

Penggunaan Aplikasi Software Iteman (*Item and Test Analysis*) untuk Analisis Butir Soal Pilihan Ganda Berdasarkan Teori Tes Klasik

Use of Application of Software Iteman (*Item and Test Analysis*) to Analysis of Multiple Choice Item Based upon Classical Test Theory

Didik Setyawarno

**Program Studi S1 Pendidikan IPA, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta**

Kampus Pusat UNY, Jl. Colombo No.1 Yogyakarta 55281, Indonesia

E-mail: didiksetyawarno@uny.ac.id

Abstrak

Artikel ini bertujuan mengkaji teori tes klasik serta aplikasi Iteman untuk menganalisis butir soal berbentuk pilihan ganda (PG) yang diperuntuk untuk guru atau mahasiswa calon guru baik bidang IPA atau Non-IPA. Metode yang digunakan dalam kajian ini adalah studi pustaka atau literatur baik buku dan jurnal baik dari dalam maupun luar negeri. Hasil kajian diperoleh konsep bahwa konsep dasar teori tes klasik menjadi landasan munculnya formula-formula penting dalam analisis butir soal yang meliputi validitas, reliabilitas, distribusi jawaban, tingkat kesukaran, dan daya pembeda. Iteman (*Item and Test Analysis*) merupakan aplikasi program komputer yang dapat digunakan untuk menganalisis butir soal berbentuk pilihan ganda berdasarkan pendekatan teori tes klasik. Setelah membaca artikel ini, diharapkan guru atau mahasiswa calon guru mempunyai pemahaman tentang konsep dasar teori tes klasik, validitas, reliabilitas, distribusi jawaban, tingkat kesukaran, dan daya pembeda, serta mempunyai kemampuan menggunakan aplikasi Iteman untuk analisis butir soal bentuk pilihan ganda sehingga memperoleh soal yang berkualitas yang digunakan untuk evaluasi pembelajaran.

Kata Kunci: Iteman, Soal Pilihan Ganda, dan Teori Tes Klasik.

PENDAHULUAN

Setiap proses pembelajaran yang telah dilakukan oleh guru selama waktu tertentu harus diketahui sejauhmana pembelajaran tersebut mampu meningkatkan kemampuan atau memberikan nilai tambah bagi siswanya. Tes merupakan salah satu alat untuk melakukan pengukuran, yaitu alat untuk mengumpulkan informasi karakteristik suatu objek. Objek ini bisa berupa kemampuan peserta didik, sikap, minat, maupun motivasi (Rosana, 2014:33). Tes tersebut dikalangan dunia pendidikan di sebut tes hasil belajar (THB).

Tes hasil belajar adalah salah satu alat ukur yang paling banyak digunakan untuk mengetahui hasil belajar seseorang dalam proses belajar-mengajar atau suatu program pendidikan. Istilah tes sering diartikan sebagai

alat penilaian yang sifatnya spesifik, dan secara sederhana diungkapkan dalam bentuk pertanyaan yang mengungkap tingkah laku, potensi maupun dikaitkan dengan hasil belajar (Suparwoto, 2005:4). Tes merupakan alat ukur yang standar dan obyektif sehingga dapat digunakan secara meluas untuk mengukur dan membandingkan keadaan psikis atau tingkah laku individu. Sebagai contoh, setiap tes yang diberikan oleh guru kepada siswa menuntut keharusan adanya respon dari subyek (orang yang dites) yang sedang dicari informasinya.

Sebagai alat untuk mengukur kemampuan siswa setelah mengikuti kegiatan pendidikan selama selang waktu tertentu, maka eksistensi tes menjadi sangat penting (Lababa, 2008:30). Sebuah tes yang baik, akan bisa mengungkapkan keadaan sebenarnya dari siswa, dan tes yang tidak baik tidak akan bisa

mengungkap apa kemampuan sebenarnya siswa. Dengan demikian identifikasi terhadap setiap butir item soal perlu dilakukan sebelum butir soal tersebut digunakan dalam pengukuran kemampuan siswa. Identifikasi dan analisis butir item soal perlu dilakukan secara rutin oleh guru untuk perbaikan, pembenahan, dan penyempurnaan kembali terhadap butir-butir soal. Dengan cara ini, guru pada masa-masa yang akan datang tes hasil belajar yang disusun atau dirancang betul-betul dapat menjalankan fungsinya sebagai alat pengukur hasil belajar yang memiliki kualitas yang tinggi. Selain itu, analisis item butir soal dilakukan untuk mengetahui berfungsi atau tidaknya sebuah soal.

Analisis item butir soal pada umumnya dilakukan melalui dua cara, yaitu analisis kualitatif dan analisis kuantitatif (Subali, 2014). Analisis butir soal secara kualitatif dilaksanakan berdasarkan kaidah penulisan soal (tes tertulis, perbuatan, dan sikap). Penelaahan ini biasanya dilakukan sebelum soal digunakan atau diujikan. Aspek yang diperhatikan dalam penelaahan secara kualitatif mencakup aspek materi, konstruksi, bahasa atau budaya, dan kunci jawaban. Analisis item butir soal secara kuantitatif merupakan penelaahan butir soal didasarkan pada bukti empirik. Salah satu tujuan utama pengujian butir-butir soal secara empirik adalah untuk mengetahui sejauh mana masing-masing butir soal membedakan antara siswa yang tinggi kemampuannya dengan siswa yang rendah kemampuannya.

Artikel ini akan memfokuskan pembahasan atau pengkajian terkait dengan konsep dasar teori tes klasik yang dijabarkan menjadi berbagai formula penting dalam analisis butir soal yang meliputi validitas, reliabilitas, distribusi jawaban, tingkat kesukaran, dan daya pembeda. Selain itu, artikel juga mengkaji aplikasi Iteman untuk analisis butir soal bentuk pilihan ganda. Setelah membaca artikel ini, guru atau mahasiswa calon guru mempunyai pemahaman tentang konsep dasar teori tes klasik, validitas, reliabilitas, tingkat kesukaran, dan daya pembeda, serta mempunyai kemampuan menggunakan aplikasi Iteman untuk analisis butir soal bentuk pilihan ganda.

METODE

Artikel ini disusun dengan menggunakan metode studi pustaka/literatur baik buku dan jurnal dari dalam maupun luar negeri dengan pendekatan deskriptif. Studi pustaka dilakukan dengan cara mengambil gagasan pokok setiap buku atau jurnal kemudian dijabarkan secara runtut dan bersifat aplikatif.

PEMBAHASAN

Hasil kajian dari berbagai sumber literature baik buku maupun jurnal, maka artikel ini akan memaparkan hasil dan pembahasan meliputi: teori tes klasik, validitas, reliabilitas, distribusi jawaban, tingkat kesukaran, dan daya pembeda, serta penggunaan aplikasi Iteman untuk analisis butir soal bentuk pilihan ganda sebagai berikut.

A. Teori Tes Klasik

Classical Test Theory yang dalam bahasa Indonesia sering disebut dengan teori tes klasik merupakan salah satu teori pengukuran yang tertua didunia pengukuran behavioral. Mamun Ali Naji Qasem menyatakan bahwa: *classical test theory introduces three concepts test score, true score, and error score. Within that theoretical framework, models of various forms have been formulated* (Qasem, 2013:78). Teori tes klasik memperkenalkan tiga konsep yaitu: skor tes, skor yang benar, dan skor galat. Model berbagai bentuk telah dirumuskan berdasarkan teori tersebut. Sebagai contoh, kita sering merujuk pada model uji klasik, yaitu model linear sederhana di mana postulat-postulas yang menghubungkan skor tes tampak (X) dengan jumlah dua variabel yang tidak dapat diamati, skor murni (T), dan skor kesalahan (E), yaitu, $X = T + E$. Ada tujuh macam asumsi yang ada dalam teori tes klasik ini. Allen & Yen (1979:57) menguraikan asumsi-asumsi teori klasik sebagai berikut.

1. Terdapat hubungan antara skor tampak (*observed score*) yang dilambangkan dengan huruf X, skor murni (*true score*) yang dilambangkan dengan T dan skor kesalahan (error) yang dilambangkan dengan E. Dalam bahasa matematika dapat dilambangkan dengan $X = T + E$.
2. Skor murni (T) merupakan nilai harapan ϵ (X). Dengan demikian skor murni adalah nilai rata-rata skor perolehan teoretis sekiranya dilakukan pengukuran berulang-ulang (sampai

tak terhingga) terhadap seseorang dengan menggunakan alat ukur.

3. Tidak terdapat korelasi antara skor murni dan skor pengukuran pada suatu tes yang dilaksanakan ($\rho_{et} = 0$).
4. Korelasi antara kesalahan pada pengukuran pertama dan kesalahan pada pengukuran kedua adalah nol ($\rho_{e1e2} = 0$).
5. Jika terdapat dua tes untuk mengukur atribut yang sama maka skor kesalahan pada tes pertama tidak berkorelasi dengan skor murni pada tes kedua ($\rho_{e1t2} = 0$).
6. Dua perangkat tes dapat dikatakan sebagai tes-tes yang paralel jika skor-skor populasi yang menempuh kedua tes tersebut mendapat skor murni yang sama ($T = T'$) dan varian skor-skor kesalahannya sama $\sigma_e^2 = \sigma_e'^2$.
7. Jika dua perangkat tes mempunyai skor-skor perolehan X_{t1} dan X_{t2} yang memenuhi asumsi 1 sampai 5 dan apabila untuk setiap populasi subyek $X_1 = X_2 + C_{12}$, dimana C_{12} adalah sebuah bilangan konstanta, maka kedua tes itu disebut tes yang paralel.

Asumsi-asumsi teori klasik sebagaimana disebutkan di atas memungkinkan untuk dikembangkan dalam rangka pengembangan berbagai formula yang berguna dalam melakukan pengukuran psikologis. Daya beda, indeks kesukaran, efektifitas distraktor, reliabilitas dan validitas adalah formula penting yang disarikan dari teori tes klasik (Lababa, 2008:30).

B. Validitas

Ellen A. Drost menyatakan bahwa “*validity is concerned with the meaningfulness of research components*” (Drost, 2012:114). Dalam hal yang sama Ronald Jay Cohen menyatakan bahwa “*validity, as applied to a test, is a judgment or estimate of how well a test measures what it purports to measure in a particular context. More specifically, it is a judgment based on evidence about the appropriateness of inferences drawn from test scores*” (Cohen, 2009:172). Pernyataan tersebut menyatakan bahwa validitas berkaitan dengan ketepatan keberartian komponen penelitian. Pengertian tersebut jika dikaitkan dengan butir soal sebagai alat ukur, maka validitas merupakan ketepatan alat ukur dengan hal yang diukur. Ada beberapa jenis validitas alat ukur, yaitu: validitas isi, validitas konstruksi, validitas ramalan, dan validitas sama saat. Validitas butir soal pilihan ganda/data dikotomi, jika butir soal di jawab benar bernilai 1 dan salah bernilai 0 di hitung dengan menggunakan teknik *Korelasi Point Biserial* (Koyan, 2012:56). Rumus yang digunakan untuk menghitung Korelasi Point Biserial sebagai berikut.

$$r_{pbi} = \frac{M_p - M_t}{s_t} \sqrt{\frac{p}{q}}$$

Keterangan:

- r_{pbi} = koefisien korelasi *point biserial*
 M_p = rerata skor dari subjek yang menjawab betul bagi butir yang dicari validitasnya
 M_t = rerata skor total
 s_t = standar deviasi dari skor total

- p = proporsi peserta didik yang menjawab betul (banyaknya peserta didik yang menjawab betul dibagi dengan jumlah seluruh peserta didik)
 q = proporsi peserta didik yang menjawab salah ($q = 1 - p$).

Suatu butir tes dinyatakan valid jika r hitung lebih besar daripada r tabel dengan taraf signifikansi atau taraf kekeliruan 5% ($r_{hit} > r_{tab}$ dengan taraf signifikansi 5%).

C. Reliabilitas

Reliabilitas berasal dari kata *reliability* yang dapat diartikan hal yang dapat dipercaya. Dalam hal yang sama, Drost menyatakan bahwa “*reliability is a major concern when a psychological test is used to measure some attribute or behaviour*” (Drost, 2012:106). Pengertian tersebut menyatakan bahwa reliabilitas adalah keterpercayaan, keterandalan, keajegan, konsistensi, atau kestabilan. Ada beberapa jenis reliabilitas, yaitu: (1) konsistensi internal, (2) stabilitas, dan (3) ekuivalen. Reliabilitas konsistensi internal alat ukur dapat dihitung dengan menggunakan rumus Koefisien Alpha-Cronbach, Kuder-Richardson (KR-20 atau KR-21), dan Teknik Belah Dua. Suparwoto menyatakan bahwa Koefisien AlphaCronbach dapat dimanfaatkan untuk analisis butir soal dengan skor benar +1 dan salah 0, atau dengan skor 1, 2, 3 berurutan dan cara ini merupakan upaya menetapkan koefisien reliabilitas instrumen/tes yang mengacu pada konsep *internal consistency* (Suparwoto, 2005:55). Rumus yang digunakan untuk menghitung Koefisien Alpha-Cronbach sebagai berikut.

$$r_{1.1} = \left(\frac{k}{k-1} \right) \left(\frac{SD_t^2 - \sum SD_i^2}{SD_t^2} \right)$$

Keterangan:

$r_{1.1}$ = koefisien reliabilitas perangkat tes

k = banyak butir tes

SD_i^2 = varians skor tiap butir (item)

SD_t^2 = varian skor total

Tingkat reliabilitas instrumen dapat diketahui dari nilai r sebagai berikut.

$r \leq 0,20 \Rightarrow$ sangat rendah

$0,20 < r \leq 0,40 \Rightarrow$ rendah

$0,40 < r \leq 0,60 \Rightarrow$ sedang

$0,60 < r \leq 0,80 \Rightarrow$ tinggi

$0,80 < r \leq 1,00 \Rightarrow$ sangat tinggi

D. Distribusi Jawaban

Berfungsi tidaknya jawaban yang tersedia dalam soal pilihan ganda, maka digunakan analisis distribusi jawaban. Satu soal pilihan ganda, suatu pilihan jawaban (pengecoh) dapat dikatakan berfungsi apabila pengecoh paling tidak dipilih oleh 5 % peserta tes/siswa dan lebih banyak dipilih oleh kelompok siswa yang belum paham materi (Depdikbud, 2005:14).

E. Tingkat Kesukaran

Indeks kesukaran butir sebagaimana dinyatakan oleh Allen & Yen (1979:120) adalah "*proportion of examinees who get that item correct*". Pernyataan tersebut menjelaskan bahwa tingkat kesukaran butir tes merupakan bilangan yang menunjukkan proporsi peserta ujian (testee) yang dapat menjawab betul butir soal tersebut. Sedangkan tingkat kesukaran perangkat tes adalah bilangan yang

menunjukkan rata-rata proporsi *testee* yang dapat menjawab seluruh (perangkat) tes tersebut (Koyan, 2012:62). Rumus yang digunakan untuk menentukan tingkat kesukaran sebagai berikut.

$$P = \frac{nB}{n}$$

Keterangan:

P = tingkat kesukaran butir tes

nB = banyaknya subyek yang menjawab soal dengan betul

n = jumlah subyek (*testee*) seluruhnya

Sebagaimana dinyatakan oleh Allen & Yen, tingkat kesukaran butir soal yang baik adalah 0,3 sampai 0,7. Butir dengan tingkat kesulitan dibawah 0,3 dianggap butir soal yang sukar sedangkan jika indeksnya diatas 0,7, butir soal tersebut dianggap mudah (Allen & Yen, 1979:121). Dengan demikian kriteria tingkat kesukaran (P) dapat dituliskan sebagai berikut.

0,00 - 0,30 \Rightarrow soal tergolong sukar

0,31 - 0,70 \Rightarrow soal tergolong sedang

0,71 - 1,00 \Rightarrow soal tergolong mudah

F. Daya Pembeda

Daya beda (diskriminasi) suatu butir tes adalah kemampuan suatu butir untuk membedakan antara peserta tes yang berkemampuan tinggi dan berkemampuan rendah (Lababa, 2008:32). Pengertian tersebut menjelaskan bahwa daya beda butir tes merupakan kemampuan butir tes tersebut membedakan antara *testee* kelompok atas (tinggi) dan *testee* kelompok bawah (lemah).

Rumus untuk menghitung daya beda butir tes adalah sebagai berikut.

$$D_B = \frac{nB_A}{n_A} - \frac{nB_B}{n_B}$$

Keterangan:

nB_A = jumlah subyek yang menjawab betul pada kelompok atas

nB_B = jumlah subyek yang menjawab betul pada kelompok bawah

n_A = jumlah subyek kelompok atas

n_B = jumlah subyek kelompok bawah

Kriteria Daya Beda (D_B) sebagai berikut.

0,40 – 1,00 => Soal baik

0,30 – 0,39 => Soal diterima dan diperbaiki

0,20 – 0,29 => Soal diperbaiki

0,00 – 0,19 => Soal ditolak

Jika " D_B " negatif, soal tersebut sangat buruk dan harus dibuang.

G. Soal Pilihan Ganda

Soal pilihan ganda merupakan bentuk soal yang jawabannya dapat dipilih dari beberapa kemungkinan jawaban yang telah disediakan. Konstruksinya terdiri dari pokok soal dan pilihan jawaban (Depdiknas, 2007:12). Pilihan jawaban terdiri atas kunci dan pengecoh. Kunci jawaban harus merupakan jawaban benar atau paling benar sedangkan pengecoh merupakan jawaban tidak benar, namun daya pengecohnya harus berfungsi, artinya siswa memungkinkan memilihnya jika tidak menguasai materinya.

Soal pilihan ganda merupakan bagian dari tes objektif. Sudjana menyatakan bahwa

ada empat macam tes objektif, yaitu tes jawaban benar-salah (*true-false*), pilihan ganda (*multiple choice*), isian (*completion*), dan penjodohan (*matching*) (Sudjana, 1987:6). Secara umum, setiap soal pilihan ganda terdiri dari pokok soal (*stem*) dan pilihan jawaban (*option*). Pilihan jawaban terdiri atas kunci jawaban dan pengecoh (*distractor*). Tes pilihan ganda terdiri dari sebuah pernyataan atau kalimat yang belum lengkap yang kemudian diikuti oleh sejumlah pernyataan atau bentuk yang dapat untuk melengkapinya. Dari sejumlah "pelengkap" tersebut, hanya satu yang tepat sedang yang lain merupakan pengecoh (*distractors*)

Tes pilihan ganda merupakan suatu bentuk tes yang paling banyak dipergunakan dalam dunia pendidikan. Bentuk ini sangat tepat digunakan untuk ujian berskala besar yang hasilnya harus segera diumumkan, seperti ujian nasional, ujian akhir sekolah, dan ujian seleksi masuk perguruan tinggi. Kelebihan dari penggunaan soal pilihan ganda sebagai alat ukur/tes diantaranya 1) lebih mudah dan cepat cara memeriksanya karena dapat menggunakan kunci tes bahkan alat-alat hasil kemajuan teknologi, 2) pemeriksaannya dapat diserahkan orang lain, dan 3) dalam pemeriksaan, tidak ada unsur subjektif yang memengaruhi (Rosana, 2014:175).

H. Analisis Butir Soal Pilihan Ganda dengan Aplikasi Iteman

Iteman (*Item and Test Analysis*) adalah perangkat lunak komputer (*software*) yang dibuat khusus untuk menganalisa butir soal atau suatu tes yang dilakukan. Iteman

merupakan analisis butir empirik dengan model pendekatan klasik yang berguna untuk menentukan kualitas butir soal atau sebuah tes. Hasil dari analisis butir soal meliputi tingkat kesukaran, daya beda, dan statistik penyebaran jawaban. Selain menghasilkan statistik butir soal/tes, program ini juga menghasilkan statistik tes yang meliputi realibilitas tes, kesalahan pengukuran atau standard error dan distribusi skor (Rosana & Setyawarno, 2016: 19). Aplikasi program ini dapat digunakan sebagai berikut.

1. Menganalisis data file (format ASCII) jawaban butir soal yang dihasilkan melalui manual *entry data* atau dari mesin *scanner*.
2. Menskor dan menganalisis data soal pilihan ganda dan skala likert (maks 250 butir).
3. Menganalisis sebuah tes yang terdiri dari 10 skala (subtes) dan memberikan informasi tentang validitas setiap butir.

Berikut langkah-langkah untuk menjalankan aplikasi *Iteman*.

1. Memasukan data (jawaban peserta tes dan kunci jawaban tes) dengan contoh format penulisan sebagai Lampiran 1.
2. Menyimpan file *notepad* tersebut dalam satu folder bersama program *iteman.exe*. Misalkan data yang baru selesai dimasukkan itu diberi nama "Tes1" dengan ekstensi file "txt".
3. Jalankan aplikasi *Iteman* dengan perintah sebagaimana Lampiran 2.
 - a. *Enter the name of the input file:* ketik nama file yang akan

dianalisis, misalnya "Tes1.txt" kemudian tekan enter.

- b. *Enter the name of the output file:* ketik nama file output (hasil) yang dikehendaki, misal "hsltes1.txt" lalu tekan enter.
- c. *Do you want the score written to a file?:* ketik Y bila dikehendaki hasil analisis direkam, ketik N bila hasil analisis tidak direkam. Bila diketik Y maka akan muncul *Enter the name of the score file:* ketik nama file untuk hasil skor, misal "scrtes1.txt" lalu tekan enter.
- d. Dalam waktu beberapa detik, akan muncul tampilan di folder (hasil ada di folder yang sama dengan file yang akan dianalisis) hasil analisis dengan *Iteman* tadi.
- e. Hasil analisis butir soal pilihan ganda sebagaimana Lampiran 3.

Sebelum menafsirkan hasil analisis butir soal sebagaimana contoh pada Lampiran 3, perlu untuk memahami beberapa istilah dalam tabel hasil analisis di atas baik kolom pertama maupun kolom kedua.

1. *Seq. No* adalah nomor urut butir soal dalam file data.
2. *Scala-item* adalah nomor urut butir soal dalam skala (tes/subtes)
3. *Prop. Correct* adalah proporsi siswa(peserta tes) yang menjawab benar butir soal. Nilai ekstrim (mendekati nol atau satu) menunjukkan bahwa butir soal tersebut terlalu sukar atau terlalu mudah untuk peserta tes. Indeks ini

disebut juga indeks tingkat kesukaran soal secara klasikal.

4. *Biser* adalah indeks daya pembeda soal dengan menggunakan koefisien korelasi biserial. Nilai positif menunjukkan bahwa peserta tes yang menjawab benar butir soal, mempunyai skor yang relatif tinggi dalam tes/skala tersebut. Sebaliknya nilai negatif menunjukkan bahwa peserta tes yang menjawab benar butir soal, memperoleh skor yang relatif rendah dalam tes/skala tersebut. Untuk statistik pilihan jawaban (alternative) korelasi biserial negatif sangat tidak dikehendaki untuk kunci jawaban dan sangat dikehendaki untuk pilihan jawaban yang lain (pengecoh).
5. *Point-biser* adalah juga indeks daya pembeda soal dan pilihan jawaban (alternatif) dengan menggunakan koefisien korelasi point-biserial. Penafsirannya sama dengan statistik biserial.
6. Statistik pilihan jawaban (alternative) memberikan informasi yang sama dengan statistik butir soal. Perbedaananya adalah bahwa statistik pilihan jawaban dihitung secara terpisah. Untuk setiap pilihan jawaban dan didasarkan pada dipilih tidaknya alternatif tersebut, bukan pada benarnya jawaban. Tanda (*) yang muncul di sebelah kanan hasil analisis menunjukkan kunci jawaban.

Istilah yang perlu dipahami untuk membaca hasil analisis sebagai berikut.

1. *N of items* adalah jumlah butir soal dalam tes.
2. *N of examines* adalah jumlah peserta tes yang digunakan dalam analisis.
3. *Mean* adalah skor rata-rata peserta tes.
4. *Variance* adalah varian dari distribusi skor peserta tes yang memberikan gambaran tentang sebaran skor peserta tes.
5. *Std. Dev* adalah deviasi standar dari distribusi skor peserta tes.
6. *Skew* adalah kemiringan distribusi skor peserta tes yang memberikan gambaran tentang bentuk distribusi skor peserta tes.
7. *Kurtosis* adalah puncak distribusi skor yang menggambarkan kelandaian distribusi skor dibanding dengan distribusi normal.
8. *Minimum* adalah skor terendah peserta tes dalam tes/skala tersebut.
9. *Maximum* adalah skor tertinggi peserta tes dalam tes/skala tersebut.
10. *Median* adalah skor tengah dimana 50% skor berada pada atau lebih rendah dari skor tersebut.
11. *Alpha* adalah koefisien reliabilitas alpha untuk tes/skala tersebut yang merupakan indeks homogenitas tes/skala. Koefisien alpha bergerak dari 0,0 sampai 1,0.
12. *SEM* adalah kesalahan pengukuran standar untuk setiap tes/skala. SEM merupakan estimat dari deviasi standar kesalahan pengukuran dalam skor tes.
13. *Mean P* adalah rata-rata tingkat kesukaran semua butir soal dalam tes

secara klasikal dihitung dengan cara mencari rata-rata proporsi peserta tes yang menjawab benar untuk semua butir soal dalam tes/skala.

14. *Mean item-Tot* nilai rata-rata indeks daya pembeda dari semua soal dalam tes/skala yang diperoleh dengan menghitung nilai rata-rata point biserial dari semua soal dalam tes/skala.
15. *Mean-Biserial* adalah juga nilai rata-rata indeks daya pembeda yang diperoleh dengan menghitung nilai rata-rata korelasi biserial dari semua butir soal dalam tes/skala.
16. *Scale intercorrelation* adalah indeks korelasi antara skor-skor peserta tes yang diperoleh dari setiap subtes/subskala.

PENUTUP

Asumsi-asumsi teori klasik merupakan dasar pengembangan berbagai formula yang berguna dalam melakukan pengukuran butir soal mencakup daya beda, indeks kesukaran, efektifitas distraktor, reliabilitas, dan validitas. Validitas butir soal merupakan ketepatan butir soal tersebut dengan aspek yang diukur dalam penilaian hasil belajar. Teknik Korelasi Point Biserial merupakan salah satu teknik untuk menentukan butir soal tersebut valid atau tidak. Reliabilitas adalah keterpercayaan, keterandalan, keajegan, konsistensi, atau kestabilan. Reliabilitas butir soal dapat dihitung salah satunya dengan menentukan koefisien Alpha-Cronbach. Tingkat kesukaran

butir tes merupakan bilangan yang menunjukkan proporsi peserta ujian (*testee*) yang dapat menjawab betul butir soal tersebut. Daya beda (diskriminasi) suatu butir tes adalah kemampuan suatu butir untuk membedakan antara peserta tes yang berkemampuan tinggi dan berkemampuan rendah. Aplikasi Itemandapat digunakan untuk analisis butir soal dengan pendekatan teori klasik tes.

Daftar Pustaka

- Allen & Yen. 1979. *Introduction to Measurement Theory*. Belmont, California: Wadsworth, Inc.
- Arif Pratisto. 2004. *Cara Mudah Mengatasi Masalah Statistik dan Rancangan Percobaan dengan SPSS*. Jakarta: Gramedia.
- Bambang Subali. Analisis Soal Baik Kualitatif Maupun Kuantitatif. *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*.
- Carmines, EG & Zeller, RA. 1979. *Reliability and Validity Assessment*. Beverly Hills, California: Sage Publications, Inc.
- Cohen, R.J. 2009. *Psychological Testing and Assessment: An Introduction to Tests and Measurement 7th Edition*. New York, USA: Mc-Graw-Hill.
- Dadan Rosana. 2014. *Evaluasi Pembelajaran Sains*. Yogyakarta: UNY Press.
- Dadan Rosana & Didik Setyawarno. 2016. *Statistik Terapan untuk Penelitian Bidang Pendidikan*. Yogyakarta: FMIPA UNY.
- Depdikbud. 2005. *Panduan Analisis Butir Soal*. Jakarta:Depdikbud.
- Pusat Penilaian Pendidikan. 2005. *Panduan Penulisan Soal Pilihan Ganda*. Jakarta: Balitbang-Depdikbud.
- Drost, E.A. Validity and Reliability in Social Science Research. *Education Research and Perspectives, Vol.38, No.1, 2012*.
- Gaur, A.S. 2006. *Statistical Methods for Practice and Research*. Delhi, India: SAGE Publications.

- I Wayan Koyan. 2012. *Konstruksi Tes*. Bali: Undiksha Press.
- Junaidi Lababa. Analisis butir soal dengan teori tes klasik: Sebuah pengantar. *Jurnal Iqra' Volume 5 Januari - Juni 2008*.
- Qasem, M.A.N. A Comparative Study of Classical Theory (Ct) and Item Response Theory (Irt) In Relation To Various Approaches of Evaluating the Validity and Reliability of Research Tools. *IOSR Journal of Research & Method in Education (IOSR-JRME) Volume 3, Issue 5, pp 77-81, 2013..*
- Suparwoto. 2005. *Penilaian Pencapaian Hasil Belajar Fisika*. Yogyakarta: FMIPA UNY.
- Zoghi, M and & Valipour, V. A Comparative Study of Classical Test Theory And Item Response Theory In Estimating Test Item Parameters In A Linguistics

Test. *Indian Journal of Fundamental and Applied Life Sciences, Vol. 4 (S4), pp. 424-435, 2014.*

Biodata Penulis

Penulis adalah Didik Setyawarno, salah satu dosen S1 Pendidikan IPA di FMIPA UNY sejak 2015 dengan bidang keahlian penilaian dan evaluasi pembelajaran sains. Penulis merupakan alumni dari S1 Pendidikan Fisika dan S2 Pendidikan Sains (Konsentrasi Fisika) di perguruan tinggi tempat mengabdikan yaitu UNY.

Lampiran Artikel

Lampiran 1. Data Butir Soal yang akan dianalisis dengan Iteman (ekstensi file .txt)

```
035 O N 04
CACDABDACADACADACADCDCCABACADBBABD => Kunci Jawaban
444444444444444444444444444444444444 =>Jumlah Pilihan Jawaban
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY =>Kode Analisis
001 CACDOAOACOBACAOBCOACDBCACOCDDBBBOBO
002 CACDAADACACCACADACAACDCCABACADBBABD
003 CACDAADACACCADADACAACDCCABACDDBBABC
004 CACDCADACACCADACADODCCADACDDBBABA
005 CACDAADACACCACADACAACDCCABACADBBABD
006 CACDAADACBAOBCAACODCAAAABAODBABAB
.....
dan seterusnya sampai peserta tes terakhir .....
```

Lampiran 2. Perintah Menjalankan Aplikasi Iteman

```
MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation

Beta-Test Version--Univ. of Pittsburgh

Item and Test Analysis Program -- ITEMAN (tm) Version 3.00

Enter the name of the input file: Test.txt
Enter the name of the output file: Hasil.txt
Do you want the scores written to a file? (Y/N) Y
Enter the name of the score file: Score.txt
```

Lampiran 3. Hasil Analisis Iteman Setiap Butir Soal

Item analysis for data from file Test.txt

Page 1

Item Statistics				Alternative Statistics			
Seq. No.	Scale	Prop. Correct	Point Biser.	Prop. Alt.	Endorsing	Biser.	Point Key
1	0-1	0.905	0.901	0.520	A	0.000	-9.000 -9.000
	B	0.000	-9.000	-9.000			
	C	0.905	0.901	0.520 *			
			D	0.048	-0.028	-0.013	
			Other	0.048	-1.000	-0.703	
2	0-2	0.905	1.000	0.660	A	0.905	1.000 0.660 *
	B	0.000	-9.000	-9.000			
			C	0.048	-0.443	-0.206	
			D	0.000	-9.000	-9.000	
			Other	0.048	-1.000	-0.703	

.....
dan seterusnya sampai peserta tes terakhir

Lampiran 4. Hasil Analisis *Scale Statistics*

```
N of Items      35
N of Examinees  21
Mean            25.476
Variance        65.583
Std. Dev.       8.098
Skew            -1.494
Kurtosis        2.118
Minimum         0.000
Maximum         33.000
Median          29.000
Alpha           0.941
SEM             1.964
Mean P          0.728
Mean Item-Tot.  0.600
Mean Biserial   0.834
```