ANALISIS BUTIR SOAL SECARA KUANTITATIF

Disusun Oleh:

Kelompok 2

Pendidikan Biologi Kelas A :

1. Raafi Nivika (1613024034)
2. Neneng Indah (1653024002)
3. Rinjani Rosbandini (1653024012)
4. Nabila Amara Putri (1813024033)
5. Viny Chintia (1813024041)
6. Eksya Fahira Putri (1813024053)
7. Dea Milliony Putri (1813024057)
8. Indira Ratna Dewanti (1813024059)

Mata Kuliah : Evaluasi Pembelajaran Biologi

Dosen : Berti Yolida, S.Pd., M.Pd.

Rini Rita T. Marpaung, S.Pd., M.Pd.

Dr. Dewi Lengkana., M.Sc.

JURUSAN PENDIDIKAN BIOLOGI

FAKULTAS KEGURUAN DAN ILMU PENDIDIKAN

UNIVERSITAS LAMPUNG

2020/2021

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: 308). Tujuan penelaahan adalah untuk mengkaji dan menelaah setiap butir soal agar diperoleh soal yang bermutu sebelum soal digunakan. Di samping itu, tujuan analisis butir soal juga untuk membantu meningkatkan tes melalui revisi atau membuang soal yang tidak efektif, serta untuk mengetahui informasi diagnostik pada siswa apakah mereka sudah/belum memahami materi yang telah diajarkan (Aiken, 1994: 63).

Analisis Butir Soal Secara Kuantitatif

Penelaahan soal secara kuantitatif adalah penelaahan butir soal didasarkan pada data empirik dari butir soal yang bersangkutan.Salah satu pendekatan pada analisis butir soal secara kuantitatif adalah pendekatan secara klasik. Pada pendekatan ini proses penelaahan melalui informasi dari jawaban siswa guna meningkatkan mutu butir soal yang bersangkutan. Kelebihan analisis butir soal secara klasik adalah murah, dapat dilaksanakan sehari-hari dengan cepat, sederhana, familier dan dapat mengunakan data dari beberapa peserta. Aspek yang perlu diperhatikan dalam analisis butir soal secara klasik adalah telaah dari segi validitas, reliabilitas, daya pembeda, dan tingkat kesukaran soal (Karim, 2018). Kemudian kelemahan dari analisis butir soal secara klasik adalah:

(1) Tingkat kemampuan dalam teori klasik adalah "*truescore*". Jika tes sulit artinya tingkat kemampuan peserta didik mudah. Jika tes mudah artinya tingkat kemampuan peserta didik tinggi.

(2) Tingkat kesukaran soal didefinisikan sebagai proporsi peserta didik dalam grup yang menjawab benar soal. Mudah/sulitnya butir soal tergantung pada kemampuan peserta didik yang dites dan kemampuan tes yang diberikan.

(3) Daya pembeda, reliabilitas, dan validitas soal/tes didefinisikan berdasarkan grup peserta didik.

Penelaahan butir soal dengan menggunakan Item ResponseTheory (IRT) atau teori jawaban butir soal. Teori ini merupakan suatu teori yang menggunakan fungsi matematika untuk menghubungkan antara peluang menjawab benar suatu soal dengan kemampuan siswa. IRT merupakan hubungan antara probabilitas jawaban suatu butir soal yang benar dan kemampuan siswa atau tingkatan/level prestasi siswa.

Kelebihan :

(1) asumsi banyak soal yang diukur pada trait yang sama, perkiraan tingkat kemampuan peserta didik adalah independen;

(2) asumsi pada populasi tingkat kesukaran, daya pembeda merupakan independen sampel yang menggambarkan untuk tujuan kalibrasi soal;

(3) statistik yang digunakan untuk menghitung tingkat kemampuan siswa diperkirakan dapat terlaksana

Kelemahan : prosesnya cukup rumit dan sulit

Penghitungan dalam penelaahan butir soal secara kuantitatif dapat menggunakan bantuan kalkulatoscientific atau program komputer. Program yang sudah dikenal secara umum adalah EXCEL, SPSS (*Statitistical Program for Social Science*), atau program khusus seperti ITEMAN (analisis secara kiasik), RASCAL, ASCAL, BILOG (analisis secara item respon teori atau IRT), FACETS (analisis model Rasch untuk data kualitatif) atau ANATES.

Aspek-aspek yang harus diperhatikan dalam menganalisis butir soal secara kuantitatif, sebagai berikut :

1. Validitas

Untuk menentukan valid atau tidak valid suatu butir soal, maka diperlukan interprestasi koefisien validitas. Interpretasi koefisien validitas bersifat relatif, artinya tidak ada batasan pasti mengenai koefisien terendah yang harus dipenuhi agar validitas dinyatakan memuaskan. Koefisien validitas yang baik, setinggi mungkin mendekati harga rxy = 1,00. Akan tetapi untuk memperoleh koefisien validitas yang tinggi lebih sulit daripada memperoleh koefisien reliabilitas yang tinggi. Hal ini menjadikan alasan setiap penulis butir soal untuk bersikap realistik dan tidak menuntut koefisien yang setinggi koefisien reliabilitas (Alwi, 2015)

Suatu kesepakatan umum menyatakan bahwa koefisien validitas dapat dianggap memuaskan apabila melebihi r_{xy} = 0,30. Siapapun boleh menerima atau menolak batasan ini karena memang penetapan angka tersebut tidak didasari logika matematika melainkan merupakan konvensi tidak tertulis yang didasari oleh pertimbangan professional dan pengalaman saja.

a. Uji Validitas Untuk Butir Soal Bentuk Pilihan Ganda menggunakan rumus ***Point Biserial***

$$r_{pbi} = \frac{\bar{X}_p - \bar{X}_q}{s}\sqrt{pq}$$ Atau $$r_{pbi} = \frac{\bar{X}_p - \bar{X}_t}{s}\sqrt{\frac{p}{q}}$$

Keterangan :

$\bar{X}p$: rata-rata skor kemampuan peserta didik yangmenjawabbenar

$\bar{X}q$: rata-rata skor kemampuan peserta didik yangmenjawabsalah

$\bar{X}t$: rata-rata skor dari skortotal

s : simpangan baku skor total

p : proporsi jawaban benar terhadap semua jawaban siswa

q :1-p

b. Uji Validitas Untuk Butir Soal Skala Kontinum (Uraian dan Non-Tes)

Skor butir instrumen atau soal tes kontinum (misalnya bentuk soal Uraian dan skala sikap dengan sekor butir 0 – 10 atau 1- 5) dan diberi symbol x_i dan sekor total instrument atau tes diberi symbol x_t, maka rumus yang digunakan untuk menghitung koefisien korelasi antara sekor butir instrumen atau soal dengan sekor total instrumen atau sekor total tes adalah rumus ***Product Moment***:

Rumus: $r_{xy} = \dfrac{N\sum X_1X_2-(\sum X_1)(\sum X_2)}{\sqrt{\{N\sum X_1{}^2-(\sum X_1)^2\}\{N\sum X_2{}^2-(\sum X_2)^2\}}}$

Di mana r_{xy} adalah koefisien korelasi yang dicari, N adalah *Number of Cases,* X_1 adalah sekor butir dan X_2 adalah sekor total.

Terdapat empat jenis validitas yaitu :

1. Validitas Isi (*Content Validity*)

   Validitas ini adalah yang ditilik dari segi isi tes sendiri sebagai alat pengukur hasil belajar yaitu sejauh mana tes hasil belajar sebagai alat pengukur hasil belajar peserta didik, isinya telah dapat mewakili secara representatif terhadap keseluruhan materi atau bahan pelajaran yang seharunya diujikan.

2. Validitas Konstruksi (*Construct Validity*)

   Validitas ini adalah yang ditilik dari segi susunan, keangka atau rekaannya.

3. Validitas Ramalan (*Predictive Validity*)

   Validitas ini adalah suatu kondisi yang menunjukkan seberapa jauhkah sebuah tes telah dapat dengan secara tepat menunjukkan kemampuannya untuk meramalkan apa yang bakal terjadi pada masa mendatang.

4. Validitas Bandingan (*Concurrent Validity*)

Validitas ini adalah kemampuan sebuah tes dalam kurun waktu yang sama engan secara tepat telah mampu menunjukkan adanya hubungan searah antara tes pertama dengan tes berikutnya. Validitas bandingan juga dikenal dengan istilah validitas sama saat, validitas pengalaman, aatau validitas sekarang (Fakhrun, 2018)

Menurut Gronlund dalam Subali (2014) hakekat validitas tes dapat ditinjau dari beberapa aspek sebagai berikut :

a. Validitas tes mengacu kepada kepantasan/ketercukupan (*appropriateness*) interpretasi hasil pengukuran yang dikenakan pada kelompok atau individu testi/peserta ujian terhadap instrumen tes yang digunakan. Berbicara tentang validitas maka tentang interpretasi yang dapat dibuat dari hasil pengukuran. Semakin valid suatu instrumen tes semakin tepat interpretasi hasil pengukuran yang diperoleh.
b. Validitas adalah sesuatu hal yang memiliki derajat sehingga dapat dibedakan antarayang benar-benar sahih (valid) dan yang benar-benar tidak sahih (invalid) atau antarayang rendah, sedang, dan tinggi validitasnya.
c. Validitas adalah mengacu kepada suatu tujuan yang spesifik. Misalnya, suatu tesaritmatik dinyatakan memiliki derajat validitas yang tinggi jika dapat menunjukkanketerampilan dalam komputasi, dinyatakan memiliki derajat validitas yang rendah jikahanya untuk menunjukkan kemampuan berpikir aritmatik, dan memiliki derajatvaliditas yang sedang jika dapat digunakan untuk memprediksi keberhasilan belajararitmatik ke depan. Akan tetapi, interpretasi tersebut tidak mampu untuk menilai ataumenggambarkan validitas guna mempertimbangkan penggunaan hasil tes yangdiperoleh.
d. Validitas adalah konsep kesatuan (*unitary*). Hakekat konsepsual validitas tes untuk bidang pendidikan dan psikologi kini didasarkan pada tiga aspek yang ditinjau berdasarkan bukti-bukti empiris yaitu aspek isi (*content*), aspek hubungannya dengankriteria (*criterion related*) dan aspek konstruk (*construct*).

5. Reliabilitas

Kriteria lain dari tes yang baik adalah memiliki tingkat kepercayaan atau reliabilitas secara memadai. Reliabilitas berkaitan dengan ketetapan hasil pengukuran. Suatu tes dinyatakan reliabel apabila tes itu dikenakan pade subyek yang sama dalam kurun waktu yang berbeda memberikan hasil (skor yang kurang lebih sama). Reliabilitas merupakan kestabilan skor yang diperoleh orang yang sama ketika diuji ulang dengan menggunakan tes yang sama pada situasi yang berbeda. Kepercayaan tes menunjuk pada pengertian apakah suatu tes dapat mengukur secara konsisten sesuatu yang diukur dari waktu ke waktu. Selanjutnya dinyatakan bahwa konsistensi itu terkait dengan hal-hal sebagai berikut: (1) tes dapat memberikan hasil yang relatif tetap

terhadap sesuatu yang diukur, (2) jawaban siswa terhadap butir-butir tes secara relatiftetap, dan (3) tes tersebut diperiksa oleh siapapun akan memberikan hasil yang kurang lebih sama. Terdapat dua macam konsistensi reliabiltas tes, yakni konsitensi internal dan eksternal. Konsistensi internal merupakan suatu kriteria di mana ketetapan ditentukan berdasarkan hasil tes itu sendiri. Sementara itu, pada konsistensi eksternal, ketetapan ditentukan dengan cara mengkorelasikan dengan hasil tes lain. Suatu tes dinyatakan reliabel apabila hasil tes pertama dan kedua memiliki korelasi tinggi. Sebaliknya, apabila tes pertama dan kedua korelasinya rendah maka dinyatakan tes itu tidak reliabel (Sujati, 2005).

Cronbach menyatakan ada tiga mekanisme untuk memeriksa reliabiltas tanggapan responden terhadap tes yaitu :

a. Teknik *test retest* adalah pengetesan dua kali dengan menggunakan suatu tes yang sama pada waktu yang berbeda.
b. Teknik belah dua, pada teknik ini ppengukuran dilakukan dengan dua kelompok item yang setara pada saat yang sama.
c. Bentuk *ekivalen*, pengukuran dilakukan dengan menggunakan dua tes yang dibuat setara kemudian diberikan kepada responden atau objek ukur tes dalam waktu yang bersamaan. Skor kedua kelompok item tersebut dikorelasikan untuk mendapatkan realibilitas tes (Fakhrun, 2018)

Di antara pendekatan konsistensi internal adalah metode *Kuder-Richardson* 20 (KR-20) dan *Alpha Cronbach*. Menurut Nitko (1983: 395) *Kuder-Richardson* 20 (KR-20) digunakan untuk menghitung nilai reliabilitas tes dalam bentuk tes objektif yang hanya menggunakan sekor dikotomi, yaitu bila benar = 1 dan salah = 0, seperti pada bentuk tes pilihan ganda. Sedangkan koefisien *Alpha Cronbach* digunakan untuk menghitung nilai reliabilitas tes dalam bentuk uraian atau skala sehingga pengukurannya tidak hanya menggunakan sekor benar = 1 dan salah = 0, seperti pada tes objektif, melainkan dapat menggunakan sekor 1 – 10 atau skala 1 – 5, dan sebagainya.

Adapun rumus:

Kuder-Richardson 20 (KR-20): $r = \frac{k}{k-1}\left(1 - \frac{\sum pq}{s^2}\right)$ dan,

Koefisien *Alpha Cronbach* : $r = \frac{k}{k-1}\left(1 - \frac{\sum S_b^2}{S_t^2}\right)$.

Selanjutnya, untuk menentukan reliabel atau tidak reliabel suatu tes, makadiperlukan interprestasi koefisien reliabilitas. Sebagaimana pada interpretasi validitas, interpretasi terhadap koefisien reliabilitas juga bersifat relatif, tidak ada

batasan pasti mengenai koefisien terendah yang harus dipenuhi agar suatu pengukuran dapat disebut reliabel. Terdapat dua kriteria empirik untuk menentukan besarnya koefisien reliabilitas yang memadai. Kriteria empirik pertama berkenaan dengan bidang ilmu, dan kriteria empirik kedua berkenaan dengan statistika.

Pada umumnya, untuk bidang ilmu yang memiliki pengukuran dengan kecermatan tinggi seperti pengukuran keberhasilan belajar matematika yang baku memiliki koefisien reliabilitas yang tinggi yakni di atas 0,90. Dengan demikian, koefisien reliabilitas yang memadai pada ujian keberhasilan matematika adalah sekitar 0,90. Menurut Ebel (1979: 275) suatu koefisien reliabilitas di sekitar 0,90 atau lebih, dapat dianggap memuaskan.

Sebaliknya, untuk bidang ilmu yang belum memiliki kecermatan pengukuran yang tinggi, koefisien reliabilitas yang rendah pun sudah dianggap memadai. Hal ini dapat diperiksa pada jurnal ilmu bersangkutan. Jika di dalam jurnal bidang ilmu itu ditemukan bahwa koefisien reliabilitas pada pengukurannya di sekitar 0,40 maka koefisien reliabilitas yang memadai adalah 0,40 (Naga, 2009: 93).

Secara statistika, koefisien reliabilitas yang memadai adalah koefisien korelasi linear yang memadai. Kriteria empirik menyatakan bahwa irisan variansi X dan variansi Y yang disebut koefisien determinasi (d) dianggap memadai apabila telah mencapai = 0,50. Koefisien determinasi berkaitan dengan koefisien korelasi linear (ρ_{xy}), makahubungan keduanya adalah $\rho_{xy} = \sqrt{d}$. Jika d = 0,50 dianggap memadai maka koefisien korelasi linear dengan nilai $\sqrt{d} = \sqrt{0{,}50} = 0{,}71$ dianggap sudah memadai. Karenakoefisien reliabilitas merupakan jenis koefisien korelasi linear, maka secara statistika koefisien reliabilitas yang memadai adalah **0,71** atau lebih.

Tinggi rendahnya koefisien reliabilitas dipengaruhi oleh berbagai faktor. Koefisien reliabilitas dipengaruhi oleh panjang tes, tingkat kesukaran, homogenitas kelompok dan daya beda butir. Yang paling banyak mempengaruhi koefisien reliabilitas adalah tingkat kesukaran. Hal ini karena menyangkut variasi jumlah soal yang dijawab benar. Menurut Saifuddin Azwar (1996) tinggi-rendahnya koefisien reliabilitas dipengaruhi oleh panjang tes dan daya beda butir. Tes yang dibangun oleh banyak butir yang berdaya beda tinggi cenderung memiliki tingkat reliabilitas tinggi. Sebaliknya, konstruksi tes yang dibangun oleh butir-butir soal yang berdaya beda rendah cenderung memiliki tingkat reliabilitas rendah atau bahkan negative (Sujati, 2005).

Reliabilitas Soal Bentuk Pilihan Ganda dengan Menggunakan Rumus *Kuder Richadson* 20 (KR-20)

Rumus *Kuder Richadson* 20 (KR-20):

$$KR-20=\frac{k}{k-1}\left[1-\frac{\Sigma p(1-p)}{(SD)^2}\right]$$

Keterangan:

k = banyaknya butir soal

p = proporsi peserta tes yang menjawab benar

q = 1 – q

SD = varians total

2. Reliabilitas Soal Bentuk Uraian dengan Menggunakan Rumus Alfa Cronbach.

Rumus:

$$r=\frac{k}{k-1}\left[1-\frac{SD_t^{\,2}-\sum(SDi)^2}{(SD_t)^2}\right]$$

Keterangan:

r = koefisien reliabilitas seluruh tes

n = jumlah soal dalam tes

SD = varian skor-skor total pada tes

ΣSD = jumlah varian butir tes

Nilai Korelasi diatas konsultasikan dengan tabel kriteria korelasi koefisien, yaitu:

- 0,00 ≤ r11 ≤ 0,20 = korelasi sangat rendah
- 0,20 ≤ DB ≤ 0,40 = korelasi rendah
- 0,40 ≤ DB ≤ 0,70 = korelasi cukup
- 0,70 ≤ DB ≤ 0,90 = korelasi tinggi
- 0,90 ≤ DB ≤ 1,00 = korelasi sangat tinggi (sempurna)

6. Tingkat Kesukaran

Tingkat kesukaran soal adalah peluang untuk menjawab benar suatu soal pada tingkat kemampuan tertentu yang biasanya dinyatakan dalam bentuk indeks. Untuk

menghitung tingkat kesukaran soal uraian berbeda dengan cara yang digunakan pada tes objektif. Untuk menghitung tingkat kesukaran ada beberapa cara yaitu:

(1) *proportion correct*,

(2) indeks kesukaran linier,

(3) indeks *Davis*, dan

(4) skala *Bivariat*.

Untuk bentuk soal Pilihan Ganda, cara yang paling mudah dan paling umum digunakan adalah dengan skala rata-rata atau proporsi menjawab benar atau *proportion correct* (*p*), yaitu jumlah peserta tes yang menjawab benar pada soal yang dianalisis dibandingkan dengan peserta tes seluruhnya. Persamaan yang digunakan untuk menentukan tingkat kesukaran (*p*) ini adalah:

$$P = \frac{\sum B}{N}$$

P : proporsi menjawab benar pada butir tertentu

$\sum B$: banyaknya peserta tes menjawab benar

N : jumlah peserta tes yang menjawab benar

Sedangkan untuk bentuk Soal Uraian, Untuk mengetahui tingkat kesukaran soal bentuk uraian digunakan rumus berikut ini:

$$Tingkat\ Kesukaran = \frac{mean}{skor\ maksimum}$$

Besarnya tingkat kesukaran antara 0 dan 1. Tingkat kesukaran dikategorikan menjadi tiga bagian seperti tabel di bawah ini:

Tabel 1. Kategori Tingkat Kesukaran

Proportion Correct (*p*)	Kategori Soal
0,71 - 1,00	Mudah
0,31 - 0,70	Sedang
0,00 - 0,30	Sukar

Tingkat kesukaran butir soal memiliki 2 kegunaan, yaitu kegunaan bagi pendidik dan kegunaan bagi pengujian dan pengajaran.

Kegunaan bagi pendidik, antara lain :

a) Sebagai pengenalan konsep terhadap pembelajaran ulang dan memberi masukan kepada peserta didik tentang hasil belajar mereka.

b) Memperoleh informasi tentang penekanan kurikulum atau mencurigai butir soal yang bias

Kegunaan bagi pengujian dan pengajaran, antara lain :

a) Pengenalan konsep yang perlu diperlukan untuk diajarkan ulang

b) Tanda-tanda terhadap kelebihan dan kelemahan pada kurikulum sekolah

c) Memberi masukan kepada peserta didik

d) Tanda-tanda kemungkinan adanya butir soal yang bias

e) Merakit tes yang memiliki ketepatan daya soal (Fakhrun, 2018: 25-26).

7. Daya Pembeda Soal

Secara umum, daya pembeda diartikan sebagai kemampuan suatu butir untuk membedakan antara peserta tes yang berkemampuan tinggi dan berkemampuan rendah. Suatu butir dikatakan baik apabila butir tersebut dapat dijawab benar oleh sebagian besar peserta tes yang berkemampuan tinggi dan hanya dapat dijawab benar oleh sebagian kecil dari peserta tes yang berkemampuan rendah. Butir tes yang dapat dijawab benar atau salah oleh peserta tes yang berkemampuan tinggi dan berkemampuan rendah menunjukkan bahwa tes tersebut tidak memiliki daya beda (Sujati, 2005).

Daya pembeda dihitung berdasarkan jumlah jawaban benar untuk setiap butir soal antara peserta tes yang berkemampuan tinggi dan berkemampuan renda. Jika terjadi kelompok tes yang berkemampuan rendah lebih banyak menjawab benar daripada peserta tes yang berkemampuan tinggi, menunjukkan bahwa butir soal tersebut perlu direvisi atau diganti. Sebaliknya, apabila kelompok peserta berkemampuan tinggi lebih banyak menjawab benar, hal itu menunjukkan bahwa butir tersebut baik (Sujati, 2005).

Daya pembeda atau daya beda suatu butir soal berfungsi untuk menentukan dapat tidaknya suatu butir soal membedakan kelompok dalam aspek yang diukur sesuai dengan perbedaan yang ada pada kelompok itu. Tujuan dari pengujian daya pembeda adalah untuk melihat kemampuan butir soal dalam membedakan antara peserta didik yang berkemampuan tinggi (M_T) dengan peserta didik yang berkemampuan rendah (M_R).

Menurut Kelley dalam Naga (2009: 89) menemukan bahwa nilai optimal penggunaan ukuran kelompok adalah $M_T = M_R = 27\%$. Sejak itulah banyak orang menentukan pilahan 27% untuk kelompok tinggi dan 27% untuk kelompok rendah pada ukuran responden yang besar. Selanjutnya pemilahan responden ke kelompok tinggi dan kelompok rendah dilakukan melalui:

untuk responden (M) < 371, M_T=M_R=50%

untuk responden (M) ≥ 371, M_T=M_R=27%

Daya diskriminasi yang baik memang pada umumnya terdapat pada item yang tidak terlalu mudah dan juga tidak terlalu sukar, yaitu apabila harga p berkisar antara 0,40 sampai dengan 0,60. Dalam seleksi item, setiap item yang memiliki indeks diskriminasi lebih besar dari 0,50 dapat langsung dianggap baik, item yang memiliki indeks diskriminasi kurang dari 0,20 dapat langsung dibuang, sedangkan item lainnya dapat ditelaah lebih lanjut untuk direvisi.

Secara empiris, minimum tingkat daya pembeda yang memadai seperti tabel di bawah ini:

Tabel 2. Daya Beda Minimum

Nama Ahli	Daya Beda Minimum
Crocker & Algina (1986: 324)	0,2
Nunnaly (1970: 202)	0,2
Aiken (1994: 65)	0,2
Mehrens & Lehmans (1991: 167)	0,2

Kesepakatan beberapa ahli di atas menyatakan bahwa koefisien indek diskriminasi dapat dianggap baik apabila melebihi 0,20.

8. Efektifitas Pengecoh

Efektifitas pengecoh merupakan salah satu faktor yang dijadikan dasar dalam penelaahan soal. Hal ini dimaksudkan untuk mengetahui berfungsi tidaknya pilihan jawaban yang tersedia selain kunci jawaban. Suatu pilihan jawaban (pengecoh) dapat dikatakan berfungsi apabila pengecoh paling tidak dipilih oleh 5 % peserta tes/siswa.

Analisis pengecoh bertujuan mengetahui tingkat keberfungsian pengecoh yang disediakan. Dengan mengetahui penyebaran jawaban dapat diketahui:

(a) pengecoh yang terlalu menyolok kesalahannya sehingga tidak ada yang memilih,

(b) pengecoh yang menyesatkan, yakni pengecoh yang lebih banyak dipilih oleh siswa kelompok atas dari pada siswa kelompok bawah, dan

(c) pengecoh yang memiliki daya tarik bagi siswa kelompok bawah. Dengan demikian dapat dinyatakan bahwa pengecoh sebenarnya berfungsi untuk membedakan antara siswa yang berkemampuan tinggi dan rendah (Sujati, 2005).

Pengecoh dikatakan berfungsi efektif apabila banyak dipilih oleh siswa yang berkemampuan rendah Baik-buruknya butir soal tidak saja dilihat dari sudut tingkat kesukaran dan daya beda butir. Dalam banyak kasus, ketidakbaikan butir soal justru dipengaruhi oleh ketidakefektifan pengecoh yang disajikan. Pengecoh yang disajikan

ternyata tidak memiliki daya pikat sehingga tidak ada satupun peserta tes yang terjebak. Untuk itu perlu dilakukan analisis penyebaran frekuensi jawaban pada alternatifjawaban yang disediakan. Apabila dari 40 peserta tes misalnya, tidak ada satupun yang memilih alternatif jawaban tertentu yang berperan sebagai pengecoh, hal ini menunjukkan bahwa alternatif tersebut tidak efektif, sehingga perlu diganti atau direvisi (Sujati, 2005).

Selain cara klasik seperti yang telah diuraikan di atas, penganalisisan tes dan penggunaan data hasil analisis dapat dilakukan dengan pendekatan teori tes modern atau *Item Response Theory*. Dalam konsep *Item Response Theory* (IRT) setiap soal diwakili oleh *Item Characteristic Curve*. Pada Item Response Theory, parameter soal yang dihitung tergantung kepada model parameter yang dikehendaki. Model satu parameter atau biasa disebut *Model Rasch* hanya mempunyai satu parameter yaitu indeks kesukaran soal. Dua parameter model, yaitu tingkat kesukaran dan daya pembeda soal, sedangkan tiga parameter model terdiri dari tingkat kesukaran, daya pembeda dan *pseudo guessing*.

Konsep *Item Response Theory* (IRT) sangat berguna untuk memecahkan masalah-masalah dalam penyeleksian soa-soal dalam rangka mendisain suatu perangkat tes tertentu. Salah satu keunggulan utama *Item Response Theory* (IRT) dibandingkan teori Tes Klasik adalah dalam konsep *Item Response Theory* statistik soal seperti tingkat kesukaran, daya pembeda terletak dalam skala yang sama dengan kemampuan siswa yang diukur.

Item Response Theory dikembangkan melihat adanya beberapa kelemahan yang terdapat pada teori tes klasik. Dalam teori tes klasik, tingkat kesukaran soal biasanya dilaporkan dalam skala 0 sampai 1 dan didefinisikan atau dihubungkan dengan populasi pengikut tes atau proporsi banyaknya pengikut tes yang menjawab soal tertentu dengan benar. Sedangkan domain kemampuan siswa, walaupun juga dilaporkan dalam skala 0 sampai 1 tetapi didefinisikan atau dihubungkan dengan populasi soal dalam tes. Jadi terlihat bahwa yang menjadi masalah adalah skala tingkat kesukaran soal tidak sama atau berbeda definisinya dengan skala domain kemampuan siswa. Dengan kata lain, kalau kita berbicara mengenai tingkat kesukaran soal maka populasinya adalah kumpulan orang- orang pengikut tes, tetapi kalau kita berbicara mengenai domain kemampuan siswa, maka populasinya adalah kumpulan materi yang dites.

Perbedaan yang utama dari kedua model teori tes tersebut adalah terletak pada kelompok sampel. Tes yang telah disusun hanya dapat diterapkan secara baik kepada individu-individu yang kondisinya sama dengan kondisi kelompok yang dijadikan subjek dalam pengembangan tes. Seperti yang dikatakan Azwar, bahwa parameter-parameter item dalam teori klasik merupakan karakter item yang tergantung pada kelompok sampel yang digunakan untuk menghitungnya.

Dalam penggunaan alat ukur yang termasuk dalam *Item Response Theory*, Crocker dan Algina (1986: 322) merekomendasikan minimal 200 responden. Wright dan Stone merekomendasikan minimal panjang tes 20 butir dan 200 responden. Hullin, Lissak

dan Drasgow (1983: 99) merekomendasikan panjang tes 30 butir dan 500 responden untuk dua parameter model (L2P), 60 butir dan 1000 sampel untuk tiga parameter butir (L3P).

Berbeda dengan pendekatan IRT, untuk mendapatkan keakuratan alat ukur dalam mendeteksi keberbedaan fungsi butir (DIF), model klasik memiliki keuntungan tidak mempersyaratkan ukuran sampel yang besar. Bagi model klasik makin besar sampel makin baik. Berbeda dengan IRT, jika ukuran sample pada kelompok fokus kecil, hal ini merupakan problem. Keakuratan penggunaan alat ukur untuk mendeteksi keberbedaan fungsi butir pada IRT, tergantung model logistik parameter yang akan digunakan. Untuk tiga parameter logistik, maka DIF akan akurat jika menggunakan 1000 responden dan 60 butir (Alwi,2015).

Langkah-langkah dalam Proses Analisis Butir Soal (Analisis Kuantitatif)

1. Mengurutkan daftar nilai hasil ulangan/ujian dari yang terbesar sampai yang terkecil setiap kelas;

2. Daftar nilai yang telah diurutkan dibagi ke dalam tiga kelompok, yaitu kelompok pandai (upergroup), kelompok kurang (lowergroup), dan kelompok sedang (middlegroup);

3. Melakukan analisis pada kelompok pandai atau kelompok atas dan kelompok kurang atau kelompok bawah, sedangkan kelompok menengah kita biarkan. Umumnya diambil kelompok

atas dan bawah masing-masing 27% - 27%, (perbandingan tersebut tidak mutlak, tergantung pada kondisi jumlah objek yang akan dianalisis sehingga bisa 25% - 25%, 33% - 33% , dst);

4. Tiap soal ditabulasikan kemudian dijumlahkan pada setiap kelompok atas dan kelompok bawah.

- Menghitung Taraf Kesukaran

adalah:

$$P = \frac{\sum B}{N}$$

P : proporsi menjawab benar pada butir tertentu

$\sum B$: banyaknya peserta tes menjawab benar

N : jumlah peserta tes yang menjawab benar

- Menghitung Daya Pembeda

Rumus :

$$DB = \frac{BA - BB}{\frac{1}{2}N}$$

Keterangan :

• DB : daya pembeda

• BA : jumlah jawab benar tiap soal kel atas

• BB : jumlah jawab benar tiap soal kel bawah

• N : Jumlah testeekel atas dan kel bawah

Indeks hasil perhitungan diatas, dikonsultasikan dengan tabel tingkat daya pembeda, yaitu:

- $0{,}40 \leq DB \leq 1{,}00$ = soal diterima baik
- $0{,}30 \leq DB \leq 0{,}39$ = soal diterima tetapi perlu diperbaiki
- $0{,}20 \leq DB \leq 0{,}29$ = soal diperbaiki
- $DB \leq 0{,}19$ = soal tidak dipakai (dibuang)
- Analisis Fungsi Pengecoh

Suatu pilihan jawaban (pengecoh) dapat dikatakan berfungsi apabila pengecoh minimal dipilih oleh 5 % peserta tes/siswa.

- Menghitung Tingkat Validitas

1. Validitas Soal Bentuk Pilihan Ganda dengan Menggunakan Korelasi PointBiserial.

$$r_{pbis} = \frac{\bar{X}b - \bar{X}s}{SD}\sqrt{pq}$$

Rumus:

Keterangan:

Xb = rata-rata skor peserta didik yang menjawab benar

Xs = rata-rata skor peserta didik yang menjawab salah

$$\sqrt{\frac{\left[\sum\left(x - \bar{x}\right)^2\right]}{N}}$$

SD = simpangan baku skor total, dengan rumus SD =

p = proporsi jawaban benar terhadap semua jawaban siswa

q = 1 – p = proporsi jawaban salah terhadap semua jawaban siswa

2. Validitas Soal Bentuk Uraian dengan Menggunakan Korelasi ProductMoment.

Rumus:

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}$$

Keterangan :

rxy = koefisien korelasi productmoment

N = banyak sampel

X = skor butir

Y = skor total

- Menghitung Reliabilitas

1. Reliabilitas Soal Bentuk Pilihan Ganda dengan Menggunakan Rumus *Kuder Richadson 20* (KR-20)

Rumus *Kuder Richadson 20* (KR-20):

$$KR-20 = \frac{k}{k-1}\left[1 - \frac{\Sigma p(1-p)}{(SD)^2}\right]$$

Keterangan:

k = banyaknya butir soal

p = proporsi peserta tes yang menjawab benar

q = 1 – q

SD = varians total

2. Reliabilitas Soal Bentuk Uraian dengan Menggunakan Rumus Alfa Cronbach.

Rumus:

$$r = \frac{k}{k-1}\left[1 - \frac{SD_t^2 - \sum (SDi)^2}{(SD_t)^2}\right]$$

Keterangan:

r = koefisien reliabilitas seluruh tes

n = jumlah soal dalam tes

SD = varian skor-skor total pada tes

ΣSD = jumlah varian butir tes

Nilai Korelasi diatas konsultasikan dengan tabel kriteria korelasi koefisien, yaitu:

- $0{,}00 \leq r11 \leq 0{,}20$ = korelasi sangat rendah
- $0{,}20 \leq DB \leq 0{,}40$ = korelasi rendah
- $0{,}40 \leq DB \leq 0{,}70$ = korelasi cukup
- $0{,}70 \leq DB \leq 0{,}90$ = korelasi tinggi
- $0{,}90 \leq DB \leq 1{,}00$ = korelasi sangat tinggi (sempurna)

Manfaat Kegiatan Menganalisis Butir Soal Secara Kuantitatif

Berdasarkan pendapat yang diungkapkan oleh Anastasia dan Urbina (1997) dalam Suprananto (2012), analisis butir soal memiliki banyak manfaat, diantaranya yakni:

1. Membantu pengguna tes dalam mengevaluasi kualitas tes yang digunakan
2. Relevan bagi penyusunan tes informal seperti tes yang disiapkan guru untuk siswa dikelas
3. Mendukung penulisan butir soal yang efektif
4. Secara materi dapat memperbaiki tes di kelas
5. Meningkatkan validitas soal dan reliabilitas

Linn dan Gronlund (1995) dalam Suprananto (2012: 163), menambahkan bahwa pelaksanaan kegiatan analisis butir soal, biasanya didesain untuk menjawab pertanyaan-pertanyaan berikut:

1. Apakah fungsi soal sudah tepat?
2. Apakah soal telah memiliki tingkat kesukaran yang tepat?
3. Apakah soal bebas dari hal-hal yang tidak relevan?
4. Apakah pilihan jawabannya efektif?

Selain itu, data hasil analisis butir soal juga sangat bermanfaat sebagai dasar untuk:

- Diskusi tentang efisien hasil tes
- Kerja remedial
- Peningkatan secara umum pembelajaran di kelas

- Peningkatan keterampilan pada kontruksi tes.

Berdasarkan uraian di atas menunjukkan bahwa analisis butir soal memberikan manfaat:

1. Menentukan soal-soal yang cacat atau tidak berfungsi dengan baik
2. Meningkatkan butir soal melalui tiga komponen analisis yaitu, tingkat kesukaran, daya pembeda dan pengecoh soal
3. Merevisi soal yang tidak relevan degan materi yang diajarkan, ditandai dengan banyaknya anak yang tidak dapat menjawab butir soal tertentu.

DAFTAR PUSTAKA

Alwi, I. 2015. Kriteria Empirik Dalam Menentukan Ukuran Sampel Pada Pengujian Hipotesis Statistika Dan Analisis Butir. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *2*(2), 141-148

Ariyana, L.T. 2011. *Analisis Butir Soal Ulangan Akhir Semester Gasal IPA Kelas IX SMP di Grobogan, 1-142*

Fakhrun, Nisya, 2018. *Analisis Kualitas Butir Soal Ujian Semester Genap Pada Siswa Kelas X Mata Pelajaran Prakarya Dan Kewirausahaan T.A. 2016/2017 Di Sekolah Menengah Atas Negeri 5 Pekanbaru.* Skripsi Thesis, Universitas Islam Negeri Sultan Syarif Kasim Riau.

Karim, A. 2018. Analisis Kualitas Soal Perlombaan Matematika Tingkat SMA. *Titian Ilmu: Jurnal Ilmiah Multi Sciences*, *10*(1), 1-8.

Subali, B. 2014. Analisis Soal Baik Kualitatif Maupun Kuantitatif. In *Disajikan pada Kegiatan Workshop Item Development Dosen Poltekes Kebidanan Politeknik Kesehatan Surakarta tanggal 18-19 Agustus 2014 di Griya Persada Conventional Hotel & Resort, Jl Boyong Kaliurang Barat.*

Sujati, H. 2005. Menganalisis Kualitas Tes Sebagai Salah Satu Kompetensi Guru Profesional. *Jurnal Ilmiah Guru 'COPE'*. 1 (IX), 39-46

Surapranata, S. 2004. *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes*. Bandung: Rosdakarya.

Suprananto, Kusaeri. 2012. *Pengukuran dan Penelitian Pendidikan*. Yogyakarta: Graha Ilmu.