode Penelitian

Pengembangan instrumen dilakukan dengan beberapa tahap. Pertama, pengembangan instrumen uji coba yaitu instrumen yang terdiri dari 20 *item* berbentuk uraian terstruktur yang merupakan tes prestasi belajar kimia siswa SMA kelas XI yang diujikan pada siswa sebagai tes persiapan sebelum pelaksanaan ujian akhir semester I (Tes U). Dengan *item-item* tes yang sama, tes U disusun dalam bentuk *testlet* (Tes T) yaitu sebanyak 20 *testlet* masing masing *testlet* terdiri dari 3 *item* tes pilihan ganda. Tes prestasi belajar kimia ini dibuat oleh peneliti bersama-sama dengan guru-guru kimia di lingkungan Dinas Pendidikan, Pemuda dan Olahraga Kabupaten Tegal. Tes U dan Tes T diujicobakan pada siswa SMA kelas XII di lingkungan Dinas Pendidikan, Pemuda dan Olahraga Kabupaten Tegal.

Untuk memperoleh bukti validitas isi atau representasi isi digunakan penilaian pakar untuk naskah tes. Untuk membuktikan validitas konstrak dan asumsi keunidimensian terhadap tes yang dikembangkan, data yang diperoleh dari hasil uji coba dianalisis menggunakan *structural equation modeling* (*SEM*). Tes U diujicobakan kepada 253 siswa kelas XII Program IPA SMA yang diambil dari 2 SMA yaitu SMA 1 SLAWI dan SMA 3 SLAWI, Kabupaten Tegal.

Sedangkan Tes T di ujicobakan kepada 311 siswa kelas XII Program IPA dari dua SMA yang sama dengan Tes U.Model *SEM* yang digunakan untuk membuktikan validitas konstrak dan asumsi keunidimensian adalah model analisis faktor konfirmatori (AFK). Penskoran kedua instrumen menggunakan metode analitik (Mehrens & Lehmann, 1973: 229) dengan menggunakan empat kriteria yaitu skor 0, 1, 2, dan 3. Bila model yang digunakan adalah model *GRM*, *item* tes diselesaikan dengan tahapan-tahapan seperti yang dijelaskan berikut ini. Jawaban pada tahapan sebelumnya mempengaruhi pada tahap berikutnya, sehingga siswa yang menjawab benar pada tahapan pertama diberi skor 1. Apabila siswa menjawab benar pada tahapan kedua, dan tahapan pertama benar, diberi skor 2. Siswa yang dapat menjawab benar keseluruhan tahapan, diberi skor 3. Siswa yang menjawab tahapan kedua benar, tetapi tahapan pertama salah atau pada kedua tahapan siswa menjawab salah diberi skor 0.

Data penelitian terdiri dari skor tunggal untuk satu jawaban yang diperoleh berdasarkan pedoman penskoran yang telah ditetapkan. Pedoman penskoran (rubrik jawaban) dibuat berdasarkan langkah penyelesaian setiap *item* dan telah didiskusikan dengan tim (peneliti dan guru). *Rater* I dan *Rater* II masing-masing adalah guru kimia yang mengajar SMA kelas XI di Kabupaten Tegal. Ketika hasil analisis menunjukkan ada perbedaan antara skor siswa yang diskor oleh *rater* I dan *rater* II untuk setiap *item*, maka dilakukan pensekoran ulang dengan melibatkan peneliti sehingga semua *rater* mempunyai kesepahaman.

Parameter *item* diestimasi dengan menggunakan program komputer *PARSCALE VERSI 3,2* (Muraki & Bock, 1998). Kedua naskah tes dikalibrasi secara terpisah, dengan asumsi distribusi kemampuan dua kelompok penempuh tes adalah normal N(0,1). Untuk melihat keefektifan, bentuk *testlet* dibandingkan secara empirik dan simulasi dengan tes uraian dengan indikator fungsi informasi *item*. Hal ini dilakukan dengan membandingkan grafik fungsi informasi *item* antara kedua bentuk tes tersebut. Dalam membandingkan grafik fungsi informasi *item* antara kedua bentuk tes tersebut, digunakan nilai rerata fungsi. Kriteria yang digunakan untuk menguji kecocokan model GRM pada kedua bentuk tes yaitu tes bentuk uraian maupun tes dalam bentuk *testlet* adalah *root mean square differences* (*RMSD*) (Kim & Cohen, 2002). Besarnya *RMSD* dihitung secara terpisah

untuk setiap parameter, satu untuk parameter diskriminasi *item* dan satu untuk setiap parameter lokasi *item*.

Penelitian empirik dilakukan terhadap 772 siswa SMA kelas XI tersebar di lima SMA di Kabupaten Tegal masing-masing SMA I Slawi, SMA 2 Slawi, SMA 3 Slawi, SMA I Pangkah, dan SMA I Balapulang. Tes uraian (tes U) dan tes dalam bentuk *testlet* (tes T) sama-sama diberikan diberikan pada siswa dilakukan pada akhir semester I dan di awal semester II dengan waktu tenggang minimal 1 bulan. Pada tes pertama, di setiap kelas peserta dibuat menjadi dua kelompok secara acak, yaitu kelompok pertama mendapatkan tes U dan kelompok kedua mendapatkan tes T. Pada tes kedua, siswa yang mengerjakan tes U pada tes pertama mendapatkan tes T demikian sebaliknya siswa yang mengerjakan tes T pada tes pertama mendapatkan tes U. Demikanlah akhirnya didapat 772 respons siswa terhadap tes uraian dan 772 respons siswa terhadap tes dalam bentuk *testlet*. Pemilihan ukuran sampel didasarkan pada ukuran minimum penempuh tes yang direkomendasikan untuk estimasi parameter *item* secara empirik yaitu sebanyak 500 penempuh tes (Hambleton & Jones, 1994: 173).

Berdasarkan parameter estimasi yang diperoleh dari data empirik, selanjutnya proses simulasi dilakukan. Simulasi digunakan untuk membangkitkan data dengan berbagai kondisi seperti ukuran sampel dan banyaknya *item* sehingga didapatkan informasi yang lebih konsisten dan komprehensif berkaitan dengan perbedaan nilai fungsi informasi *item* pada bentuk uraian dan *testlet*. Replikasi data dilakukan sebanyak 25 kali. Hasil estimasi dari penelitian empirik pada Tes T digunakan untuk membangkitkan data secara simulasi yang dianggap sebagai respons penempuh tes pada bentuk *testlet* secara simulasi. Demikian pula hasil estimasi dari penelitian empirik pada Tes U digunakan untuk membangkitkan data secara simulasi yang dianggap sebagai respons penempuh tes pada bentuk tes uraian secara simulasi.

Ukuran minimum sampel yang dibutuhkan untuk estimasi parameter *item* sebanyak 500 penempuh tes (Hambleton & Jones, 1994: 173) sedangkan ukuran sampel yang dianggap cukup memuaskan untuk estimasi parameter *item* sebanyak 2000 penempuh tes (Lord, 1980: 258). Dalam penelitian ini, dua ukuran sampel digunakan yaitu 400 dan 2000 per kelompok penempuh tes. Hal ini dilakukan dengan mencontoh apa yang telah dipersyaratkan

untuk penyetaraan tes model GRM sedikitnya 500 penempuh tes (Reise & Yu, 1990) dan 400 penempuh tes (Bastari, 2000). Berkaitan dengan banyaknya *item*, digunakan 10 dan 20 *item* untuk tes uraian maupun tes dalam bentuk *testlet*. Berdasarkan kondisi bentuk tes, ukuran sampel dan banyaknya *item* didapat 8 kondisi dalam simulasi yang telah dilakukan. Data pada penelitian simulasi dibangkitkan dari parameter *item* hasil estimasi pada penelitian empirik dengan program *WinGen 2* (2007). Pada Bentuk tes uraian, data dibangkitkan dengan paramater hasil estimasi tes bentuk uraian secara empirik. Demikian pula untuk bentuk *testlet*, data dibangkitkan dengan paramater hasil estimasi tes bentuk *testlet* secara empirik. Data dibangkitkan berdasarkan ukuran sampel 400 dan 2000. Data kemampuan (*ability*) dari penempuh tes secara simulasi dibangkitkan berdistribusi normal baku. Selanjutnya dengan bantuan program *MAPPLE Versi* 9.5, dilakukan penentuan besarnya fungsi informasi *item* beserta kurvanya untuk rentang kemampuan penempuh tes (θ) sebesar -4 sampai 4. Untuk menilai ketepatan simulasi yang dilakukan dengan mencari *RMSD* dengan membandingkan parameter hasil estimasi secara empirik dan hasil estimasi secara simulasi.

Hasil Penelitian dan Pembahasan

Hasil Penelitian

Tes yang unidimensi merupakan asumsi yang harus dipenuhi bagi suatu perangkat tes sebelum digunakan untuk mengukur suatu kemampuan. Untuk membuktikan validitas konstrak dan asumsi keunidimensian suatu tes, data yang diperoleh dari hasil uji coba dianalisis menggunakan Model Persamaan Struktural (MPS). MPS yang digunakan adalah Analisis Faktor Konfirmatori (AFK).

Dukungan terhadap model yang dikembangkan dari data empirik (sampel), dapat dilihat dari besarnya nilai RMSEA (*root mean square error of approxmation*). RMSEA ini mengukur penyimpangan nilai parameter pada suatu model dengan matriks kovarian populasi (Imam Ghozali & Fuad, 2005: 31). Mac Callum et al (1996) menyatakan bahwa RMSEA berkisar

antara 0,08 sampai dengan 0,1 merupakan model yang memiliki fit cukup, bahkan untuk Tes U lebih baik dari Tes T, yaitu memiliki nilai RMSEA sebesar 0,083. Dukungan terhadap model yang lain dapat dilihat dari nilai NFI (*Normed Fit Index*) sebesar 0,90 (*cut- off* sebesar 0,9), NNFI (*Non – Normed Fit Index*) sebesar 0,92 (*cut- off* sebesar 0,9), CFI(Camparative Fit Index) sebesar 0,93 (*cut- off* sebesar 0,9) dan IFI (*Incremental Fit Index*) sebesar 0,93 (*cut- off* sebesar 0,9), RFI (*Relative Fit Index*) sebesar 0,90 (mendekati 1), PGFI (*Parsimony Goodnes Of Fit Index*) sebesar 0, 64 (*cut- off* sebesar 0, 6)(Imam Ghozali & Fuad, 2005: 29-33).

Pada Tes T memiliki nilai RMSEA sebesar 0,090. Dukungan terhadap model yang lain dapat dilihat dari nilai NFI (*Normed Fit Index*) sebesar 0,92 (*cut- off* sebesar 0,9), NNFI (*Non – Normed Fit Index*) sebesar 0,92 (*cut- off* sebesar 0,9), CFI(*Comparative Fit Index*) sebesar 0,93 (*cut- off* sebesar 0,9) dan IFI (*IncrementalFit Index*) sebesar 0,93 (*cut- off* sebesar 0,9), RFI (*Relative Fit Index*) sebesar 0,90 (mendekati 1), PGFI (*Parsimony Goodnes Of Fit Index*) sebesar 0,62 (*cut- off* sebesar 0, 6)(Imam Ghozali & Fuad, 2005: 29-33). Berdasarkan hasil tersebut, validitas konstrak Tes T dan Tes U terbukti kebenaranya dan asumsi keunidimensian tes dipenuhi dalam penelitian ini.

Untuk parameter *item* tes dalam uji coba dalam bentuk *testlet* menunjukkan bahwa *item* 1 mempunyai tingkat kesukaran rendah, *item* nomor 2, 3, 4, 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 mempunyai tingkat kesukaran sedang dan *item* nomor 6, 7, 8, 9 mempunyai tingkat kesukaran tinggi. Bila dilihat dari parameter *slope* dan mengacu pada Emons, Meijer & Sijtsma (2002: 93), tampak bahwa *item-item* pada umumnya mempunyai indeks diskriminasi sedang kecuali *item* nomor 6, 7, 8, 9, 10, 11 mempunyai indeks diskriminasi rendah dan *item* nomor 16, 17, 18, 19, 20 mempunyai indeks diskriminasi sedang.

Untuk parameter *item* tes uji coba dalam bentuk uraian menunjukkan bahwa berdasarkan parameter *location,* ke 20 *item* tersebut menunjukkan bahwa *item* 1, 2, 3, 4, 5 mempunyai tingkat kesukaran mudah, *item* 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 mempunyai tingkat kesukaran sedang dan *item* nomor 6, 7, 8, 9 mempunyai tingkat kesukaran tinggi. Bila dilihat dari parameter *slope* dan mengacu pada Emons, Meijer & Sijtsma (2002: 93),

tampak bahwa *item-item* pada umumnya mempunyai indeks diskriminasi rendah kecuali *item* nomor 11, 12, 13, 14, 15 mempunyai indeks diskriminasi tinggi dan *item* nomor 20 mempunyai indeks diskriminasi sedang.

Data dari kedua naskah Tes U dan Tes T, masing-masing diestimasi untuk memperoleh parameter *item* tes dengan menggunakan program *parscale* versi 3,20 dengan kondisi distribusi kemampuan kedua kelompok adalah normal (N(0,1)). Estimasi parameter dilakukan terhadap 20 *item* tes pada bentuk tes uraian dan bentuk *testlet* dengan model *GRM*. Selanjutnya dari parameter empirik yang didapat, digunakan untuk menentukan besarnya fungsi informasi item dengan bantuan program *mapple* versi 9,5.

Tabel 1 menunjukkan reliabilitas tes pada bentuk tes uraian dan bentuk *testlet* sedangkan Tabel 2 dan Tabel 3 menunjukkan parameter dan besarnya fungsi informasi *item* pada bentuk tes uraian dan bentuk *testlet*. Tabel 4 menunjukkan perbedaan nilai I pada uraian dan I pada *testlet.* Gambar 1 hingga Gambar 4 menujukkan contoh kurva fungsi informasi *item* hasil komputasi dengan program *mapple versi 9,5*.

Tabel 1. Reliabilitas Tes Bentuk Uraian dan Bentuk *Testlet*

Bentuk tes	Reliabilitas	
	Alpha	Spearman-Brown
Uraian	0,9547	0,9659
Testlet	0,8255	0,7729

Tabel 2. Karakteristik Parameter *Item* pada Tes Bentuk Uraian dengan Model *GRM*

No *item*	b1	b2	b3	a	I
1	-1,973	-1,847	1,537	3,382	1,276
2	-2,603	-1,448	-0,752	0,56	0,160
3	-0,015	0,218	1,808	3,159	1,161
4	1,201	1,368	2,03	2,385	0,986
5	-1,977	-1,633	-0,551	2,665	1,371
6	-1,577	-1,096	0,504	4,186	2,516

No *item*	b1	b2	b3	a	I
7	-1,484	-0,405	0,248	4,014	2,482
8	-2,224	-1,851	-1,394	1,76	0,666
9	-2,242	0,436	1,46	2,126	1,290
10	-1,786	-1,399	-0,148	0,946	0,355
11	-1,738	-1,12	-0,574	5,499	3,412
12	-1,071	-0,426	1,062	3,397	2,063
13	-1,565	-1,431	0,346	6,787	3,573
14	-2,315	-0,838	2,526	0,526	0,201
15	-1,853	-0,917	-0,557	2,212	1,065
16	-2,034	-0,254	3,02	1,866	1,158
17	-1,908	0,119	0,847	2,652	1,584
18	-1,996	-0,572	0,189	2,482	1,465
19	-0,359	0,575	1,806	1,899	1,073
20	-1,937	-0,098	1,085	3,693	2,348
I tes					31,205

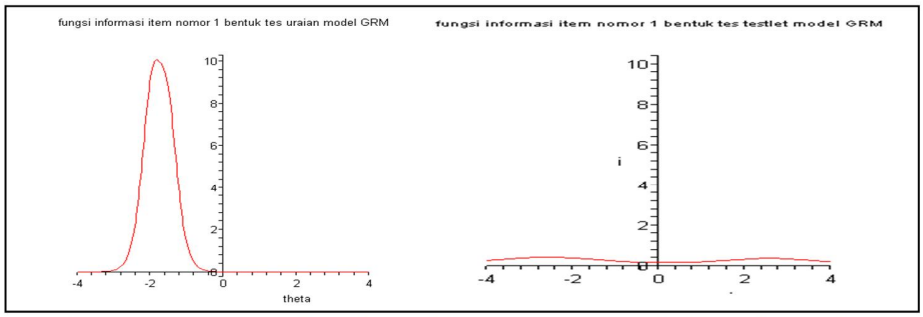
Tabel 3. Karakteristik Parameter *Item* pada Tes Bentuk *Testlet* dengan Model *GRM*

No Item	b1	b2	b3	a	I
1	-2,933	-2,276	2,578	0,697	0,273
2	-3,109	1,705	3,234	0,656	0,266
3	-2,102	-0,2	1,529	0,569	0,223
4	-0,173	2,108	3,736	0,423	0,121
5	-0,763	1,048	1,707	0,603	0,212
6	-4,96	-3,469	2,249	0,326	0,078
7	-2,207	0,722	2,256	1,068	0,594
8	-0,91	1,558	1,869	1,447	0,683
9	-1,448	-0,349	2,301	0,921	0,457
10	-1,855	1,459	4,424	0,278	0,064
11	-1,99	-1,438	1,277	1,111	0,533
12	-2,039	0,227	1,987	1,197	0,696
13	-3,384	-1,99	0,226	0,27	0,054
14	-2,124	1,841	4,216	0,307	0,076

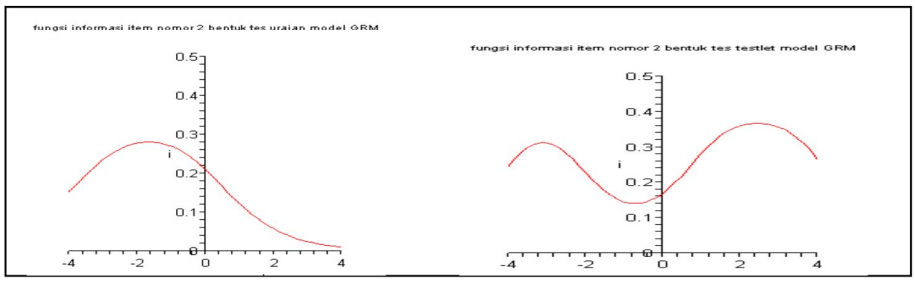
No Item	b1	b2	b3	a	I
15	-2,795	2,633	3,144	0,567	0,189
16	-0,574	2,309	3,66	0,575	0,208
17	0,215	1,804	2,496	1,035	0,454
18	0,337	1,613	1,929	1,486	0,650
19	-1,762	-1,353	0,649	0,807	0,320
20	-1,935	0,831	2,245	0,44	0,145
I (tes)					6,296

Tabel 4. Perbedaan Nilai I pada Uraian dan I pada *Testlet*

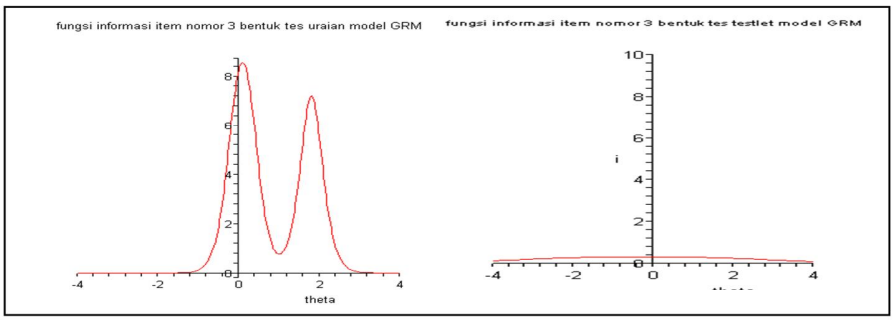
No Item	I(uraian)	I(*testlet*)	Perbedaan I
1	1,276	0,273	1,003
2	0,160	0,266	-0,160
3	1,161	0,223	0,938
4	0,986	0,121	0,865
5	1,371	0,212	1,159
6	2,516	0,078	2,438
7	2,482	0,594	1,888
8	0,666	0,683	-0,017
9	1,290	0,457	0,833
10	0,355	0,064	0,291
11	3,412	0,533	2,879
12	2,063	0,696	1,367
13	3,573	0,054	3,519
14	0,201	0,076	0,125
15	1,065	0,189	0,876
16	1,158	0,208	0,950
17	1,584	0,454	1,130
18	1,465	0,650	0,815
19	1,073	0,320	0,753
20	2,348	0,145	2,203



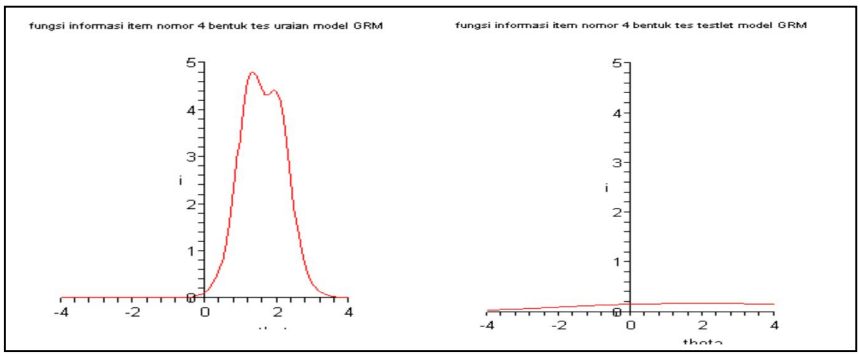
Gambar 1. Kurva Fungsi Informasi *Item* Nomor 1 pada Bentuk Tes Uraian dan Bentuk *Testlet* secara Empirik Model *GRM*



Gambar 2. Kurva Fungsi Informasi *Item* Nomor 2 pada Bentuk Tes Uraian dan Bentuk *Testlet* secara Empirik Model *GRM*



Gambar 3. Kurva Fungsi Informasi *Item* Nomor 3 pada Bentuk Tes Uraian dan Bentuk *Testlet* secara Empirik Model *GRM*



Gambar 4. Kurva Fungsi Informasi *Item* Nomor 4 pada Bentuk Tes Uraian dan Bentuk *Testlet* secara Empirik Model *GRM*

Simulasi digunakan untuk membangkitkan data dengan berbagai kondisi seperti ukuran sampel dan banyaknya *item* sehingga didapatkan informasi yang lebih komprehensif berkaitan dengan perbandingan nilai fungsi informasi *item* pada tes dalam uraian dan tes dalam bentuk *testlet*. Ada 8 kondisi pada simulasi yang dilakukan yaitu: (1) tes dalam bentuk uraian dengan 10 *item* dan 2000 penempuh tes (U21), (2) tes dalam bentuk uraian dengan 20 *item* dan 2000 penempuh tes (U22), (3) tes dalam bentuk uraian dengan 10 *item* dan 400 penempuh tes (U41), (4) tes dalam bentuk uraian dengan 20 *item* dan 400 penempuh tes (U42), (5) tes dalam bentuk *testlet* dengan 10 *item* dan 2000 penempuh tes (T21), (6) tes dalam bentuk *testlet* dengan 20 *item* dan 2000 penempuh tes (T22), (7) tes dalam bentuk *testlet* dengan 10 *item* dan 400 penempuh tes (T41) dan (8) tes dalam bentuk *testlet* dengan 20 *item* dan 400 penempuh tes (T42). Untuk melihat perbandingan fungsi informasi tes pada empirik dan simulasi dapat dilihat pada Tabel 5 sedangkan untuk melihat ketepatan simulasi yang dilakukan dengan melihat nilai RMSD seperti yang terlihat pada Tabel 6.

Tabel 5. Fungsi Informasi Tes pada Tes Secara Empirik dan Simulasi

Bentuk Tes	Empirik	10 *item* Ukuran sampel=2000	20 *item* Ukuran sampel=2000	10 *item* Ukuran sampel=400	20 *item* Ukuran sampel=400
uraian	31,205	18,321	27,681	17,423	23,761
testlet	6,296	3,303	6,810	3,391	5,755

Tabel 6. Nilai *RMSD* a dan Nilai *RMSD* b dari Tiap Kondisi Simulasi

RMSD	T21	T22	T41	T42	U21	U22	U41	U42
a	0,041	0,368	0,110	0,152	0,518	0,422	0,732	0,754
b1	0,522	0,175	0,469	0,753	0,188	0,113	0,294	0,736
b2	0,514	0,131	0,322	0,398	0,128	0,133	0,190	0,510
b3	0,518	0,170	0,255	0,496	0,057	0,212	0,105	0,596
rerata	0,399	0,212	0,290	0,45	0,223	0,221	0,331	0,614

Pembahasan

Hasil analisis reliabilitas menunjukkan siswa yang mengerjakan bentuk uraian lebih konsisten dalam menjawab setiap *item* dibanding siswa yang mengerjakan bentuk *testlet*. Hal ini ditunjukkan oleh dua hal yaitu: (1) estimasi reliabilitas dengan *Spearman Brown* (*SB)* maupun Apha untuk bentuk uraian mempunyai nilai yang lebih tinggi dibandingkan pada bentuk *testlet*, (2) estimasi dengan *SB* untuk bentuk *testlet* menghasilkan nilai reliabilitas yang lebih rendah dibandingkan estimasi dengan Alpha, hal ini memberikan informasi bahwa jika tes dalam bentuk *testlet* hanya dibagi 2 subtes mempunyai korelasi yang rendah. Hal ini disebabkan reliabilitas dengan formula Alpha berdasarkan rasio varian *item* terhadap varian skor total sedangkan formula *SB* berdasarkan korelasi skor total dua subtes yaitu nomor-nomor gasal dan nomor-nomor genap.

Bila konsep reliabilitas tes secara klasik dapat didekati dengan konsep fungsi informasi tes dalam *Item Response Theory* (*IRT*), maka informasi yang didapat pada Tabel 4 memberikan informasi bahwa baik secara empirik

maupun simulasi, fungsi informasi tes pada tes yang disajikan dalam bentuk uraian lebih tinggi bila dibandingkan tes disajikan dalam bentuk *testlet*. Bahkan secara empirik maupun simulasi, nilai fungsi informasi tes dalam bentuk uraian mempunyai nilai sebesar 4 - 5 kali dari nilai fungsi informasi tes dalam bentuk *testlet*. Hal ini dapat disimpulkan secara empirik maupun simulasi tes bentuk uraian lebih efektif dibandingkan tes bentuk *testlet*.

Hasil penelitian secara empirik menunukkan hanya ada 2 *item* dari 20 *item* yang disajikan dalam bentuk uraian memiliki nilai fungsi informasi lebih rendah dibanding dalam bentuk *testlet*. Secara empirik ada perbedaan sebesar 0,160 pada *item* nomor 2 dan sebesar 0,01 pada *item* nomor 8, dimana nilai fungsi informasi *item* lebih besar dalam bentuk *testlet* dibanding dalam bentuk uraian. Secara empirik kecuali *item* nomor 2, semua *item* yang disajikan dalam bentuk uraian memiliki daya beda yang lebih tinggi bila disajikan dalam bentuk *testlet*. Hal ini membuktikan nilai daya beda *item* sangat nyata pengaruhnya terhadaap besarnya fungsi informasi *item*.

Hasil penelitian secara simulasi, menunjukkan: (1) semua simulasi dengan kondisi10 *item* menunjukkan semua *item* yang disajikan dalam bentuk uraian memiliki nilai fungsi informasi *item* yang lebih besar dibanding disajikan dalam bentuk *testlet*, (2) Pada kondisi 20 *item*, untuk ukuran sampel 2000, *item* nomor 2 dan 8 memiliki fungsi informasi *item* yang lebih besar bila disajikan dalam bentuk *testlet* bila dibandingkan dengan uraian, (3) Pada kondisi 20 *item*, untuk ukuran sampel 400, *item* nomor 2, 8 dan 11 memiliki fungsi informasi *item* yang lebih besar bila disajikan dalam bentuk *testlet* bila dibandingkan dengan uraian.

Hasil penelitian secara simulasi menunjukkan bahwa pada semua kondisi banyaknya *item* dan ukuran sampel, *item-item* yang disusun dalam bentuk uraian cenderung memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding *item-item* dalam bentuk *testlet.* Demikian pula secara simulasi menunjukkan bahwa ada pengaruh banyaknya *item* dan ukuran sampel terhadap kecenderungan perbedaan nilai fungsi informasi *item* pada bentuk tes uraian dan bentuk *testlet*. Semakin sedikit *item*, semakin meningkat kecenderungan *item-item* dalam bentuk uraian memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding *item-item* dalam bentuk *testlet*. Semakin besar ukuran sampel, semakin meningkat kecenderungan

item-item dalam bentuk uraian memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding *item-item* dalam bentuk *testlet*. Secara simulasi, dari rerata nilai *RMSD* untuk nilai a dan b, maka bentuk T22 dan U22 memiliki nilai rerata terkecil. Oleh sebab itu, kondisi tes dengan banyaknya *item* 20 dan ukuran sampel 2000 dapat dijadikan dasar untuk mengetahui perbedaan nilai rerata fungsi informasi *item* antara *item* yang dibentuk dalam bentuk tes uraian dan dalam bentuk *testlet*. Seperti halnya pada kajian empirik, pada simulasi juga nampak bahwa *item* nomor 2 dan nomor 8 berbeda dengan kecenderungan *item-item* yang lain. Bila 18 *item* yang lain konsisten dalam hal nilai rerata fungsi informasi *item* pada bentuk tes uraian lebih besar daripada nilai rerata fungsi informasi item pada bentuk *testlet*, maka *item* nomor 2 dan 8 berlaku sebaliknya.

Dapat disimpulkan bahwa fungsi informasi tes pada tes yang terdiri dari *item-item* dalam bentuk uraian cenderung lebih tinggi daripada fungsi informasi tes pada tes yang terdiri dari *item-item* dalam bentuk *testlet*. Hal ini disebabkan fungsi informasi tes adalah akumulasi dari nilai fungsi informasi item yang menyusun tes tersebut.

Sebagai salah satu jenis karakteristik *item*, tentunya sangat diinginkan nilai fungsi informasi tes maksimal seperti halnya tes yang dianalisis dengan teori tes klasik menginginkan nilai reliabilitas yang tinggi, walaupaun secara konseptual ada perbedaan antara reliabilitas pada konsep teori tes klasik dan fungsi informasi tes secara IRT. Di dalam teori tes klasik, skor *item* yang membentuk koefisien reliabilitas tes tidaklah independen satu dan lainya. Perubahan satu *item* saja akan mengubah semua nilai pada koefisien reliabilitas. Hal ini tidak terjadi pada fungsi informasi tes. Pada *IRT*, *item-item* bersifat independen satu dengan yang lain sehingga perubahan suatu *item* hanya mengubah pada fungsi informasi tes dan tidak mengubah nilai fungsi informasi *item* dari *item-item* yang lain. Menurut Hambleton dan Swaminathan(1985: 236), pengukuran fungsi informasi tes lebih akurat bila dibandingkan dengan penggunaan reliabilitas karena: (1) bentuknya tergantung hanya pada *item-item* dalam tes, (2) mempunyai estimasi kesalahan pengukuran pada setiap level kemampuan.

Fungsi informasi pada IRT berhubungan secara terbalik dengan ketidakpastian. Ini berarti bahwa makin tinggi ketidakpastian, maka makin

rendah nilai fungsi informasi tes. Sebaliknya, makin rendah ketidakpastian, maka makin tinggi tinggi nilai fungsi informasi tes. Dengan demikian bentuk tes uraian akan lebih kaya informasi berkaitan dengan kemampuan penempuh tes dibanding dengan bentuk *testlet* yang merupakan kumpulan dari bentuk pilihan ganda. Hal ini dapat dipahami karena jika tes dimaksudkan untuk melihat pola pikir yang kompleks dari penempuh tes, maka bentuk tes uraian lebih tepat dibandingkan dengan bentuk tes pilihan ganda. Bentuk tes uraian, memberikan kebebasan kepada setiap penempuh tes untuk mengekspresikan daya nalarnya, sehingga jawaban yang diberikan oleh setiap peserta akan menunjukkan kemampuan berpikir secara kompleks.

Melalui studinya, Zidner (1987: 607) menyimpulan bahwa pada bentuk tes uraian, membutuhkan kemampuan yang tinggi untuk mengorganisasi jawaban, membutuhkan kemampuan mengingat kembali terhadap materi, membutuhkan pengetahuan yang integratif dan kemampuan menulis dengan baik. Pada tes pilihan ganda tidak didapatkan hal seperti itu, oleh karena peserta tinggal memilih opsi yang telah disiapkan. Bila seorang peserta menjawab benar untuk *item* yang sama pada tipe tes pilihan ganda, sangat sukar untuk menduga bahwa pilihanya didasarkan pada hasil berpikir yang kompleks.

Rendahnya nilai fungsi informasi *item* pada bentuk pilihan ganda disebabkan pula karena adanya peluang menebak dalam menjawab. Hal ini meningkatkan kesempatan penempuh tes dalam menebak jawaban. Semakin tingginya tebakan penempuh tes ini tentunya akan memperlemah daya beda *item*. Hal ini diperparah lagi dengan kenyataan bahwa peluang bekerja sama antara penempuh tes pada tes objektif sangatlah tinggi. Semakin rendah daya beda *item* akan menyebabkan semakin homogennya skor yang diperoleh, dan semakin homogennya skor yang diperoleh penempuh tes akan menurunkan nilai fungsi informasi *item*.

Hasil penelitian terdahulu yang membandingkan bentuk tes uraian dan bentuk pilihan ganda juga memperkuat hasil penelitian dimana ada kecenderungan tes bentuk uraian lebih efektif dibanding tes dalam bentuk *testlet*. Hasil penelitian yang dilakukan oleh Kuechler dan Simkin, (2003: 394) menyimpulkan bahwa bentuk tes pilihan ganda memberikan

kesempatan siswa menebak jawaban benar lebih besar dibanding pada bentuk tes uraian. Shepard, (2008: 604), melalui kajian yang dilakukan oleh *National Mathematics Advisory Panel* terhadap lebih dari 15 penelitian, menyimpulkan bahwa: (1) adanya kesalahan bila bentuk pilihan ganda dan bentuk uraian digunakan untuk mengukur kompetensi yang sama, (2) bentuk tes uraian lebih baik digunakan untuk mengukur kemampuan siswa yang lebih tinggi, dan (3) bentuk tes uraian memiliki informasi yang lebih tinggi dibandingkan bentuk pilihan ganda.

Simpulan

Hasil penelitian menunjukkan bahwa: (1) secara empirik dan simulasi, tes yang disajikan dalam bentuk uraian cenderung memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding dengan tes yang disajikan dalam bentuk *testlet.* Dengan demikian dapat disimpulkan bentuk tes uraian cenderung lebih efektif dibandingkan bentuk tes *testlet,* (2) secara simulasi, ada pengaruh banyaknya *item* terhadap perbandingan nilai fungsi informasi *item p*ada bentuk tes uraian *d*an bentuk *testlet.* Bila tes menggunakan *item* yang lebih sedikit, ada kecenderungan bentuk tes uraian memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding dengan tes yang disajikan dalam bentuk *testlet, (*3) secara simulasi, ada pengaruh ukuran sampel terhadap perbandingan nilai fungsi informasi *item p*ada bentuk tes uraian *d*an *testlet.* Bila tes menggunakan ukuran sampel yang lebih besar, ada kecenderungan bentuk tes uraian memiliki nilai fungsi informasi *item* yang lebih tinggi dibanding dengan tes yang disajikan dalam bentuk *testlet,* dan (4) pemodelan GRM pada bentuk tes uraian dan *testlet* paling akurat pada kondisi banyaknya *item* 20 dan ukuran sampel 2000.

Daftar Pustaka

Bastari (2000). *Linking multiple-choice and construct-response items to a common proficiency scale*. Disertasi tidak diterbitkan.University of Massachusetts, Amherst.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Dalam F. M. Lord & M.R., Novick (Eds.). *Statistical theories of mental test score* (Bab *17-20).* Reading, MA: Addison- Wesley.

De Ayala, R., J. (1993). An introduction to polytomous item respons theory models. *Measurement and Evaluation in Counseling and Development* , 25, 172-189.

Ebel, R., L. & Frisbie, D., A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall, Inc.

Embretson, S., E. & Reise, S., P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrance Erlbaum Associates.

Emons, W.H.M, Meijer, R., R & Sijtsma, K.(2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement*, 26, 88-108.

Feldt, L., S & Chorter, R., A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Method,* 8 (1), 102-109.

Gronlund, N., E. (1990). *Measurement and evaluation in teaching.* New York: Macmillan.

Hambleton, R., K. (1989). Principles and selected applications of item response theory. Dalam R.L. Linn (Ed.). *Educational Measurement hal. 147-200.* UK: Macmillan..

Hambleton, R., K. & Jones, R., W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, 7(3), 171-186.

Hambleton, R., K. & Swaminathan, H. (1985). *Item response theory.* Boston, MA: Kluwer.

Han, K., T. & Hambleton, R., K. (2007). *User's manual for WinGen*. Amherst, MA: University of Masssachusetts Amherst.

Imam Ghozali & Fuad. (2005). *Structural equation modeling*, Semarang: Badan Penerbit Universitas Diponegoro.

Johnson, D., W. & Johnson, R., T. (2002). *Meaningful assessment*. Boston, MA: A Pearson Education Company.

Joreskog, K., G. & Sorbom, D., Toit, S., & Toit, M. (2000). LISREL 8: *New statistical features*. Chicago, IL:SSI, Inc.

Kaufman, R., & Thomas, S. (1980). *Evaluation without fear.* New York: New Viewpoints.

Kim, S. & Cohen, A. (2002). A comparison of linking and concurrent calibrated under the graded response models. *Applied Psychological Measurement*, 26(1), 25-41.

Muraki, E. & Bock, R., D. (1997). *Parscale*: *IRT item analysis and test scoring for rating- scale data*. Chicago: Scientific Software International, Inc.

Nonny Swediati. (1997). *Equating tests under the generalized partial credit model*. Disertasi tidak dipublikasikan. University of Massachussets at Amherst.

Reise, S.P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 145-163.

Setyo H. Wijayanto. (2002). *Structural equation modeling & lisrel 8.51 for window*. Tidak diterbitkan. Jurusan akuntansi, Fakultas Ekonomi UI.

Shepard, L.A. (2008). Commentary on the national mathematics advisory panel recommendations on assessment. *Educational Reserarcher*, 37 (9), 602-609.

Thissen, D. & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Zedner, M. (1987). Essay versus multiple-choice type classroom exam: the student perspective. *Journal of Educational Research*, 80 (6), 352-358.