KOMPARASI KARAKTERISTIK BUTIR TES PILIHAN GANDA DITINJAU DARI TEORI TES KLASIK

Imam Suseno
Dosen Metodologi Penelitian pada FBS
Universitas Indraprasta PGRI Jakarta

Abstract: This study compares the characteristics of test items (item difficulty level, different power and reliability) test Multiple Choice Association (PGA), multiple choice tests Ordinary (PGB) and test Multiple Choice relationship between things (PGH) in descriptive statistics. Using 30 items of each type of test show the reliability coefficient of the most consistent PGH test measures the ability of students, as well as the level of hard grain and different power index included in the criteria of a good test. So that both types decent test PGA and PGH are used to measure students' abilities in addition to the type of test used PGB. Despite the complexity of the process of answering each item in the test, but the test proved to be the type of PGA and PGH had a good enough criterion and decent used to measure students' abilities in understanding various subjects.

Keywords: The level of difficult Item, the diffrence Power, and Reliability Tests

Abstrak: Penelitian ini membandingkan karakteristik butir tes (taraf sukar butir, daya beda dan reliabilitas) tes Pilihan Ganda Asosiasi (PGA), tes Pilihan Ganda Biasa (PGB) dan tes Pilihan Ganda Hubungan antara hal (PGH) secara deskriptif statistik. Menggunakan 30 butir soal tiap tipe tes menunjukkan koefisien reliabilitas tes PGH paling konsisten mengukur kemampuan siswa, demikian pula taraf sukar butir dan indeks daya beda termasuk dalam kriteria tes yang baik. Sehingga kedua tipe tes PGA dan PGH layak digunakan untuk mengukur kemampuan siswa disamping tes tipe PGB yang biasa digunakan. Meskipun memiliki kompleksitas dalam proses menjawab tiap butir tes, namun terbukti tipe tes PGA dan PGH memiliki kriteria yang cukup baik dan layak digunakan untuk mengukur kemampuan siswa pada mata pelajaran berjenis pemahaman.

Kata Kunci: *Taraf Sukar Butir, Daya Beda, dan Reliabilitas Tes.*

PENDAHULUAN

Permendikbud No.23 Tahun 2016 tentang Standar Penilaian pendidikan Pasal 14 **point pertama** disebutkan bahwa instrumen penilaian yang digunakan oleh pendidik dalam bentuk penilaian berupa tes, pengamatan, penugasan perseorangan atau kelompok, dan bentuk lain yang sesuai dengan karakteristik kompetensi dan tingkat perkembangan peserta didik. Pada **point kedua** disebutkan bahwa instrumen penilaian yang digunakan oleh satuan pendidikan dalam bentuk penilaian akhir dan atau ujian sekolah/madrasah memenuhi persyaratan substansi, konstruksi, dan bahasa, serta memiliki bukti validitas emipirik.

Karakteristik kompetensi pada siswa tingkat SMA meliputi pengetahuan faktual, konseptual, prosedural dan matakognitif. Kompetensi metakognitif merupakan pengetahuan tentang kekuatan dan kelemahan diri sendiri dan menggunakannya dalam mempelajari pengetahuan teknis, detail, spesifik, kompleks, konseptual dan

kondisional berkenaan dengan ilmu pengetahuan, teknologi, seni dan budaya terkait dengan masyarakat dan lingkungan alam sekitar, bangsa, negara, kawasan regional dan internasional.

Sesuai dengan SKL (Standar Kompetensi Lulusan), sasaran pembelajaran tingkat SMA sederajat mencakup pengembangan ranah sikap, pengetahuan, dan keterampilan yang dielaborasi untuk setiap satuan pendidikan. Pada ranah pengetahuan diperoleh melalui aktivitas mengingat, memahami, menerapkan, menganalisis, mengevaluasi, dan mencipta. Evaluasi hasil pembelajaran dilakukan saat proses pembelajaran dan diakhir satuan pembelajaran dengan menggunakan metode dan alat: tes lisan/perbuatan, dan tes tertulis. Salah satu bentuk tes tertulis adalah adalah tes pilihan ganda.

Penggunaan instrumen tes pilihan ganda (PG) dapat ditemui pada ujian skala besar (misal Ujian Nasional) maupun ujian berskala kecil (misal Ujian Formatif, Ujian Sumatif), karena sifatnya objektif dan mudah dalam pensekoran. Groundlund and Linn (1990: 168) bahwa soal tes PG dapat digunakan mengukur hasil belajar atau kemampuan siswa dalam berfikir sederhana sampai berfikir kompleks dan hal tersebut disesuaikan dengan materi pembahasan.

Kekuatan tes PG adalah memuat beberapa jawaban yang berbeda yang saling berhubungan, namun perbedaan dibuat hampir tidak kentara diantara pilihan jawaban namun beberapa yang mungkin menjadi sedikit benar. Sudjono (2013: 119-120) menyebutkan tes pilihan PG dibedakan atas: (1) melengkapi lima pilihan, (2) asosiasi dengan lima pilihan atau empat pilihan, (3) hal kecuali, (4) analisis hubungan antar hal, (5) analisis khusus, (6) perbandingan kuantitatif, (7) hubungan dinamik, (8) melengkapi berganda, (9) pemakaian diagram, gambar dan grafik.

Sehingga dalam melakukan evaluasi kemampuan siswa, guru diharapkan tidak hanya terpaku pada jenis tes PG biasa saja namun lebih bervariasi dalam menggunakan instrumen tes hasil belajar khususnya pada bentuk tes PG. Namun penggunaan berbagai bentuk tes PG perlu mempertimbangkan beberapa hal.

Hasil penelitian Hajaroh (2011: 141) menyimpulkan bahwa beragamnya bentuk soal objektif, maka tidak bisa memberikan *jagement* bahwa soal pilihan ganda (dibandingkan bentuk *matching tes*t) memiliki tingkat kesukaran yang lebih tinggi. Meskipun pada bentuk tes pilihan ganda terdapat dua parameter yang menyebabkan tingginya kesukaran yaitu adanya pengecoh (*distractor*) dan adanya peluang responden untuk menebak jawaban. Selanjutnya daya beda tes pilihan ganda tidak lebih tinggi dibandingkan bentuk *matching* tes, artinya beragamnya bentuk tes objektif yang ada tes bentuk PG bukanlah satu-satunya bentuk tes yang lebih baik dalam membedakan siswa yang berkemampuan tinggi dan siswa yang berkemampuan rendah. Hal ini dapat memberikan masukan bagi para pendidik bahwasanya sebelum membuat tes perlu adanya analisis, sebab untuk menentukan bentuk tes hendanya disesuaikan dengan karakteristik dari bidang studi yang akan diujikan.

Terdapat hasil penelitian yang menyebutkan bahwa item pilihan ganda memiliki semua persyaratan sebagai tes yang baik, yakni dilihat dari validitas, reliabilitas dan daya pembeda anatar siswa yang berhasil dengan siswa yang tidak berhasil, tidak semuanya benar jadi selum memberikan keputusan dalam membuat soal hendanya melakukan penyelidikan terlebih dahalu terhadp butir-butir soal yang akan diujikan .

Hajaroh (2011 : 142) keberhasilan proses belajar mengajar tidak dapat dipantau tanpa adanya evaluasi hasil belajar. Akan tetapi evaluasi yang baik akan mungkin jika alat evaluasinya juga baik, maka guru dituntut untuk menguasai cara dan kaidah dalam menyusun tes yang baik. Untuk mengetahui apakah alat evaluasi tersebut baik atau tidak, maka perlu adanya analisis butir. Melalui analisis butir soal guru akan mendapatkan informasi untuk memberikan umpan balik baik kepada siswa maupun pendidik itu sendiri

hasil dan dapat dipakai untuk mengupayakan butir soal tersebut. Dalam memilih bentuk soal, pendidik hendaknya menyesuaikan dengan karakteristik bidang dan mengetahui kualitas butir soal

Karmel and Karmel (1978: 406) terdapat sepuluh kriteria tes yang baik, yakni: (1) tes harus relevan, (2) ada keseimbangan antara tujuan yang ingin dicapai dengan jumlah butir tes yang mewakilinya, (3) efisiensi waktu yang digunakan untuk melakukan tes, pensekoran dan pengadministrasian sekor tes, (4) objektivitas dalam memberikan skor dan interpretasinya, (5) kekhususan tes yang mengukur materi pelajaran yang diajarkan dikelas, (6) tingkat kesukaran setiap butir tes, (7) kemampuan butir membedakan kelompok siswa yang memiliki kemampuan tinggi dan rendah, (8) reliabilitas, (9) kejujuran dan pemerataan kesempatan, dan (10) kecepatan menyelesaikan tes.

Dalam penelitian ini bertujuan melihat perbedaan taraf sukar butir, daya beda dan reliabilitas dari ketiga bentuk tes PG diantaranya tes PGB (pilihan Ganda Biasa), Tes PGA (Pilihan Ganda Asosiasi) dan Tes PGH (Pilihan Ganda Hubungan antar hal).

Teori Tes Klasik (*Classical Test Theory*)

Teori tes klasik sering disebut teori skor murni klasik (Allen & Yen, 1979:57) yaitu skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran. Secara matematis adalah

$$X = T + E \quad \text{..... (1)}$$

Dengan X adalah skor amatan, T adalah skor sebenarnya dan E adalah skor kesalahan pengukuran (*error score*). *Error score* yang dimaksud merupakan kesalahan acak atau tidak sistematis. Kesalahan ini merupakan penyimpangan secara teoritis dari skor amatan yang diperoleh dengan skor amatan yang diharapkan. Sedangkan kesalahan pengukuran sistematis dianggap bukan merupakan kesalahan pengukuran.

Asumsi yang menyertai dalam teori tes klasik antara lain: 1) skor kesalahan pengukuran tidak berinteraksi dengan skor sebenarnya, 2) skor kesalahan tidak berkorelasi dengan skor sebenarnya dan skor-skor kesalahan pada tes-tes yang lain untuk peserta tes (*testee*) yang sama, 3) rata-rata dari skor kesalahan ini sama dengan nol.

Ketiga asumsi pada teori tes klasik ini dijadikan dasar untuk mengembangkan formula-formula dalam menentukan validitas dan reliabilitas tes. Validitas dan reliabilitas pada perangkat tes digunakan untuk menentukan kualitas tes selain dari indeks kesukaran dan daya beda butir.

Taraf Sukar Butir

Taraf sukar butir berkaitan dengan responden. Butir tes bisa dirasakan sukar oleh suatu responden namun dirasakan tidak sukar oleh responden lain. Karena itu taraf sukar butir dan kelompok responden saling berkait (Dali, 2013: 280). Taraf sukar butir dapat berbentuk proporsi (P_i), berbentuk distribusi probabilitas normal baku (z), berbentuk delta (Δ). Untuk menentukan indeks kesukaran dari suatu butir pada perangkat tes pilihan ganda, digunakan persamaan sebagai berikut:

$$P_i = \frac{banyaknya\ jawaban\ betul}{banyaknya\ seluruh\ jawaban} \quad \text{.....(2)}$$

Skor responden pada masing-masing perangkat tes dalam penelitian dikelompokkan menjadi kelompok tinggi (27% skor responden tertinggi) dan kelompok rendah (27% skor responden rendah). Oleh sebab itu untuk menentukan tingkat kesukaran butir, maka rumus yang digunakan adalah :

$$IK = \frac{JB_A + JB_B}{JS_A + JS_B} \quad \text{.....(3)}$$

Dimana:

IK = Indeks kesukaran
JSA = jumlah siswa kelompok atas
JSB = jumlah siswa kelompok bawah
JBA = jumlah siswa yang menjawab benar kelompok atas
JBB = jumlah siswa yang menjawab benar kelompok bawah.

Taraf sukar butir disimbolkan P_i (proporsi jawaban benar). Makin sedikit jawaban betul maka makin kecil P_i makin sukar butir itu, makin besar P_i dan makin tidak sukar butir tersebut. Artinya P_i mendekati 0, maka soal terlalu sukar, sedangkan jika P_i mendekati 1, maka soal tersebut terlalu mudah.

Taraf sukar butir dalam bentuk proporsi memiliki rentang diantara 0 s/d 1. Nilai sedang terletak ditengah-tengah pada $P_i = 0{,}50$ yang memisahkan butir tidak sukar dari butir sukar. Butir tidak sukar $P_i > 0{,}50$ dan butir sukar memiliki $P_i < 0{,}50$ (Dali, 2013: 282-283). Allen & Yen (1979: 122) menyatakan bahwa secara umum indeks kesukaran butir sebaiknya terletak pada interval 0,3 sampai 0,7. Pada interval ini informasi tentang kemampuan siswa akan diperoleh secara maksimal.

Daya Beda

Daya beda soal ialah kemampuan butir soal dalam membedakan antara peserta tes kelompok rendah dengan peserta tes kelompok tinggi. Rumus yang digunakan untuk memperoleh nilai daya beda butir tes (Dali, 1992: 68) adalah :

$$D_{tj} = \left(\frac{1}{M_T} f_{iT}\right) - \left(\frac{1}{M_R} f_{iR}\right) \quad \text{.....(4)}$$

Dimana

D_{tj} = indeks daya beda butir soal ke-j
M_T = jumlah peserta kelompok tinggi
f_{jT} = jumlah peserta kelompok tinggi yang menjawab benar soal ke-j
M_R = jumlah peserta kelompok rendah
f_{jR} = jumlah peserta kelompok rendah yang menjawab benar soal ke-j

Dalam hal ini, rumus diatas sama dengan:

$$DP = \frac{JB_A - JB_B}{JS_A} \text{ atau } DP = \frac{JB_A - JB_B}{JS_B} \quad \text{.....(5)}$$

Dimana :

DP = indeks daya beda
JSA = jumlah siswa kelompok atas
JSB = jumlah siswa kelompok bawah
JBA = jumlah siswa yang menjawab benar kelompok atas
JBB = jumlah siswa yang menjawab benar kelompok bawah.

Klasifikasi interpretasi untuk daya beda dapat disusun sebagai berikut:

DP < 0,00 sangat jelek
0,00 < DP < 0,20 jelek
0,20 < DP < 0,40 cukup
0,40 < DP < 0,70 baik
0,70 < DP < 1,00 sangat baik

Pada suatu butir soal, indeks daya beda dikatakan baik jika lebih besar atau sama dengan 0,3. Indeks daya pembeda suatu butir yang kecil nilainya akan menyebabkan butir tersebut tidak dapat membedakan siswa yang kemampuannya tinggi dan siswa yang kemampuannya rendah. Pada analisis tes dengan *content-referenced measures*, indeks daya pembeda butir tidak terlalu perlu jadi perhatian, asalkan tidak negatif (Ebel and Frisbie, 1986; Frisbie, 2005). Jika nilainya kecil, menunjukkan bahwa kemencengan distribusi skor dari populasi, yang juga mengakibatkan validitas tes menjadi rendah.

Reliabilitas Tes

Reliabilitas merupakan derajat keajegan (*consistency*) diantara dua buah hasil pengukuran pada objek yang sama (Mehrens and Lehmann, 1973: 102). Sedangkan Allen & Yen (1979: 62) menyatakan dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Artinya bahwa suatu tes dinyatakan reliabel jika hasil pengukuran mendekati keadaan peserta tes yang sebenarnya. Karena pengukuran dalam pendidikan tidak dapat langsung dilakukan pada ciri atau karakter yang akan diukur. Karena ciri atau karakter ini bersifat abstrak. Oleh karena itu sulit memperoleh alat ukur yang stabil untuk mengukur karakteristik seseorang (Mehrens and Lehmann, 1973: 103).

Reliabilitas tes dalam hal ini adalah suatu nilai yang mampu menggambarkan sejauhmana konsistensi hasil ukur bila dilakukan pengukuran berulang-ulang terhadap responden yang memiliki gejala sama dengan alat ukur yang sama. Rumus yang digunakan untuk menentukan koefisien

reliabilitas pada masing-masing instrumen diperoleh dari analisis butir soal yakni Kuder Richardson-20 (KR-20), yaitu:

$$r_{tt} = \left(\frac{n}{n-1}\right) - \left(\frac{SD_T - \Sigma pq}{SD_T}\right) \quad \text{.....(5)}$$

Penentuan formula reliabilitas diatas didasarkan pada bentuk tes yang digunakan adalah tes pilihan ganda (*multiple choice*) yang menghasilkan skor dikotomi. Meskipun tidak ada perjanjian secara tegas, tes yang digunakan untuk membuat keputusan pada siswa secara perseorangan sebaiknya memiliki koefisen reliabilitas minimal sebedar 0,85.

Fokus penelitian ini adalah membedakan tiga tipe tes Pilihan Ganda berdasarkan analisis butir soal dan bukan mengkaji hasil belajar siswa sebagai responden. Dengan alasan tersebut maka perangkat ketiga tes hanya divalidasi isi dan tidak diujicoba untuk melakukan validasi empirik dengan pertimbangan: (1) setiap butir soal yang ada pada perangkat tes PGA (Pilihan Ganda Asosiasi), PGB (Pilhan Ganda Biasa), dan PGH (Pilihan Ganda Hubungan antar Hal) harus setara dalam hal materi/konsep, kategori kognitif, jumlah soal dan nomor urut soal; (2) hasil penelitian yang diharapkan adalah adanya perbedaan taraf kesukaran butir, daya beda dan reliabilitas tes pada ketiga perangkat tes.

METODE PENELITIAN

Penelitian ini bertujuan membandingkan karakteristik butir tes Pilihan Ganda (PG) yaitu tes PGA, PGB dan PGH. Penelitian dilaksanakan di beberapa Sekolah Menegah Atas (SMA) DKI Jakarta diantaranya: SMA 48, SMA 71, SMA SULUH dan SMA 91.

Jenis penelitian ini termasuk penelitian survey dengan menggunakan instrumen yang dirancang dan dibuat peneliti setelah melalui validasi ahli dan empiris maka digunakan sebanyak masing-masing 30 butir tes pilihan ganda khusus pada matapelajaran sosiologi siswa kelas sebelas (XI). Sampel yang digunakan sebanyak 224 siswa yang diperoleh dengan menggunakan teknik acak secara proporsional atau *proporsional random sampling.*

Selain dianalisis tingkat kesukaran butir tes dan daya beda butir tes, dalam penelitian ini juga dianalisis perbedaan rerata kelompok koefisien reliabilitas tes PGB (μ_{PGB}) , Tes PGH (μ_{PGH}) dan Tes PGA (μ_{PGA}). Ketiga kelompok data reliabilitas diperoleh dari pengacakan terhadap skor responden masing-masing tipe tes. Pengacakan tersebut menggunaan teknik random dengan pengembalian (*replace randomized*), yaitu mengambil secara acak 30 dari 224 skor responden tes melengkapi kemudian dianalisis untuk menentukan koefisien reliabilitas tes. Program yang digunakan dalam teknik random data tes adalah Minitab versi 16.

Hal tersebut dilakukan sebanyak 30 (tiga puluh) kali, sehingga diperoleh 30 gugus data yang merupakan kelompok koefisien reliabilitas tes Pilihan Ganda Biasa (n_{PGB} = 30). Hal yang sama dilakukan pada skor responden tes pilihan ganda asosiasi, sehingga terbentuk 30 gugus data yang merupakan kelompok koefisien tes pilihan ganda asosiasi (n_{PGA}), dan tes pilihan ganda hubungan antar hal (n_{PGH}).

Penentuan nilai dari tingkat kesukaran, daya beda dan reliabilitas, mengandalkan skor responden dari instrumen tes hasil belajar sosiologi pada ketiga tipe tes. Untuk itu hasil pengukuran yang diperoleh harus dinyakini bentuk keabsahannya, dengan kata lain semaksimal mungkin terhindar dari kesalahan ukur berikut: (a) objek/individu yang diukur; (b) alat ukur; (c) petugas/pengumpul data dilapangan (Agung, 1992: 46).

Bilamana kesalahan terjadi pada alat ukur/instrumen tes, maka penelitian akan gagal secara keseluruhan. Untuk mencegah itu, maka sebelum naskah diambil sebagai perangkat tes atau disahkan sebagai instrumen penelitian, perlu dilihat validitas isi/teoritik/konsep juga validitas empiriknya. Validitas empirik instrumen dalam perhitungan menggunakan rumus *korelasi biserial* (Djaali, Muljono dan Ramly, 2000: 77).

Perangkat tes dalam penelitian ini telah dikalibrasi melalui validasi isi oleh team penilai yang berkompeten dalam bidang

sosiologi. Validasi isi dilakukan melalui penelaahan kisi-kisi untuk memastikan bahwa soal-soal yang tercakup dalam perangkat tes sudah mampu mengukur secara proporsional isi materi yang seharusnya dikuasai oleh siswa sesuai jenjang kelasnya.

HASIL PENELITIAN

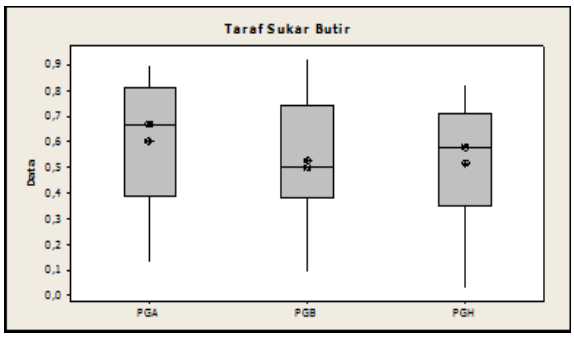
Hasil analisis diskriptif pada taraf sukar butir ketiga tipe tes pilihan ganda dari masing-masing tes sebanyak 30 butir soal terlihat pada tebel berikut:

Tabel 1. Nilai Taraf Sukar Butir Tes PGA, PGB dan PGH

	PGA	PGB	PGH
Mean	,5997	,5253	,5170
Median	,6650	,5000	,5800
Modus	,81	,39[a]	,58
Varians	,052	,055	,060

Sumber: Data diolah, 2016

Tabel 1 terlihat nilai rerata dari taraf sukar butir ketiga tipe tes pilihan ganda tidak jauh berbeda, ketiga tipe tes termasuk kriteria taraf sukar butir ideal karena terletak antara 0,3-0,7. Hal itu juga didukung dengan besaran nilai varian yang tidak jauh berbeda. Nilai varians dari taraf sukar butir menunjukkan perbedaan nilai diantara taraf sukar butir tiap tipe tes pilihan ganda. Namun yang membedakan ketiga tipe tes pilihan ganda terlihat pada nilai modus, PGA memiliki nilai tertinggi yaitu 0,81, kemudian PGH dan PGB. Hasil analisis taraf sukar butir digambarkan:



Gambar 1. Box Plot Taraf Sukar Butir Tes PGA, PGB dan PGH

Dari boxplot menunjukkan tes PGA memiliki nilai mean, median dan modus tertinggi yang berarti taraf sukar butir paling unggul (mudah) dibandingkan dengan kedua tes PGH dan PGB. Kemudian nilai mean, dari taraf sukar butir tes PGB lebih baik (mudah) dibandingkan tes PGH.

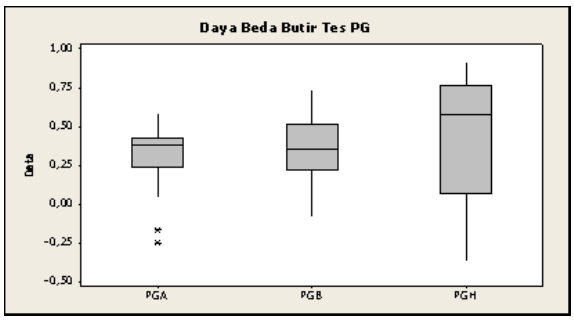
Hasil perhitungan indeks daya beda ketiga tes Pilihan Ganda dari masing-masing tes terdiri atas 30 butir tes diperoleh nilai sebagai berikut :

Tabel 2. Indeks Daya Beda Butir Tes PGA, PGB, dan PGH

	PGA	PGB	PGH
Mean	,3059	,3514	,4596
Median	,3770	,3610	,5740
Modus	,38	,33[a]	,89
Varians	,043	,048	,156

Sumber: Data diolah, 2016

Indeks daya beda antar tes pilihan ganda secara statistik deskriptif menunjukkan nilai mean dan varians terkecil pada tes PGA, sebaliknya nilai mean dan varians terbesar berada pada tes PGH. Hal ini menunjukkan kemampuan membedakan siswa yang kemampuannya tinggi dengan berkemampuan rendah pada tes PGA tergolong kecil, sebaliknya pada tes PGH tergolong besar. Meskipun kedua nilai rerata dari indeks daya beda butir tes tergolong baik karena diatas dari nilai 0,30.



Gambar 2. Box Plot Indeks Daya Beda Butir Tes PGA, PGB dan PGH

Gambar 2 boxplot indeks daya beda butir tes memperlihatkan sebaran indeks daya beda

butir tes, pada tes PGH sebaran paling tinggi, kemudian tes PGB, disusul tes PGA yang paling kecil dalam membedakan kemampuan siswa tinggi dengan siswa berkemampuan rendah. Ketiga tipe tes pilihan ganda termasuk dalam kriteria tes memiliki indeks daya beda yang baik.

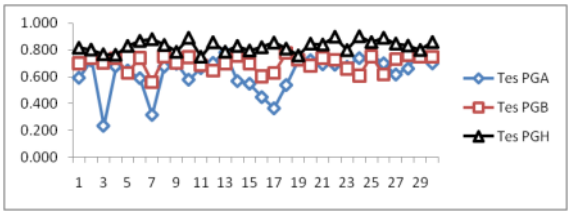
Hasil pembentukan 30 gugus data pada masing-masing tes pilihan ganda, kemudian dihitung koefisien reliabilitas diperoleh nilai terlihat pada tabel berikut :

Tabel 3. Statistik Koefisien Reliabilitas Tes PGA, PGB, dan PGH

	PGA	PGB	PGH
Mean	,6267	,7010	,8306
Median	,6750	,7160	,8330
Modus	,70	,74[a]	,86[a]
Varians	,018	,003	,002

Sebaran nilai (varians) dari koefisien reliabilitas paling konsisten yaitu tes PGH dengan nilai varians sebesar 0,002, kemudian tes PGB dengan nilai varians sebesar 0,0031 dan konsistensi terendah pada tes PGA yaitu sebesar 0,0173. Rupanya konsistensi nilai koefisien reliabilitas pada tes PGH juga diiringi dengan besaran nilai rerata koefisien reliabiltas terbesar yaitu μ_{PGH} = 0,831, kemudian tes PGB yaitu μ_{PGB} = 0,701 dan diikuti dengan tes PGA yaitu μ_{PGA} = 0,627.

Jika mengikuti kriteria umum menurut Naga (2013, 240-241) maka tes PGA dengan nilai rerata 0,627 termasuk dalam kriteria koefisien reliabilitas yang “bermasalah”, sedangkan tes PGB termasuk kriteria “dapat diterima” dan tes PGH termasuk kriteria “baik”. Konsistensi dari koefisien reliabilitas ketiga tipe tes pilihan ganda terlihat pada gambar berikut:



Gambar 3.
Koefisien Reliabilitas Tes PGA, PGB dan PGH

Tes PGH berdasarkan nilai rerata dari koefisien reliabilitas termasuk kriteria “baik” atau paling stabil dan besaran nilai varians termasuk paling kecil, hal ini menandakan bahwa tes pilihan ganda tipe hubungan antar hal dapat menguji dengan stabil kemampuan siswa.

PEMBAHASAN

Taraf sukar butir tes PGH berada posisi terendah diantara dua tipe tes pilihan ganda lainnya, hal ini tidak terlepas dari keunikan dalam menjawab butir tes PGH, dimana siswa harus menentukan kalimat pertama, dan kalimat kedua benar atau salah, baru kemudian mengolah pemikirannya secara rasional maupun teori kedua pernyataan tersebut memiliki kaitan ataupun tidak. Sejalan tujuan tes PGH bahwa siswa diharapkan dapat menunjukkan kemampuan mereka pada hubungan diantara fakta yang merupakan dasar dalam pengembangan kemampuan pemahaman (*understanding*), kemampuan berfikir dan kemampuan lainnya yaitu *ability to intepret causeand-effect relationship* (Surapranata, 2005: 148).

Ketidak-konsistensi siswa dalam menjawab butir-butir tertinggi pada tes PGA, namun ukuran tersebut tidak dapat dianggap bahwa bentuk tes ini bermasalah. Sebab merujuk hasil deskripsi nilai dari taraf sukar butir dan daya beda kualitas tes PGA memenuhi persyaratan sebagai salah satu tipe tes pilihan ganda yang baik. Thorndike & Hagen (1969: 114-116) menyatakan tes asosiasi pilihan ganda sebagai variasi butir tes pilihan ganda yang kompleks dan menggunakan pasangan pernyataan. Dikatakan sebagai tes PG kompleks karena variasi dari bentuk dan kombinasi pilihan ganda yang digunakan sebagai pilihan jawaban. Siswa tidak hanya diberikan satu pilihan jawaban yang benar saja, namun dapat dua atau bahkan tiga pilihan jawaban yang benar. Hal ini menjadi sebab variansi tes PGA lebih sulit dan lebih mampu memberikan perbedaan hasil belajar yang dicapai oleh siswa. Hal senada diungkap Earlyanti (2012:

231) menyatakan bahwa butir-butir dalam tes pilihan ganda asosiasi memiliki akurasi yang sangat baik dalam mengukur kemampuan siswa.

PENUTUP

Hasil analisis deskriptif menunjukkan nilai koefisien reliabilitas dari tes PGH paling konsisten mengukur kemampuan siswa, demikian pula taraf sukar butir dan indeks daya beda termasuk dalam kriteria tes yang baik. Sehingga kedua tipe tes PGA dan PGH layak digunakan untuk mengukur kemampuan siswa disamping tes tipe PGB yang biasa digunakan. Meskipun memiliki kompleksitas dalam proses menjawab soal ujian berdasarkan data terbukti tipe tes PGA dan PGH memiliki kriteria yang cukup baik untuk mampu mengukur kemampuan siswa pada mata pelajaran berjenis pemahaman.

DAFTAR PUSTAKA

Agung, I Gusti Ngurah Agung. (1992). *Metode Penelitian Sosial Pengertian dan Pemakaian Praktis*. Jakarta: Gramedia.

Allen, M.J. & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Djaali, Pudji Muljono dan Ramly. (2000). *Pengukuran Dalam Bidang Pendidikan.* Jakarta: PPS UNJ.

Earlyanti, Novi Indah. (2012). *Komparasi Fungsi Informasi Butir Model Logistik Dua Parameter Ditinjau dari Ragam Bentuk Soal Pilihan Ganda dan Model Pensekoran*. Disertasi: Universitas Negeri Jakarta.

Ebel, R.L. & Frisbie, D.A. (1986). *Essenstials of Educational Measurement*. Englewood Cliftfs, NJ:Prentice-Hall, Inc

Grondlund, Norman. E & Linn, Robert L. (1990). *Measurement and evaluation in teaching*. New York: Mc Millan Publishing Company.

Hajaroh, Siti. (2011). *Komparasi Bentuk Tes Pilihan Ganda Dengan Tes Menjodohkan (Matching Test) Ditinjau dari Tingkat Kesukaran, Daya Beda dan Koefisien Reliabilitas*. Tesis: Universitas Negeri Jakarta.

Karmel, Louis J. & Karmel, Marylin O. (1978). *Measurement and Evaluation in Research*. New York: Macmillan, Publisher, Inc, 1978.

Mehrens, W.A & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. New York: Hold, Rinehart and Wiston, Inc.

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

Surapranata, Sumarna. (2005). *Panduan Penulisan Tes Tertulis Implementasi Kurikulum 2004.* Bandung: PT Remaja Rosdakarya.

Sudjono, Anas. (2013). *Pengantar Evaluasi Pendidikan*. Jakarta: PT. RajaGrafindo Persada.

Permendikbud No.23 Tahun 2016 tentang Standar Penilaian Pendidikan.

Thorndike, Robert L and Hagen, Elizabeth. (1969). *Measurement and Evaluation in Psychology and Education*. New York: John Wiley & Sons, 1969.