Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 2, December 2017 (12-23)
Online: http://journal.uny.ac.id/index.php/jpep

ANALISIS KUALITAS SOAL DI PERGURUAN TINGGI BERBASIS APLIKASI *TAP*

Akbar Iskandar [1] *****, *Muhammad Rizal* [1]
[1]STMIK AKBA Makassar
[1]Jln. Perintis Kemerdekaan, Tamalanrea, Tamalanrea Jaya, Makassar, Sulsel, Indonesia
***** Corresponding Author. Email: akbar.iskandar06@gmail.com

Abstrak

Tujuan penelitian ini adalah untuk menemukan butir instrumen yang berkualitas berdasarkan tingkat validitas, reliabilitas, tingkat kesukaran, daya beda dan pengecoh. Jenis penelitian ini adalah jenis penelitian ex post facto. Objek penelitian semua jawaban hasil tes calon mahasiswa baru tahun 2014-2016. Data dikumpulkan dengan metode Observasi, wawancara dan dokumentasi. Teknik analisis data secara kualitatif dan kuantitatif. Hasil penelitian menunjukkan bahwa untuk validitas isi ditemukan nilai validitas isi (vi) sebesar 0,42 termasuk kategori sedang. Selanjutnya tampak bahwa nilai koefisien reliabilitas instrumen sebesar 0,514. Jumlah butir soal yang sukar sebanyak 57,5%, kategori sedang sebanyak 42,5% dan tidak terdapat soal kategori mudah. Selain itu, butir soal yang memiliki daya beda sangat baik sebanyak 5%, baik sebanyak 20%, perlu revisi sebanyak 13,75%, tidak baik sebanyak 61,25%. Sedangkan option yang tidak berfungsi dengan baik pada saat dijadikan sebagai pengecoh sebanyak 5 butir soal, akan tetapi terdapat 40 butir soal yang harus direvisi karena option pengecoh malah dianggap sebagai kunci jawaban oleh peserta yang pintar.

Kata kunci: *validitas, reliabilitas, tingkat kesukaran, daya beda,pengecoh,TAP*

ANALYSIS OF EXAMINATION INSTRUMENTS QUALITY AT UNIVERSITY BASED ON TAP APPLICATION

Abstract

This study is aimed at finding the items of quality instruments based on the level of validity, reliability, difficulty, differentiation, and distraction. This study was an ex post facto research. The object of research was all the answers of enrollment test results of prospective students in 2014-2016. The data were collected by using observation, interview, and documentation methods, and the data were analyzed using both qualitative and quantitative technique. The results of the research show that the value of the content validity found is 0.42, which belongs to medium category. Furthermore, it appears that the value of the instrument reliability coefficient is 0.514. The number of the question items which belongs to difficult category is 57.5%, medium category is 42.5%, and no item is in easy category. In addition, the percentage of the question items that belong to the category of very good, good, need a little revision, and not good, in terms of the items' differentiation, are respectively 5%, 20%, 13.75%, and 61.25%. Further, the options that do not work properly when used as distractors are 5 items, but there are 40 items that must be revised because the distractive option is even considered as the key answer by smart participants.

Keywords: *validity, reliability, difficulty level, differentiation, distraction, TAP*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v22i1.15609

Pendahuluan

Aplikasi *Test Analisys Program* (TAP) merupakan salah satu program yang dapat digunakan dalam bidang pengukuran untuk menganalisis kualitas butir sebuah instrumen. Hasil analisisnya dapat dijadikan sebagai sumber informasi akurat dan sebagai dasar pengambilan keputusan, apakah instrumen tersebut baik atau tidak.

Salah satu penyebab rusaknya mutu pendidikan adalah hasil tes masuk yang tidak akurat. Untuk itu, penilaian yang benar akan memberikan informasi yang tepat serta mondorong dalam meningkatkan motivasi dan prestasi dalam pembelajaran mahasiswa. Hal ini dijelaskan oleh Iskandar (2013, p. 37) yang menyatakan bahwa sistem tes dan penilaian yang baik akan mondorong mahasiswa dalam meningkatkan motivasi dan prestasi dalam pembelajaran.

Namun yang sering terjadi dalam dunia pendidikan, kita sering dihadapkan pada masalah pengambilan keputusan, apakah seorang mahasiswa harus mengulang materi tertentu, pantas lulus ataukah harus tidak lulus. Hal tersebut bukanlah pekerjaan yang mudah. Dibutuhkan pertimbangan yang matang agar dapat menghasilkan suatu keputusan yang benar dan tepat sehingga tidak merugikan mahasiswa. Untuk itu, keputusan yang tepat dan benar sangat dipengaruhi oleh kualitas instrumen yang digunakan. Jika kualitas instrumen jelek maka pengambilan keputusan juga dipastikan akan jelek.

Instrumen merupakan suatu alat yang digunakan untuk mengukur suatu obyek penelitian, oleh karena itu instrumen tersebut harus memenuhi kriteria yang baik. Persyaratan instrumen yang baik setidaknya memenuhi syarat valid dan reliabel. Disamping memenuhi syarat valid dan reliabel juga harus memperhatikan karakteristik butir yaitu tingkat kesukaran, daya beda, dan keberfungsian pengecoh. Hal tersebut sesuai dengan pendapat Mansyur, Rasyid, & Suratno (2015, p. 30) yang mengatakan bahwa untuk memperoleh informasi yang akurat maka dibutuhkan instrumen yang sahih dan handal.

Menurut (Sudrajat, 2008) penilaian *(assessment)* adalah penerapan berbagai cara dan penggunaan beragam alat penilaian untuk memperoleh informasi tentang sejauh mana hasil belajar peserta didik atau ketercapaian kompetensi (rangkaian kemampuan) peserta didik. Sehingga hasil dari proses penilaian melahirkan keputusan-keputusan yang berkaitan dengan mahasiswa meliputi penempatan mahasiswa pada program pendidikan yang berbeda, pemberian nilai pada mahasiswa, membimbing dan mengarahkan mahasiswa, pemilihan mahasiswa untuk mengikuti program-program pendidikan, pemberian penghargaan dan sertifikat terhadap kompetensi mahasiswa.

Kata tes berasal dari bahasa latin *testum,* yang berarti alat untuk mengukur tanah. Sehingga Tes dapat didefinisikan sebagai sejumlah pertanyaan yang membutuhkan jawaban atau sejumlah pernyataan yang harus diberikan tanggapan guna mengukur tingkat kemampuan seseorang atau mengungkap aspek tertentu dari orang yang dikenai tes.

Tes didefinisikan sebagai suatu instrumen atau prosedur sistematik untuk mengobservasi dan menjelaskan satu atau beberapa karakteristik siswa dengan menggunakan suatu skala numerik atau skema klasifikasi (Nitko & Brookhart, 2007, p. 7). Selanjutnya Sax (1980, p. 13) berpendapat bahwa *" a test may be defined as a task or series of tasks used to obtain systematic observations presumed to be representative of educational or psychological traits or attributes".*

Analisis butir tes pada umumnya dimaksudkan untuk mengetahui besar kecilnya indeks tingkat kesulitan, indeks daya beda dan efektivitas pengecoh butir-butir soal yang bersangkutan. Analisis tes dapat dilakukan dengan menggunakan salah satu dari dua cara, tergantung teori tes mana yang digunakan. Teori tes tersebut dapat berupa teori tes klasik atau teori tes modern (Suryabrata, 2002, p. 24).

Analisis kualitas tes merupakan suatu tahap yang harus ditempuh untuk mengetahui derajat kualitas suatu tes, baik secara keseluruhan maupun butir soal yang menjadi bagian dari tes tersebut. Dalam penilaian hasil belajar, tes diharapkan dapat meng-

gambarkan sampel perilaku dan menghasilkan nilai yang objektif serta akurat. Jika tes yang digunakan dosen kurang baik, maka hasil yang diperoleh pun tentunya kurang baik pula. Hal ini dapat merugikan mahasiswa itu sendiri, artinya hasil yang diperoleh mahasiswa menjadi tidak objektif. Oleh sebab itu, tes yang digunakan harus memiliki kualitas yang baik. Tes hendaknya disusun berdasarkan prinsip dan prosedur penyusunan tes. Setelah digunakan perlu diketahui apakah tes tersebut berkualitas baik atau tidak maka perlu dilakukan analisis kualitas tes (Arifin, 2012, p. 22).

Hasil observasi pendahuluan pada lokasi penelitian ditemukan bahwa soal yang digunakan dalam penerimaan calon mahasiswa baru tidak melalui analisis secara empirik (uji validitas, reliabilitas, tingkat kesukaran, daya beda,dan pengecoh). Hal tersebut diungkapkan salah seorang petugas pelaksana tes. Sehingga bisa dipastikan bahwa informasi yang dikumpulkan dari tes yang diberikan, mengandung bias atau tidak sesuai dengan apa yang diinginkan.

Salah satu syarat instrumen tes yang baik harus melihat daya beda (diskriminasi) suatu butir tes dimana daya beda ini digunakan untuk membedakan antara peserta tes yang berkemampuan tinggi dan berkemampuan rendah. Daya beda butir dapat diketahui dengan melihat besar kecilnya indeks diskriminasi. Adapun fungsi dari daya pembeda tersebut adalah mendeteksi perbedaan individual yang sekecil-kecilnya di antara para peserta tes. Ramdani (2012, p. 28) juga mengungkapkan daya pembeda sebuah soal bertujuan untuk menunjukkan kemampuan soal tersebut atau membedakan antara mahasiswa yang pandai dengan yang kurang pandai.

Suatu soal yang dapat dijawab benar oleh mahasiswa pandai maupun oleh mahasiswa kurang pandai, maka soal itu tidak baik karena tidak mempunyai daya beda. Demikian pula jika semua mahasiswa, baik pandai maupun kurang pandai tidak dapat menjawab dengan benar maka soal tersebut juga tidak memiliki daya pembeda. Soal yang baik dan mempunyai daya pembeda adalah soal yang dapat dijawab benar oleh siswa-siswa yang pandai saja (Arikunto, 2009, p. 26). Butir soal yang tidak memiliki daya pembeda diduga terlalu mudah atau terlalu sulit maka perlu diperbaiki atau diganti dengan pertanyaan lain.

Penentuan daya beda butir biasanya dilakukan dengan menggunakan indeks korelasi, diskriminasi, dan indeks keselarasan item. Dari ketiga cara tersebut yang paling sering digunakan adalah indeks korelasi. Ada dua macam teknik korelasi yang biasa digunakan untuk menghitung nilai daya beda, yaitu: (1) teknik *point biserial*, (2) teknik *biserial* (Mansyur, Rasyid, & Suratno, 2009, p. 155). Untuk memudahkan perhitungan daya pembeda butir soal, Suprananta (2004) (Mansyur & Rasyid, 2007, p. 161) memberikan formula umum dengan rumus sebagai berikut :

$$D = \frac{\Sigma A}{n_A} - \frac{\Sigma B}{n_B}$$

Keterangan :

D : Indeks daya beda butir soal
$\sum X_A$: Banyaknya peserta tes yang menjawab benar kelompok atas
$\sum X_B$: Banyaknya peserta tes yang menjawab benar kelompok bawah
n_A : Banyaknya peserta tes pada kelompok atas
n_B : Banyaknya peserta tes pada kelompok bawah

Dari rumus di atas dapat dimaknai bahwa daya beda adalah perbedaan antara proporsi kelompok atas yang menjawab benar butir tes dengan proporsi kelompok bawah yang menjawab benar butir tes. Rumus tersebut dapat digunakan untuk menghitung daya beda butir soal dalam bentuk pilihan ganda.

Koefisien daya beda butir soal bergerak dari -1,00 sampai +1,00 maksudnya adalah jika suatu butir memiliki korelasi negatif, maka dapat dikatakan bahwa butir tersebut menyesatkan, karena subjek yang terdiri dari kelompok pandai menjawab salah soal yang ada daripada subjek pada kelompok kurang

pandai, sehingga harus didrop atau dibuang. Untuk menyatakan bahwa besaran daya beda dapat berfungsi dengan baik, ada beberapa patokan yang dapat digunakan.

Butir soal yang diterima harus memiliki indeks daya beda > 0,30 atau lebih. Butir dengan indeks daya beda kurang dari antara 0,10 sampai 0,30 perlu direvisi, dan jika daya bedanya < 0,10 maka butir tersebut harus dibuang. Sejalan dengan hal ini, Crocker & Algina (1986) (Mansyur & Rasyid, 2007, p. 155) memberikan patokan indeks daya beda seperti Tabel 1.

Tabel 1. Indeks Daya Beda Butir

Indeks Daya Beda	Kriteria Butir
$0,40 \leq D \leq 1,0$	Sangat Baik
$0,3 \leq D < 0,4$	Baik
$0,2 \leq D < 0,3$	Cukup dan perlu sedikit revisi
$D < 0,2$	Tidak Baik

Selain syarat daya beda juga harus memperhatikan tingkat kesukaran butir karena soal yang baik adalah soal yang tidak terlalu mudah atau tidak terlalu sukar. Soal yang terlalu mudah tidak merangsang mahasiswa untuk mempertinggi usaha memecahkannya. Sebaliknya, soal yang terlalu sukar akan menyebabkan mahasiswa berputus asa dan tidak mempunyai semangat untuk mencoba lagi karena berada di luar jangkauannya. Bilangan yang menunjukkan sukar dan mudahnya suatu soal disebut dengan indeks kesukaran. Adapun besarnya indeks kesukaran adalah antara 0,00 sampai dengan 1,00 (Mansyur et al., 2009, p. 20; Rofiah, Aminah, & Ekawati, 2013, p. 4). Indeks kesukaran ini menunjukkan taraf kesukaran soal. Soal dengan indeks kesukaran 0,00 menunjukkan bahwa soal itu terlalu sukar, sebaliknya indeks 1,00 menunjukkan bahwa soalnya terlalu mudah.

Proporsi menjawab benar *p* (*proportion correct*) adalah indeks kesukaran soal yang paling sederhana dan sering digunakan dalam menentukan besaran indeks. Rumus untuk menentukan besarnya indeks kesukaran secara matematis dirumuskan oleh Mansyur et al. (2009, p. 21) sebagai berikut:

$$Pi = \frac{\sum Xi}{Smi.N}$$

Pi : tingkat kesukaran butir soal
$\sum Xi$: Jumlah peserta tes yang menjawab benar
Smi : Skor maksimum
N : Jumlah peserta tes

Kriteria yang digunakan untuk menentukan jenis tingkat kesukaran butir soal disajikan dalam Tabel 2.

Tabel 2. Tingkat Kesukaran Butir Soal

Nilai *P*	Kategori
$p < 0,30$	Sukar
$0,30 \leq p \leq 0,70$	Sedang
$p > 0,70$	Mudah

Dari penjelasan di atas ada beberapa hal yang bisa disimpulkan berkaitan dengan indeks kesukaran butir yaitu bahwa nilai *p* bagi suatu butir hanya menunjukkan indeks bagi kelompok yang diuji. Harga *p* ini bisa berubah jika tes diujikan pada kelompok yang berbeda. Selain itu, indeks kesukaran yang dihasilkan dari rumus ini adalah indeks kesukaran yang berlaku bagi kelompok secara keseluruhan, bukan perorangan. Indeks kesukaran bagi tiap peserta tes tidak bisa disimpulkan dengan melihat indeks proporsi menjawab benar *p*.

Setiap tes pilihan ganda memiliki satu pertanyaan serta beberapa pilihan jawaban. Di antara pilihan jawaban yang ada, hanya satu yang benar. Selain jawaban yang benar tersebut, juga ada jawaban salah, yang dikenal dengan *distractor* (pengecoh). Dengan demikian, efektivitas pengecoh adalah seberapa baik pilihan yang salah tersebut dapat mengecoh peserta tes yang memang tidak mengetahui kunci jawaban yang tersedia. Semakin banyak peserta tes yang memilih pengecoh tersebut, maka distaktor itu dapat menjalankan fungsinya dengan baik. Kriteria pengecoh yang baik adalah apabila pengecoh tersebut dipilih oleh paling sedikit 5% dari peserta tes (Hamzah & Koni, 2012, p. 120).

Distractor (pengecoh) berfungsi untuk mengidentifikasi peserta tes yang berkemampuan tinggi. Pengecoh dikatakan berfungsi efektif apabila dipilih lebih banyak oleh peserta tes yang berasal dari kelompok bawah (berkemampuan rendah), sebaliknya apabila pengecoh itu dipilih lebih banyak oleh peserta tes yang mempunyai kemampuan tinggi, maka pengecoh itu tidak berfungsi sebagaimana mestinya. Bila pengecoh dipilih secara merata, maka termasuk pengecoh yang baik, apabila pengecoh lebih banyak dipilih oleh peserta tes dari kelompok atas dibandingkan dengan kelompok bawah maka termasuk pengecoh yang menyesatkan (Supranata, 2006, p. 32).

Dengan demikian maka pengecoh yang tidak memenuhi kriteria sebagai pengecoh yang baik, karena tidak satupun diantara peserta tes yang memilihnya sebaiknya diganti dengan pengecoh lain yang lebih menarik untuk dipilih oleh peserta tes. Agar semua opsi dalam setiap butir soal dapat berfungsi secara efektif, maka penyusunan pengecoh harus dilakukan sedemikian rupa sehingga tidak terlalu mencolok sebagai opsi yang salah. Pengecoh-pengecoh yang baik adalah yang serupa tetapi tidak sama dengan opsi benar sehingga mempunyai peluang untuk dipilih oleh peserta tes yang tidak berhati-hati.

Metode Penelitian

Penelitian ini termasuk dalam kategori jenis penelitian Ex Post Facto yaitu penelitian empiris yang sistematis dimana peneliti tidak mengendalikan variabel bebas secara langsung karena eksistensi dari variabel tersebut telah terjadi, atau karena pada dasarnya variabel tersebut tidak dapat dimanipulasi, Karlinger (Emzir, 2011, p. 18). Penelitian ini dilakukan pada perguruan tinggi STMIK AKBA yang ada di Kota Makassar Propinsi Sulawesi Selatan dengan objek penelitian semua jawaban hasil tes calon mahasiswa baru tahun 2014-2016.

Untuk memperoleh data di lapangan, peneliti menggunakan beberapa teknik yaitu: (1) observasi, yaitu pengumpulan data yang dilakukan dengan cara mengadakan pengamatan langsung di lapangan, yang ada hubungannya dengan masalah penelitian ini; (2) wawancara, yaitu pengumpulan data yang dilakukan dengan cara mengadakan tanya jawab secara langsung, yang ada hubungannya dengan masalah penelitian ini; (3) dokumentasi, yaitu pengumpulan daya melalui referensi-referensi tertulis berupa jawaban hasil tes mahasiswa, buku-buku, bahan ajar, dan lain-lain yang sangat relevan.

Teknik analisis data dilakukan secara kualitatif dan kuantitatif. Analisis kualitatif menggunakan format penelaahan oleh pakar dan analisis secara kuantitatif dengan menampilkan hasil analisis secara klasik yaitu dengan melihat, tingkat kesukaran, daya beda dan efektivitas pengecoh setiap soal atau item melalui program aplikasi TAP.

Adapun kriteria yang digunakan dalam membedakan daya beda merujuk pada pada tabel 1, sedangkan tingkat kesukaran butir soal merujuk pada pada Tabel 2. Pengecoh dinyatakan telah dapat menjalankan fungsinya dengan baik apabila dipilih oleh sekurang-kurangnya 5 % dari seluruh peserta tes (Sudijono, 2009, p. 14). Tingkat validitas tes dapat dilihat dengan mengikuti kriteria validitas isi dari (Gregory, 2007, p. 221) yaitu sebagai berikut.

0,8 – 1	*= Validitas sangat tinggi*
0,6 – 0,79	*= Validitas tinggi*
0,40 – 0,59	*= Validitas sedang*
0,20 – 0,39	*= Validitas rendah*
0,00 – 0,19	*= Validitas sangat rendah*

Selanjutnya kriteria yang digunakan dalam menentukan reliabilitas tes yaitu jika hasil analisis memiliki nilai reliabilitas lebih besar atau sama dengan 0,70 maka dikatakan reliabel (Linn dalam Mansyur et al., 2009, p. 24).

Hasil Penelitian dan Pembahasan

Analisis kualitatif dilakukan untuk mereview butir soal dari aspek materi, konstruksi, dan bahasa sehingga diketahui validitas instrumen tes berdasarkan pandangan para pakar. Aspek-aspek yang diperhatikan dalam memvalidasi instrumen ini adalah:

petunjuk, cakupan soal, bahasa. Tabel 3 adalah rangkuman hasil validasi instrumen tes untuk setiap aspek pengamatan. Dari hasil penilaian para ahli yang berjumlah 2 orang dosen sebagai pakar bidang tes, sehingga instrumen tersebut dapat digunakan dengan melakukan revisi terlebih dahulu.

Tabel 3. Hasil Penilaian Ahli Tes

Rater 1	Rater 2	Hasil Tabulasi Silang
1	3	C
2	2	A
3	3	D
4	2	B
3	3	D
3	2	B
2	3	C

Berdasarkan hasil tabulasi silang 2x2 tersebut maka selanjutnya dimasukkan ke dalam rumus Gregory $\mathrm{Vi} = \frac{\mathrm{D}}{\mathrm{A+B+C+D}}$ maka hasilnya sebesar 0,28. Sehingga instrumen ini memenuhi kriteria validitas isi pada kategori rendah, (Gregory, 2007, p. 221). Hasil penilaian dari 2 orang dosen sebagai pakar IT dapat dilihat seperti pada Tabel 4.

Tabel 4. Hasil Penilaian Ahli IT

Rater 1	Rater 2	Dari Tabulasi Silang
2	1	A
3	2	B
3	4	D
2	3	C
3	3	D
3	2	B
3	3	D

Berdasarkan hasil analisis validitas isi dengan menggunakan rumus Gregory ditemukan nilai Vi sebesar 0,42. Sehingga instrumen ini memenuhi validitas isi kategori sedang. Setelah melewati tahap uji validitas dilanjutkan dengan uji reliabilitas yang bertujuan untuk melihat tingkat kesepakatan validator ahli melalui analisis *ICC (Intraclass Correlation Coefficients)* dan hasil analisis ditemukan sebesar K=0,514.

Berdasar pada hasil analisis tersebut tampak bahwa nilai koefisien reliabilitas instrumen ini lebih kecil dari batas bawah r-eliabilitas yang telah ditentukan yaitu sebesar 0,70 menurut Linn (Mansyur et al., 2009, p. 24), sehingga instrumen tersebut tidak memenuhi kriteria reliabel dan berada pada level reliabilitas cukup (*fair*) (0,40 ≤ K ≤ 0,60: cukup), (Fleiss dalam Widhiarso, 2012, p. 15).

Selanjutnya analisis kuantitatif dilakukan dengan menggunakan Program TAP, yang secara otomatis menganalisis butir instrumen tes seperti tingkat kesukaran, daya beda, efektivitas pengecoh, reliabilitas tes serta beberapa statistik data lainya (ukuran dari data hasil tes). Hasil analisis secara deskriptif untuk semua butir soal dapat dilihat pada Tabel 5.

Tabel 5. Hasil Analisis Dekriptif Skor Peserta Tes

Kriteria	Hasil analisis
Jumlah peserta tes	300
Kemungkinan skor total	80
Skor maksimal	35
Skor minimum	11
Median	22
Mean	22.33
Standar deviasi	4.546
Variance	20.66

Berdasarkan hasil analisis pada Tabel 5, tampak bahwa jumlah responden yang mengikuti tes ini sebanyak 300 orang. Jika seorang peserta tes menjawab semua soal dengan benar maka skor maksimal yang mungkin diperoleh sebesar 80. Akan tetapi dari hasil tes tersebut, skor maksimal yang diperoleh responden sebesar 35, skor minimal sebesar 11, median 22, mean 22.33, standar deviasi 4.54 dan variance sebesar 20.66. Selanjutnya hasil analisis butir soal secara deskriptif dapat dilihat pada Tabel 6.

Terkait dengan Tabel 6. tampak bahwa jumlah butir soal yang dianalisis sebanyak 80 butir dan tidak ada butir yang hilang. Kemudian dari hasil analisis tersebut ditemukan rerata tingkat kesulitan butir soal

sebesar 0,279 yang menandakan bahwa soal yang digunakan dalam penerimaan calon mahasiswa baru, termasuk dalam kategori sulit, selanjutnya rerata daya beda butir sebesar 0,130. Daya beda dihitung berdasarkan pembagian dua kelompok peserta tes yaitu kelompok atas dan kelompok bawah.

Tabel 6. Hasil Analisis Deskriptif Butir Soal

Kriteria	Hasil Analisis
Jumlah butir soal	80
Rerata tingkat kesulitan butir	0,279
Rerata daya beda butir	0,130
Koefisisen reliabilitas (KR20)	0,281
Kesalahan pengukuran	3,83

Kelompok atas dikategorikan sebagai *testee* yang tergolong sebagai anak yang pandai sedangkan peserta tes yang berada pada kelompok bawah dikategorikan sebagai *testee* yang tergolong kurang pandai, sedangkan angka 0,130 tersebut menunjukkan bahwa rata-rata butir soal tidak mampu membedakan antara calon mahasiswa pandai dengan yang kurang pandai. Karena besar kecilnya daya beda dapat diketahui melalui hasil analisis diskriminasi butir dengan membandingkan hasil analisis dengan kriteria yang telah ada.

Lebih lanjut, dengan menggunakan KR 20 diperoleh nilai kesalahan baku pengukuran sebesar 3,83 dan tingkat reliabilitas instrumen tes sebesar 0,281 hal ini menandakan bahwa tingkat keajekan instrumen tes yang digunakan dalam kategori buruk karena indeks reliabilitas lebih kecil dari 0,4 (K < 0,4) atau masuk kategori *Bad,* Fleiss (Fleiss dalam Widhiarso, 2012, p. 15). Untuk melengkapi penyataan sebelumnya, maka hasil analisis tingkat kesukaran setiap butir soal dapat dilihat pada Tabel 7.

Tingkat kesukaran berdasar pada besarnya indeks korelasi yang berkisar antara 0 sampai 1. Makin tinggi indeks korelasi maka butir soal tersebut semakin mudah dan semakin kecil indeks korelasi maka butir soal tersebut semakin sulit. Mansyur et al. (2009, p. 20) membedakan tingkat kesukaran soal ke dalam tiga kategori yaitu soal yang memiliki $p \leq 0,3$ biasanya disebut sebagai soal sukar, soal yang memiliki $p \geq 0,7$ biasanya disebut soal mudah, adapun soal yang memiliki p antara 0,3 sampai 0,7 disebut sebagai soal yang sedang.

Merujuk pada Tabel 7, jumlah butir soal yang masuk kategori sukar sebanyak 46 butir soal (57,5%), butir soal yang berada pada kategori sedang sebanyak 34 butir atau sebesar (42,5%) sedangkan butir soal yang berada pada kategori mudah tidak ada. Dari hasil analisis tersebut kelihatan jumlah soal yang sukar lebih banyak dibandingkan dengan butir soal yang sedang, tetapi perlu diketahui bahwa hasil analisis tersebut bukan satu-satunya indikator bahwa soal yang sukar atau mudah adalah jelek, karena dalam analisis butir soal terdapat beberapa kategori yang harus diperhatikan seperti tingkat kesukaran dan daya beda untuk menilai apakah butir tersebut baik atau tidak. Hasil analisis daya beda dapat dilihat pada Tabel 8.

Tabel 7. Hasil Analisis Tingkat Kesukaran

Kategori	Butir soal	Jumlah
Sukar *P < 0,30*	3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 16, 19, 20, 22, 23, 24, 26, 28, 29, 31, 34, 27, 38, 29, 43, 45, 47, 48, 49, 50, 52, 53, 54, 56, 57, 60, 61, 64, 67, 69, 70, 71, 72, 75, 77, 78.	46 (57,5%)
Sedang $0,30 \leq p \leq 0,70$	1, 2, 5, 8, 15, 17, 18, 21, 25, 27, 30, 32, 33, 35, 36, 40, 41, 42, 44, 46, 51, 55, 58, 59, 62, 63, 65, 66, 68, 73, 74, 76, 79, 80.	34 (42,5%)
Mudah *P > 0,70*	-	-
Total soal	80	80 (100%)

Tabel 8. Hasil Analisis Daya Beda

Kategori	Butir Soal	Jumlah
Sangat baik $0{,}40 \leq D \leq 1{,}0$	2, 32, 33, 50	4 (5%)
Baik $0{,}3 \leq D < 0{,}4$	8, 9, 24, 25, 30, 35, 36, 45, 48, 52, 54, 63, 68, 74, 79, 80	16 (20%)
Perlu sedikit revisi $0{,}2 \leq D < 0{,}3$	1, 15, 17, 42, 44, 51, 55, 56, 58, 62, 66.	11 (13,75%)
Tidak baik $D < 0{,}2$	3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 31, 34, 37, 38, 39, 40, 41, 43, 46, 47, 49, 53, 57, 59, 60, 61, 64, 65, 67, 69, 70, 71, 72, 73, 75, 76, 77, 78.	49 (61, 25%)
Total soal	80	80 (100%)

Berdasarkan hasil analisis pada Tabel 8, tampak bahwa jumlah butir soal yang memiliki daya beda yang sangat baik sebanyak 4 butir soal (5%) yang berarti bahwa butir soal tersebut mampu membedakan antara peserta tes yang pandai dan peserta tes yang kurang pandai. Selanjutnya, terdapat 20% yang memiliki daya beda pada kategori baik yang berarti butir-butir soal tersebut dapat membedakan antara peserta tes yang pandai dengan peserta tes yang kurang pandai.

Selain butir soal yang memiliki daya beda kategori sangat baik dan kategori baik juga terdapat butir soal yang memiliki daya beda pada kategori cukup, sebanyak 11 butir (13,75%) yang berarti harus melewati tahap revisi. Jika butir-butir soal ini telah direvisi maka butir soal tersebut dapat digunakan. Sedangkan jumlah butir soal yang berada pada kategori tidak baik sebanyak 49 butir yaitu sebesar (61,25%) sehingga harus dibuang.

Selanjutnya, untuk melihat pengecoh butir soal yang berfungsi pada Tabel 9.

Rendahnya daya beda biasanya disebabkan oleh tingkat keberfungsian pengecoh butir soal, selain itu pengecoh juga memberikan dampak terhadap tingkat kesukaran butir soal karena jika terdapat satu atau dua pengecoh pada suatu butir soal yang tidak berfungsi maka indeks tingkat kesukaran butir soal akan menurun, karena peluang peserta tes untuk menjawab dengan benar semakin meningkat.

Berdasarkan hasil analisis yang tertera pada Tabel 9, tampak bahwa jumlah butir soal yang memiliki pengecoh yang baik untuk option A yaitu sebanyak 68 butir soal dan terdapat 2 butir soal yang memiliki option A tidak berfungsi sebagai pengecoh yaitu butir soal 27, 28 dan 10 butir soal menempatkan option A sebagai kunci. Selanjutnya pada option B terdapat 55 butir soal yang berfungsi sebagai pengecoh dan 25 butir soal yang menempatkan kunci berada pada option B.

Lebih lanjut, terdapat 62 butir soal yang menggunakan option C berfungsi sebagai pengecoh dan 3 butir soal yang menggunakan option C tidak berfungsi dengan baik yaitu pada butir 1, 6, 31 dan 15 butir soal yang menjadikan option C sebagai kunci. Lebih lanjut, terdapat 51 butir soal yang menggunakan option D sebagai pengecoh dan semuanya berfungsi dengan baik. Selain itu, 29 butir soal yang menjadikan option D sebagai kunci. Untuk melengkapi penyajian keberfungsian pengecoh pada Tabel 9 dapat pula dilihat Tabel 10 untuk melihat letak kunci jawaban pada setiap butir soal.

Tabel 9. Keberfungsian Pengecoh

Menjadi kunci	Pengecoh tidak berfungsi dengan baik	Butir soal	Jumlah	Pengecoh yang berfungsi dengan baik	Butir soal	Jumlah
10	A	27, 28.	2	A	1, 2, 3, 5, 6, 8, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 29, 30, 31, 32, 33, 35, 36, 38, 39, 41, 42, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80.	68
25	B	-	-	B	2, 3, 4, 6, 7, 9, 10, 12, 13, 14, 16, 18, 19, 20, 22, 23, 25, 26, 28, 30, 31, 32, 34, 36, 37, 38, 40, 41, 43, 45, 46, 48, 49, 50, 52, 53, 54, 56, 57, 58, 60, 61, 63, 64, 66, 67, 68, 69, 71, 72, 74, 75, 77, 78, 79.	55
15	C	1, 6, 31.	3	C	2, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 20, 21, 23,24, 25,26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 54, 55, 56, 58, 59, 60, 62, 63, 65, 66, 68,70, 71, 73, 74, 76, 77, 79, 80.	62
29	D	-	-	D	1, 3, 4, 5, 7, 8, 10, 11, 13, 14, 15, 16, 17, 19, 21, 22, 24, 26, 27, 29, 31, 33, 34, 35, 37, 39, 40, 42, 43, 44, 46, 47, 49, 51, 53, 55, 57, 59, 61, 62, 64, 65, 67, 69, 70, 72, 73, 75, 76, 78, 80.	51

Tabel 10. Tampilan Kunci Jawaban dalam Aplikasi TAP

```
============================
CORRECT ANSWERS (Item#-Key):
============================
 # 1-2  # 2-4  # 3-3  # 4-1  # 5-2  # 6-4  # 7-1  # 8-2  # 9-4 #10-1
 #11-2 #12-4 #13-3 #14-1 #15-2 #16-3 #17-2 #18-4 #19-3 #20-4
 #21-2 #22-3 #23-4 #24-2 #25-4 #26-1 #27-2 #28-4 #29-2 #30-4
 #31-2 #32-4 #33-2 #34-1 #35-2 #36-4 #37-1 #38-4 #39-2 #40-1
 #41-4 #42-2 #43-1 #44-2 #45-4 #46-1 #47-2 #48-4 #49-3 #50-4
 #51-2 #52-4 #53-3 #54-4 #55-2 #56-4 #57-3 #58-4 #59-2 #60-4
 #61-3 #62-2 #63-4 #64-3 #65-2 #66-4 #67-3 #68-4 #69-3 #70-2
 #71-4 #72-3 #73-2 #74-4 #75-3 #76-2 #77-4 #78-3 #79-4 #80-2
```

Berdasarkan Tabel 10 dalam penyajian kunci jawaban setiap butir soal yang ada pada *CORRECT ANSWERS (Item#-Key)* , seperti pada butir soal 1 letak kunci jawab berada pada option B (2), sedangkan pada soal butir nomor 2 letak kunci jawaban berada pada option D (4) dan seterusnya. Akan tetapi pada setiap peletakan kunci jawaban pada setiap soal, terkadang ada hal-hal yang menyebabkan letak kunci jawaban tersebut tidak sesuai dengan teori pengembangan tes seperti pada contoh butir soal nomor 3 yang disajikan pada Tabel 11.

Dari hasil analisis pada soal nomor 3 di atas tampak bahwa letak kunci jawaban yang telah ditentukan berada pada option 3

(C), akan tetapi peletakan kunci jawaban yang terletak pada option C tersebut tidak sesuai dan dianggap bahwa option 2 (B) lebih pantas menjadi kunci daripada option 3 (C) karena sebagian besar peserta tes yang pintar menganggap option 2 (B) sebagai kunci jawabannya, berbeda dengan option 3 (C) yang banyak dipilih oleh peserta tes kurang pintar sehingga harus direvisi/dibuang. Selain butir soal nomor 3 juga terdapat butir soal yang lain memiliki hal yang sama, seperti yang tampak pada Tabel 11.

Tampak pada Tabel 11. Jumlah butir soal yang memiliki kunci jawaban yang tidak sesuai yaitu sebanyak 40 butir soal karena pengecoh yang sedianya dijadikan sebagai alat untuk mengelabui peserta tes malah dianggap bisa menjadi kunci jawaban karena pengecoh lebih banyak dipilih oleh peserta tes yang pintar daripada peserta tes yang kurang pintar, sedangkan untuk kunci jawaban dalam setiap butir soal malah banyak dipilih oleh peserta yang kurang pintar.

Tabel 11. Letak Kunci Jawaban

Item	Group	Option 1	Option 2	Option 3	Option 4
3	TOTAL	95 (0.317)	112 (0.373)	46*(0.153)	47 (0.157)
	High	28 (0.337)	41 (0.494)	7 (0.084)	7 (0.084)
	Low	36 (0.343)	28 (0.267)	17 (0.162)	24 (0.229)
	Diff	-8(-0.006)	13#(0.227)	-10(-0.078)	-17(-0.144)

Tabel 12. Pemilihan Option sebagai Letak Kunci Jawaban

Butir soal	Letak kunci (*)	Saran pengganti (#)	Butir soal	Letak kunci (*)	Saran pengganti (#)
4	Option 1	Option 4	46	Option 1	Option 2, 4
6	Option 4	Option 2	47	Option 2	Option 1
7	Option 1	Option 3, 4	49	Option 3	Option 2
10	Option 1	Option 2	53	Option 3	Option 1, 2
11	Option 2	Option 3, 4	57	Option 3	Option 1, 2
13	Option 3	Option 1,2	59	Option 2	Option 4
14	Option 1	Option 4	60	Option 4	Option 2, 3
16	Option 3	Option 4	61	Option 3	Option 4
19	Option 3	Option 2	64	Option 3	Option 1, 2
21	Option 2	Option 1	65	Option 2	Option 4
22	Option 3	Option 1	67	Option 3	Option 1, 2
26	Option 1	Option 2	69	Option 3	Option 2
27	Option 2	Option 1, 4	70	Option 2	Option 4
28	Option 4	Option 2	71	Option 4	Option 2
29	Option 2	Option 4	72	Option 3	Option 4
31	Option 2	Option 1	73	Option 2	Option 1, 3
34	Option 1	Option 4	75	Option 3	Option 2
37	Option 1	Option 4	76	Option 2	Option 3, 4
39	Option 2	Option 1	78	Option 3	Option 2, 4
43	Option 1	Option 3, 4	3	Option 3	Option 2

Simpulan

Berdasarkan hasil analisis dan pembahasan sebelumnya maka dapat disimpulkan bahwa (1) hasil validitas isi ditemukan nilai Vi sebesar 0,42 kategori sedang. (2) indeks relibilitas sebagai kesepakatan validator ahli melalui analisisi ICC diperoleh nilai K sebesar 0,514 < 0,70, yang berarti berada pada kategori cukup (fair). (3) jumlah butir soal yang sukar sebanyak 57,5%, kategori sedang sebanyak 42,5% dan soal mudah tidak ada.

(4) butir soal yang memiliki daya beda kategori sangat baik sebanyak 5%, kategori baik sebanyak 20%, perlu sedikit revisi sebanyak 13,75%, dan kategori tidak baik sebanyak 61,25% dan (5) pengecoh yang tidak berfungsi dengan baik terdapat pada 5 butir soal, serta terdapat 40 butir soal yang harus direvisi/dibuang karena memiliki pengecoh yang justru dianggap sebagai kunci jawaban oleh peserta yang pintar.

Berdasarkan hasil penelitian dan simpulan yang disampaikan, penlitian memberikan saran sebagai berikut: (1) instrumen tes yang baik sebelum digunakan seharusnya didiskusikan dalam *Focus Group Discussion* (FGD) serta sosialisasi dengan pelaku bidang pendidikan dan ahli dalam bidang ilmu pengukuran dan evaluasi; (2) pemahaman dan kemampuan dalam penyusunan instrumen tes sangat diperlukan untuk menghasilkan butir instrumen yang baik; (3) pelaksana tes harus bersifat adil pada setiap peserta tes.

Daftar Pustaka

Arifin, Z. (2012). *Evaluasi pembelajaran prinsip, teknik prosedur*. Bandung: PT: Remaja Rosdakarya.

Arikunto, S. (2009). *Dasar-dasar evaluasi pendidikan*. Jakarta: PT. Bumi Aksara.

Emzir. (2011). *Metodologi penelitian pendidikan kuantitatif dan kualitatif*. Jakarta: Rajawali Pers.

Gregory, R. . (2007). *Psychological testing: history, principles, and applications* (5th ed.). New York: Pearson Education Group, Inc.

Hamzah, B. U., & Koni, S. (2012). *Assessment pembelajaran*. Jakarta: Bumi Aksara.

Iskandar, A. (2013). Pengembangan perangkat penilaian psikomotor di sekolah menengah kejuruan (SMK). *Inspiration Jurnal Teknologi Informasi Dan Komunikasi*, *3*(1). Retrieved from http://jurnal.akba.ac.id/index.php/inspiration/article/view/30

Mansyur, & Rasyid, H. (2007). *Penilaian hasil belajar*. Bandung: Wacana.

Mansyur, Rasyid, H., & Suratno. (2009). *Assesmen pembelajaran di sekolah*. Yogyakarta: Multi Pressindo.

Mansyur, Rasyid, H., & Suratno. (2015). *Asesmen pembelajaran disekolah. Panduan bagi guru dan calon guru*. Yogyakarta: Pustaka Pelajar.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students*. New Jersey: Pearson Education.

Ramdani, Y. (2012). Pengembangan instrumen dan bahan ajar untuk meningkatkan kemampuan komunikasi, penalaran, dan koneksi matematis dalam konsep integral. *Jurnal Penelitian Pendidikan*, *13*(1). Retrieved from http://jurnal.upi.edu/file/6-yani_ramdhani.pdf

Rofiah, E., Aminah, N. S., & Ekawati, E. Y. (2013). Penyusunan instrumen tes kemampuan berpikir tingkat tinggi fisika pada siswa SMP. *Jurnal Pendidikan Fisika*, *1*(2).

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2nd ed.). Belmont: Wadsworth Publishing Company.

Sudijono, A. (2009). *Pengantar statistik pendidikan*. Jakarta: Rajawali Pers.

Sudrajat, A. (2008). Pengembangan perangkat penilaian psikomotor. Retrieved January 20, 2012, from http://akhmadsudrajat.files.wordpress.

com/2008/08/penilaian-psikomotor.pdf

Supranata, S. (2006). *Analisis, validitas, reliabilitas dan interpretasi hasil tes*. Bandung: PT. Remaja Rosdakarya.

Suryabrata, S. (2002). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi.

Widhiarso, W. (2012). Mengestimasi reliabilitas. Retrieved February 12, 2012, from http://widhiarso.staff.ugm.ac.id/files/bab_2_estimasi_reliabilitas_via_spss.pdf