Jurnal Evaluasi Pendidikan
Volume 15, Nomor 1, 2024

Analisis Tes dan Butir Soal pada Moodle

Suwanda Priyadi
Fakultas Psikologi, Universitas Muhammadiyah surakarta

Alamat Korespondensi
alamat email: sp822@ums.ac.id

ABSTRACT
Moodle, a popular learning platform in Indonesia, offers quizzes to assess learning outcomes. Unfortunately, there is limited literature that covers the items and test analyses of Moodle. As a result, the purpose of this article is to provide a review of Moodle item and test analysis using classical test theory. This paper shows that the estimation formulas for Moodle items and tests are the same as the basic parameters of classical test theory. For example, 1) the coefficient of internal consistency (CIC) is the same as the coefficient alpha (α) and standard error is the same as the Standard Error of Measurement (SEM); 2) the facility index (FI) is the same as the item difficulty; and 3) the discrimination index (DI) is the same as the point-biserial correlation or item discrimination. Thus, Moodle's item and test analysis can be interpreted in terms of classical test theory.

Keywords
item analysis, moodle, test reliability, classical test theory

1. Pendahuluan

Salah satu platform pembelajaran yang diminati di Indonesia adalah Moodle (Modular Object-Oriented Dynamic Learning Environment). Moodle merupakan sistem manajemen pembelajaran sumber terbuka (*open source*) yang memungkinkan pengguna untuk menciptakan pengalaman belajar daring yang powerful, fleksibel, dan menarik (Rice, 2015). Berdasarkan statistik pengguna Moodle (2024), tercatat 5.764 situs pembelajaran berbasis Moodle berasal dari Indonesia. Hal ini membuat Indonesia menduduki peringkat 6 negara pengguna Moodle terbanyak di dunia.

Hal menarik yang dimiliki Moodle adalah fitur kuis. Fitur ini dirancang untuk memfasilitas penilaian hasil belajar siswa. Menurut Tim Pusdiklat Pegawai Kemendikbud (2016), terdapat lima fungsi penilaian hasil belajar, diantaranya: 1) memberikan umpan balik bagi guru dan siswa terkait capaian standar kompetensi dasar (formatif); 2) mengetahui sejauh mana tujuan pembelajaran telah dicapai diakhir pembelajaran (sumatif); 3) mengungkap kesulitan siswa dalam proses belajar (diagnostik); 4) menyesuaikan input (siswa) dengan fasilitas ruangan, tempat duduk serta fasilitas lainnya (selektif); 5) memberikan dorongan kepada siswa untuk meningkatkan prestasinya (motivasi).

Fitur kuis Moodle dilengkapi dengan laporan statistik kuis (*quiz statistics report*) (Collman, 2023). Laporan ini dapat diakses melalui menu Quiz kemudian dari menu navigasi Quiz pilih Results, dan pilih Statistics. Atau bisa melalui menu Administration, kemudian pilih Quiz administration, kemudian pilih Results, dan pilih Statistics. Laporan statistik kuis terdiri dari Overall quiz statistics dan Quiz question statistics. Overall quiz statistics terdiri dari: 1) Quiz information yang berisi informasi tentang tes dan hasil analisis tes; 2) Quiz structure analysis yang berisi hasil analisis butir soal; dan 3) Quiz statistics chart yang berisi grafik batang hasil analisis butir soal. Sementara itu, Quiz question statistics berisi informasi yang sama seperti Overall quiz statistics yang disajikan berdasarkan masing-masing butir soal.

Meskipun telah banyak artikel yang mengkaji pemanfaatan Moodle (Aulia & Waspada, 2019; Febliza dkk., 2021; Ikawati, 2015; Pramita dkk., 2021; Ramdani dkk., 2019; Setiawan & Rahayu, 2022; Waspada dkk., 2019), masih sedikit sekali artikel yang membahas hasil analisis tes dan butir soal pada Moodle. Salah satu artikel tersebut adalah penelitian Ramdani dkk. (2019) membandingkan hasil analisis tes dan butir soal pada Moodle dengan parameter yang ada di teori tes klasik dan model Rasch. Hasil penelitian ini menunjukkan bahwa hasil analisis tes dan butir soal pada Moodle memiliki kesamaan dengan parameter pokok dalam teori tes klasik. Seperti indeks kesukaran butir, indeks daya diskriminasi butir, dan koefisien

reliabilitas beserta estimasi kesalahan pengukuran (Hayat, 2021). Meskipun demikian, penelitian tersebut tidak memberikan penjelasan lebih lanjut mengenai formula estimasi dan kriteria interpretasi laporan statistik kuis yang ada pada Moodle.

Penjelasan lebih lanjut mengenai formula estimasi dan kriteria interpretasi hasil analisis tes dan butir soal pada Moodle akan memberikan manfaat teoretis dan praktis. Manfaat teoritis yang diberikan adalah tersedianya literatur yang secara khusus mengulas formula estimasi dan interpretasi hasil analisis tes dan butir soal pada Moodle. Hal ini akan menjadi referensi dalam meneliti lebih lanjut tentang analisis tes dan butir soal pada Moodle. Sementara itu, manfaat praktis yang diberikan adalah dasar interpretasi laporan statistik kuis bagi para guru yang memanfaatkan Moodle sebagai sistem manajemen pembelajaran. Analisis tes dan butir soal merupakan bagian dari tahapan mengembangkan tes yang baik dan benar. Pengembangan tes yang baik dan benar akan membuat fungsi penilaian hasil belajar siswa (formatif, sumatif, diagnostik, selektif, dan motivasi) menjadi sangat penting dan sangat terasa manfaatnya (Tim Pusdiklat Pegawai Kemendikbud, 2016).

Oleh karena itu, artikel ini bertujuan untuk mengulas analisis tes dan butir soal pada Moodle. Ulasan tersebut meliputi formula estimasi analisis tes dan butir soal yang digunakan Moodle beserta interpretasinya. Berdasarkan penelitian terdahulu, hasil analisis tes dan butir soal pada Moodle relatif mirip dengan parameter utama yang ada dalam teori tes klasik (Ramdani dkk., 2019). Sehingga, penulis membatasi pembahasan dalam artikel ini pada topik koefisien reliabilitas tes beserta estimasi kesalahan pengukuran, indeks kesukaran butir soal, dan indeks daya diskriminasi butir soal.

Koefisien reliabilitas adalah properti psikometri yang sangat penting untuk dilaporkan dari sebuah pengukuran psikologis (kognitif, afektif, dan konatif). Reliabilitas terkadang disebut keterandalan, konsistensi, derajat kepercayaan, kestabilan, keajekan, dan sebagainya, yang pada dasarnya memiliki gagasan pokok sejauh mana hasil suatu proses pengukuran dapat dipercaya (Azwar, 2012). Sayangnya, konsep ini masih sering dipertukarkan dalam penggunaannya ketika menyebut reliabilitas alat ukur dan reliabilitas hasil ukur. Padahal kedua istilah ini berbeda secara konsep dan pemaknaannya.

Perbedaan makna dalam penggunaan istilah reliabilitas alat ukur dan reliabilitas hasil ukur perlu diperhatikan (Azwar, 2012). Konsep reliabilitas alat ukur berkaitan erat dengan masalah eror pengukuran yang merujuk pada sejauh mana tes yang dikenakan berulang pada sampel yang sama menghasilkan ukuran yang berbeda. Sementara itu, konsep reliabilitas hasil ukur berkaitan erat dengan eror pengambilan sampel subjek ukur yang merujuk pada sejauh mana tes yang dikenakan pada sampel yang berbeda, namun berasal dari populasi yang sama menghasilkan ukuran yang berbeda. Estimasi reliabilitas yang dilakukan dalam praktik pengukuran psikologis lebih sesuai disebut sebagai reliabilitas hasil ukur dibandingkan dengan reliabilitas alat ukur karena estimasi yang dilakukan bersumber dari skor tes bukan dari fungsi tes.

Tujuan utama dilaporkannya estimasi reliabilitas hasil pengukuran adalah untuk memberikan gambaran mengenai konsistensi skor hasil pengukuran. Skor adalah respons subjek terhadap butir soal atau pertanyaan dalam instrumen pengukuran psikologis yang dinyatakan dalam bentuk angka (Azwar, 2015). Menurut pandangan teori tes klasik, skor kuantitatif yang langsung diperoleh sebagai hasil pengukuran disebut skor tampak (diberi simbol huruf X). Skor tampak tersebut sebenarnya terdiri atas skor sesungguhnya atau dapat disebut skor murni (diberi simbol T) dan juga eror pengukuran (diberi simbol E), dengan besaran eror untuk setiap skor yang dimiliki individu tidak diketahui.

Terdapat tiga cara yang dapat dilakukan untuk estimasi reliabilitas skor tes menurut prosedur dan sifat koefisien yang dihasilkannya (Azwar, 2012). Pertama adalah metode tes ulang (*test-retest*), metode ini dilakukan dengan cara menyajikan sebuah tes sebanyak dua kali dengan adanya tenggat waktu di antara keduanya. Kedua adalah metode bentuk-paralel

(*parallel-forms*), metode ini dilakukan dengan cara menyajikan tes bersama suatu tes alternatif yang juga memiliki tujuan ukur yang sama serta butir soal tes yang setara secara kualitas dan kuantitas. Ketiga adalah metode penyajian tunggal (*single-trial administration*), metode ini dilakukan dengan cara menyajikan tes hanya sekali saja pada satu kelompok subjek. Dari ketiga metode ini, metode penyajian tunggal menjadi metode yang paling populer dan banyak digunakan oleh pengembang tes.

Salah satu hal penting dalam melaporkan reliabilitas sebuah tes adalah pelaporan estimasi kesalahan pengukuran. Estimasi kesalahan pengukuran atau eror standar dalam pengukuran (SEM; *Standard Error of Measurement*) merupakan estimasi variasi skor yang diperoleh (X; skor tampak) terhadap skor sesungguhnya (T; skor murni) jika dilakukan tes berulang kali (Hayat, 2021). Semakin kecil kesalahan pengukuran yang terjadi, maka semakin cermat hasil pengukuran tersebut.

Selain reliabilitas, analisis butir soal adalah dasar evaluasi yang penting terhadap keputusan butir soal atau pertanyaan mana saja yang diikutkan dalam sebuah tes. Evaluasi terhadap butir soal dilakukan setidaknya terhadap tiga parameter yaitu: (a) daya diskriminasi butir; (b) tingkat kesukaran butir; dan (c) efektivitas distraktor (Azwar, 2016). Estimasi ketiga parameter ini hanya dapat dilakukan melalui *field-test* yang menghasilkan data respons subjek terhadap butir soal yang ada dalam suatu tes.

Indeks kesukaran butir adalah parameter yang mendeskripsikan seberapa sukarkah bagi sekelompok subjek yang dites untuk memberikan jawaban yang benar terhadap suatu butir soal (Azwar, 2016). Parameter kesukaran butir tidak berlaku secara umum melainkan hanya berlaku bagi kelompok subjek yang di tes. Dibandingkan daya diskriminasi butir, taraf kesukaran butir selalu disesuaikan dengan tujuan pembuatan tes. Terkadang soal-soal yang sulit lebih disukai ketika tujuan tes untuk seleksi dan terkadang soal-soal yang memiliki tingkat kesukaran yang sedang lebih disukai karena mampu menghasilkan variansi skor tes yang optimal. Parameter ini diestimasi dengan cara membagi jumlah subjek yang menjawab benar terhadap suatu soal dengan jumlah subjek yang menjawab soal tersebut.

Sementara itu, daya diskriminasi butir adalah sejauh mana butir soal mampu membedakan individu satu dengan individu lainnya berdasarkan atribut yang diukur oleh tes (Azwar, 2016). Dengan kata lain, daya diskriminasi butir adalah parameter yang menunjukkan keberfungsian butir soal dalam membedakan kelompok yang terkategori memiliki kemampuan (skor) tinggi dengan kemampuan (skor) rendah. Parameter ini dapat diestimasi melalui formula indeks daya diskriminasi, *point-biserial correlation, biserial correlation coefficient, phi coefficient, dan tetrachoric correlation coefficient* (Crocker & Algina, 2008).

Menurut Azwar (2016), butir soal yang baik tidak hanya memiliki daya diskriminasi yang tinggi, tingkat kesukaran yang sesuai akan tetapi juga harus memiliki distraktor-distraktor yang efektif. Distraktor adalah opsi yang bukan menjadi kunci jawaban pada soal yang berbentuk pilihan ganda. Distraktor yang efektif ditandai dengan banyaknya kelompok yang terkategori memiliki kemampuan (skor) rendah memilih opsi tersebut dibandingkan dengan kelompok yang memiliki kemampuan (skor) tinggi. Secara umum, analisis butir soal adalah properti yang sangat penting dalam menunjang performansi dan kualitas sebuah tes.

2. Metode Penelitian

Artikel ini adalah ulasan naratif (*narrative review*) tentang analisis tes dan butir soal pada Moodle. Ulasan ini meliputi formula estimasi analisis tes dan butir soal pada Moodle beserta interpretasinya. Secara khusus, ulasan ini secara kritis membandingkan formula estimasi analisis tes dan butir soal pada Moodle dengan parameter utama yang ada dalam teori tes klasik. Ulasan naratif (Ferrari, 2015) dianggap paling tepat (misalnya, dibandingkan dengan ulasan sistematis dan/atau metaanalisis) karena fokus utamanya konseptual.

3. Hasil Penelitian dan Pembahasan

Secara umum, hasil dan pembahasan dalam artikel ini dibagi ke dalam dua subbahasan. Pertama, bahasan Quiz information yang membahas hasil analisis tes. Kedua, bahasan Quiz structure analysis yang membahas hasil analisis butir soal.

3.1. Quiz Information (Analisis Tes)

Pada bagian ini berisi beberapa informasi dasar tentang tes yang telah disajikan kepada siswa. Berdasarkan pendekatan teori tes klasik, beberapa informasi penting pada bagian ini adalah *standard deviation, coefficient of internal consistency, error rasio* dan *standard error*. Rangkuman dari informasi tersebut dapat dilihat pada Tabel 1.

Tabel 1. Rangkuman Quiz Information

Informasi Tes	Formula Estimasi	Hasil Estimasi
Standard Deviation	$SD = \sqrt{V(t)}$	13,14% (0,1314)
Coefficient of Internal Consistency	$CIC = 100\frac{P}{P-1}\left(1 - \frac{1}{V(T)}\sum_{p \in P} V(x_p)\right)$	76,92% (0,7692)
Error Ratio	$ER = 100\sqrt{1 - \frac{CIC}{100}}$	48,05% (0,4805)
Standard Error	$SE = \frac{ER}{100}SD$	6,31% (0,0631)

Keterangan. Sumber formula adalah Hunt (2013) dan sumber hasil estimasi adalah dokumentasi pribadi.

Standard deviation (SD) adalah ukuran sebaran data dari nilai rerata skor. Semakin besar nilai SD maka dapat dikatakan bahwa skor individu memiliki sebaran yang luas, sebaliknya semakin kecil nilai standar deviasi maka nilai tersebut akan mendekati nilai rerata, artinya skor individu berada di sekitar nilai rerata sehingga dapat dikatakan bahwa kemampuan individu yang dites cenderung homogen. Informasi statistik lain yang memiliki fungsi memberikan informasi mengenai sebaran data adalah *skewness* dan *kurtosis* yang pada kesempatan ini tidak dibahas.

Tidak ada kriteria tunggal dalam menginterpretasikan nilai SD, namun nilai ini harus dilaporkan agar pembaca mengetahui sebaran data yang disajikan. Menurut Butcher (2022), nilai SD yang disukai berada dalam rentang 12% (0,12) hingga 18% (0,18). Meskipun demikian, SD bukanlah bagian dari properti psikometri tes, akan tetapi SD merupakan estimasi statistik yang penting karena selalu digunakan dalam estimasi properti psikometris.

Coefficient of internal consistency (CIC) adalah estimasi yang terkadang disebut dengan koefisien reliabilitas. Formula estimasi yang digunakan Moodle adalah koefisien alfa (α) yang diperkenalkan Cronbach (1951, hlm. 299).

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \qquad (1)$$

Pada formula (1) diketahui n adalah jumlah butir soal tes, V_t adalah variansi skor tes dan V_i adalah variansi skor butir soal ke-i. Formula ini memiliki kemiripan dengan formula yang digunakan Moodle (Collman, 2023):

$$CIC = 100\frac{P}{P-1}\left(1 - \frac{1}{V(T)}\sum_{p \in P} V(x_p)\right) \qquad (2)$$

Jika kita sederhanakan dan menghilangkan bentuk persentasenya maka formula (2) menjadi:

$$CIC = \frac{P}{P-1}\left(1 - \frac{1}{V(T)}\sum_{p \in P} V(x_p)\right) \qquad (3)$$

Pada formula (3) diketahui p adalah jumlah butir soal tes, $V(T)$ adalah variansi skor tes dan $V(x_p)$ adalah variansi skor butir soal ke-p. Formula ini identik dengan formula (1) sebagaimana yang kita bicarakan sebelumnya.

Formula alfa (α) sebenarnya diderivasi dari formula Kuder-Richardson (KR 20) berikut ini (L. J. Cronbach, 1951, hlm. 299):

$$r_{tt(KR20)} = \frac{n}{n-1}\left(1 - \frac{\sum_i p_i q_i}{\sigma_t^2}\right);\ (i = 1,2, \dots n) \qquad (4)$$

Pada formula (4) diketahui i adalah butir soal, p_i adalah proporsi mendapatkan skor 1, dan q_i proporsi mendapatkan skor 0 pada sebuah butir soal. Sementara itu, σ_t^2 adalah variansi skor tes.

Formula yang dikenalkan Cronbach (1951) bukanlah formula yang baru. Hanya saja sitasi yang terus diberikan kepada artikel Cronbach tahun 1951 membuat formula alfa (α) menjadi terasosiasi dengan nama Cronbach dan dikenal luas dengan Cronbach's alpha (α). Cronbach sendiri melalui artikelnya tahun 2004 merasa kurang nyaman dengan nama *Cronbach's alpha* dan berkata *"It is an embarrassment to me that the formula became conventionally known as Cronbach's α"* (L. Cronbach & Shavelson, 2004, hlm. 397).

Beberapa ahli merekomendasikan agar menggunakan formula (4) atau KR21 untuk estimasi reliabilitas tes yang datanya dikotomi (jawaban benar diberi skor 1 dan salah diberi skor 0). Studi yang dilakukan Wibowo (2016) menunjukkan bahwa penggunaan koefisien alfa (α) pada data dikotomi menghasilkan nilai reliabilitas yang *underestimate*. Meskipun demikian, perbedaan hasil estimasi kedua formula ini dirasa tidak terpaut jauh karena formula estimasinya relatif identik.

Terlepas dari perdebatan tepat atau tidaknya penggunaan koefisien alfa (α) untuk data dikotomi, terlihat pada tabel 1. nilai CIC hasil tes X sebesar 76,92% (0,7692). Nilai ini setidaknya sudah cukup memadai untuk penilaian *class-room test*. Menurut Wells dan Wollack tes yang memiliki risiko tinggi (*high stake*) yang biasanya digunakan untuk keperluan asesmen ataupun seleksi minimal memiliki koefisien sebesar 0,9 (90%) (Azwar, 2016). Adapun tes yang digunakan untuk keperluan selain itu minimal memiliki koefisien paling tidak sebesar 0,8 (80%) atau 0,85 (85%). Sedangkan untuk keperluan class-room test yang dibuat oleh guru minimal memiliki koefisien sebesar 0,7 (70%). Secara khusus, Moodle memberikan saran setidaknya nilai CIC di atas 75% (0,75) dan apabila nilai CIC di bawah 64% (0.64) maka tes tersebut harus direvisi (Butcher, 2022).

Dalam setiap pelaporan estimasi reliabilitas perlu disertai estimasi eror standar dalam pengukuran (*standard error of measurement*) (Azwar, 2016). Informasi mengenai eror standar dalam pengukuran pada Moodle terdapat pada informasi *standard error*. Estimasi yang dihasilkan standard error sama dengan estimasi yang dihasilkan formula standard error of measurement sebagai berikut (Azwar, 2016, hlm. 191):

$$S_E = s_X\sqrt{1 - r_{xx'}} \qquad (5)$$

Pada formula (5) diketahui s_X adalah *standard deviation* skor tes dan $r_{xx'}$ adalah koefisien reliabilitas tes. Sehingga, nilai S_E didapatkan sebagai berikut.

$$S_E = 0{,}1314\sqrt{1 - 0{,}7692}$$
$$S_E = 0{,}0631$$

Berdasarkan hasil estimasi di atas, maka didapatkan nilai S_E sebesar 0,0631 (6,31%), nilai ini sama dengan nilai standard error pada tabel 1. Artinya, apabila skor tampak (X) yang

didapatkan seseorang adalah 60, maka prakiraan skor murni (T) orang tersebut adalah berkisar ditambah 60. Dengan kata lain, semakin kecil nilai S_E maka akan semakin cermat hasil pengukuran yang dilakukan.

3.2. *Quiz Structure Analysis* (Analisis Butir Soal)

Pada bagian ini terdapat beberapa informasi yang berkaitan dengan analisis butir soal. Informasi yang umum digunakan adalah Facility Index (FI) dan Discrimination Index (DI) (Gamage dkk., 2019; Gómez-Soberón dkk., 2013; López-Tocón, 2021). FI menggunakan formula estimasi sebagai berikut (Hunt, 2013):

$$F_p = 100\frac{\bar{x}_p - x_p(min)}{x_p(max) - x_p(min)} \qquad (6)$$

Karena $x_p(min)$ selalu bernilai 0, maka formula ini dapat disederhanakan menjadi:

$$F_p = 100\frac{\bar{x}_p}{x_p(max)} \qquad (7)$$

Pada formula (7) diketahui $\bar{x}_p$ adalah nilai rerata peserta tes pada butir soal, sedangkan $x_p(max)$ adalah nilai maksimal yang didapatkan dari jawaban benar butir soal tersebut (=1). Dari formula (7) diketahui bahwa sebenarnya nilai rerata peserta tes pada butir soal sama dengan proporsi peserta tes menjawab benar (p). Dalam teori tes klasik hal ini dikenal dengan tingkat kesukaran butir soal. Misalnya terdapat 10 soal, 10 responden, dan 5 responden menjawab benar. Ketika jawaban benar diberi skor 1 dan jawaban salah diberi skor 0. Maka nilai $\bar{x}_p$ = [(5 x 1) + (5 x 0)] / 10 = 0,5. Sementara itu, nilai p = 5/10 = 0,5.

Menurut Azwar (2016), tingkat kesukaran yang mendekati 0,5 (50%) dianggap terbaik karena akan menghasilkan variansi butir soal terbesar. Tes yang berisi butir soal seperti ini akan menghasilkan variansi skor yang lebih besar dan perbedaan kemampuan (skor) individu akan lebih tajam tercermin dari perbedaan skor dalam kelompok bersangkutan. Hal ini menjadi penting ketika tes yang dikembangkan tersebut diinterpretasikan secara normatif.

Penentuan kriteria kesukaran butir soal atau FI biasanya didasarkan dari tujuan pengembangan tes. Apabila tes dikembangkan untuk keperluan seleksi dalam arti mencari individu terbaik maka butir soal yang sukar akan lebih disukai. Sementara itu, butir soal yang cenderung mudah akan disukai apabila tes yang dirancang akan digunakan untuk evaluasi formatif. Kesukaran butir soal atau FI juga memungkinkan guru memberikan umpan balik dalam proses belajar siswa berdasarkan topik yang tidak dipahami (sukar) (López-Tocón, 2021). Secara khusus, tim Moodle memberikan saran dalam interpretasi nilai FI sebagai berikut.

Informasi penting lainnya yang disajikan quiz structure analysis adalah Discrimination Indeks (DI). Adapun formula DI sebagai berikut (Hunt, 2013):

$$D_p = 100r(x_p, X_p) = 100\frac{C(x_p, X_p)}{\sqrt{V(x_p)V(X_p)}} \qquad (8)$$

Pada formula (8) diketahui x_p adalah skor butir soal, X_p adalah skor total butir soal tersisa, C adalah kovariansi, dan V adalah variansi. Formula ini identik dengan rumus korelasi *product moment* dari Karl Pearson sebagai berikut (Jacobs & Viechtbauer, 2017, hlm. 162).

Tabel 2. Kriteria interpretasi nilai *Facility Index* (FI)

Nilai FI (%)	Interpretasi
5 atau kurang	Sangat sulit atau ada sesuatu yang salah dengan butir soal
6-10	Sangat sulit
11-20	Sulit
21-34	Agak sulit
35-65	Rerata
66-80	Agak mudah
81-89	Mudah
90-94	Sangat mudah
95-100	Amat sangat mudah

Catatan. Sumber kriteria adalah Butcher (2022)

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (9)$$

Adapun bentuk ekuivalennya pada tataran sampel sebagai berikut.

$$r = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}} \quad (10)$$

Pada formula (10) diketahui $cov(x,y)$ adalah kovariansi antara skor x dan y, s_x^2 adalah variansi skor x, dan s_y^2 adalah variansi skor y. Formula korelasi product moment Spearman ini memiliki bentuk ekuivalensi khusus yang disebut korelasi *point-biserial* (*ρpbis*) yang digunakan apabila salah satu data yang dikorelasikan adalah data dikotomi. Formula tersebut adalah sebagai berikut (Crocker & Algina, 2008, hlm. 317).

$$\rho_{pbis} = \frac{\mu_+ - \mu_X}{\sigma_X}\sqrt{p/q} \quad (11)$$

Di mana μ_+ adalah rerata skor dari orang yang menjawab benar butir soal. μ_X adalah rerata skor dari seluruh peserta tes dan σ_X adalah standar deviasinya. p adalah tingkat kesukaran butir soal, dan q adalah (1 – p).

Kekurangan yang dimiliki formula (11) adalah tidak adanya koreksi terhadap efek spurious overlap. Efek ini terjadi karena skor tes X_p melibatkan skor butir soal x_p, sehingga korelasi antara skor butir soal x_p dan skor tes $X_p$adalah korelasi skor butir soal x_p dengan skor butir soal itu sendiri ditambah dengan skor butir soal lainnya. Hal ini mengakibatkan estimasi korelasi yang dihasilkan cenderung tinggi dari korelasi yang sesungguhnya (overestimasi). Semakin sedikit butir soal yang digunakan dalam sebuah tes maka akan semakin besar efek ini. Oleh karena itu, interpretasi daya diskriminasi yang dihasilkan formula (11) harus lebih hati-hati.

Menurut Azwar (2016) jika jumlah butir soal yang digunakan dalam tes lebih dari 30 maka umumnya efek ini dapat diabaikan. Namun, apabila jumlah butir soal yang digunakan dalam tes kurang dari 30 maka efek *spurious overlap* akan menjadi permasalahan serius dan perlu dilakukan koreksi. Berikut adalah formula yang dapat digunakan untuk koreksi efek *spurious overlap* (Crocker & Algina, 2008, hlm. 317).

$$\rho_{i(X-i)} = \frac{\rho_{Xi}\sigma_X - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_X^2 - 2\rho_{Xi}\sigma_X\sigma_i}} \quad (12)$$

Di mana $\rho i(\mathrm{x}-i)$ adalah korelasi antara skor butir soal dengan skor total tes yang telah

dikurangi dengan skor butir soal. σx adalah standar deviasi skor total tes dan σi adalah standar deviasi skor butir soal.

Formula (12) identik dengan formula (8). Formula (8) sudah melakukan upaya koreksi terhadap efek *spurious overlap.* Hal ini dapat dilihat dari formula estimasi X_p sebagai berikut (Hunt, 2013).

$$X_p(s) = T_s - x_p(s) \qquad (13)$$

Pada formula (13) $X_p(s)$ adalah skor total butir soal yang tersisa, T_s adalah skor total peserta tes, dan $x_p(s)$ adalah skor butir soal.

Moodle memberikan saran bahwa untuk mencapai diskriminasi butir yang maksimal setidaknya memerlukan nilai FI yang berkisar 30% - 70% (walaupun nilai tersebut tidak menjamin indeks diskriminasi yang tinggi) (Butcher, 2022). Secara umum, Moodle memberikan saran dalam interpretasi nilai DI sebagai berikut.

Tabel 3. Kriteria interpretasi nilai Discrimination Index (DI)

Nilai DI (%)	Interpretasi
Di atas 50	Daya diskriminasi sangat baik
30 – 50	Daya diskriminasi cukup
20 - 29	Daya diskriminasi lemah
0 - 19	Daya diskriminasi sangat lemah
Negatif (-)	Butir soal mungkin invalid

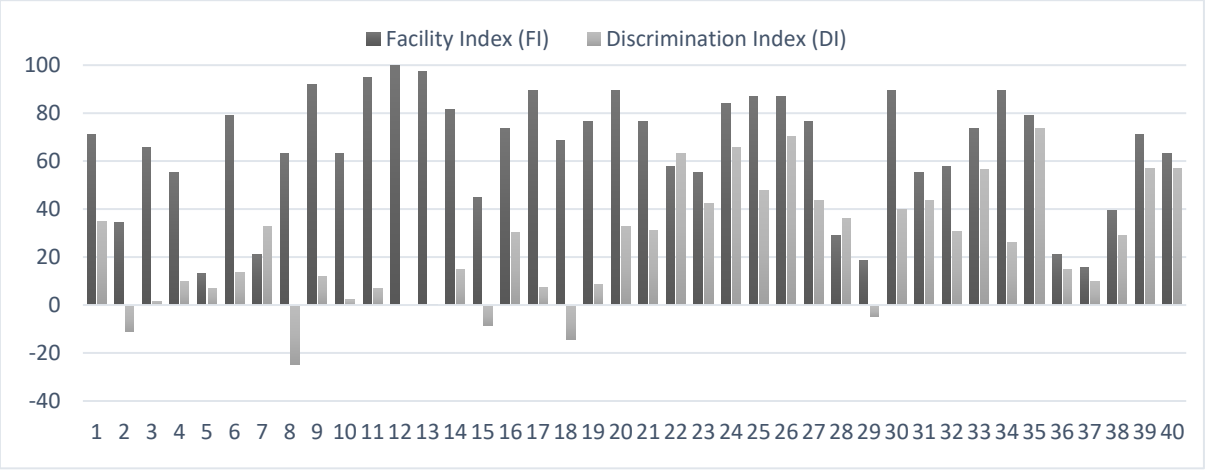
Catatan. Sumber kriteria adalah Butcher (2022).

Sebagai ilustrasi, pada gambar 1. terlihat bahwa terdapat beberapa butir soal yang memiliki daya diskriminasi atau nilai DI yang negatif yaitu butir soal 2, 8, 15, 18, dan 29. Butir soal yang bernilai negatif ini adalah butir soal yang buruk sehingga harus dikeluarkan dari tes atau direvisi. Salah satu penjelasan mengapa nilai DI pada butir soal tersebut menjadi negatif karena butir soal tersebut berfungsi terbalik, artinya lebih banyak individu yang tergolong memiliki kemampuan (skor) yang tinggi menjawab salah pada butir soal tersebut, dibandingkan individu yang memiliki kemampuan (skor) yang rendah.

Berdasarkan gambar 1. terlihat bahwa hanya butir soal 1, 7, 16, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 35, 39, dan 40 (berdasarkan kriteria tabel 3.) yang memiliki nilai DI yang lebih besar dari 30% (0,3). Artinya butir soal ini adalah butir yang dapat dipilih dan digunakan dalam sebuah tes. Indeks kesukaran seperti FI juga menjadi pertimbangan penyaji tes ataupun developer tes sesuai dengan kebutuhan penyusunan tes.

Salah satu kekurangan dari artikel ini adalah pembahasan mengenai efektivitas distraktor. Pembahasan mengenai efektivitas distraktor tidak disertakan karena fitur yang disediakan Quiz question statistics hanya memuat frekuensi subjek yang memilih dari setiap opsi. Pada umumnya, efektivitas distraktor dievaluasi berdasarkan frekuensi subjek yang memilih opsi dan daya diskriminasi opsi. Bagi pembaca yang tertarik dalam pembahasan efektivitas distraktor dapat membaca artikel yang ditulis oleh Testa, Rosano, dan Toscato (2018).

Secara umum, analisis tes dan butir soal pada Moodle dapat dimanfaatkan untuk mengevaluasi tes yang akan digunakan menilai hasil belajar siswa. Baik itu tes yang sifatnya sumatif, maupun tes yang sifatnya formatif. Interpretasi yang dilakukan dapat menggunakan pendekatan teori tes klasik. Selain itu, analisis tes dan butir soal pada Moodle memiliki nilai praktis dan ekonomis bagi para guru. Mengingat analisis tes dan butir soal membutuhkan komputasi yang rumit. Selain itu, perangkat lunak komersil untuk analisis butir soal dan tes dijual dengan harga yang relatif mahal.



Gambar 1. Nilai Facility Index (FI) dan Discrimination Index (DI) 40 Butir Soal
Keterangan. Sumber data adalah dokumentasi pribadi.

4. Kesimpulan

Artikel ini menunjukkan bahwa beberapa formula estimasi yang digunakan dalam analisis tes dan butir soal pada Moodle identik dengan parameter pokok yang ada dalam teori tes klasik. Pertama, formula estimasi *Coefficient of Internal Consistency* (CIC) identik dengan koefisien alfa (α) dan *standard error* yang identik dengan *Standard Error of Measurement* (SEM). Kedua, Formula estimasi Facility Index (FI) identik dengan tingkat kesukaran butir (*p*). Ketiga, formula estimasi dan Discrimination Index (DI) identik dengan *point-biserial correlation* atau daya diskriminasi butir. Dengan demikian, analisis tes dan butir soal pada Moodle sebenarnya dapat diinterpretasikan menggunakan pandangan teori tes klasik.

5. Daftar Pustaka

Aulia, D., & Waspada, I. (2019). The design of exploratory application and preprocessing of event log data in LMS Moodle-based online learning activities for process mining. *Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika*, *5*(2), 124–133. https://doi.org/10.23917/khif.v5i2.8023

Azwar, S. (2012). *Reliabilitas dan validitas* (4 ed.). Pustaka Pelajar.

Azwar, S. (2015). *Dasar-dasar psikometrika* (2 ed.). Pustaka Pelajar.

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif* (1 ed.). Pustaka Pelajar.

Butcher, P. (2022, November 6). *Quiz report statistics*. MoodleDocs. https://docs.moodle.org/dev/Quiz_report_statistics

Collman, C. (2023, Februari 2). *Quiz statistics report*. MoodleDocs. https://docs.moodle.org/404/en/Quiz_statistics_report

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L., & Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418. https://doi.org/10.1177/0013164404266386

Febliza, A., Afdal, Z., Copriady, J., & Futra, D. (2021). Enhancement of quiz Moodle feature in developing e-self assessment communication skill. *2021 Universitas Riau International Conference on Education Technology (URICET)*, 356–362. https://doi.org/10.1109/URICET53378.2021.9865899

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, *24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Gamage, S. H. P. W., Ayres, J. R., Behrend, M. B., & Smith, E. J. (2019). Optimising

Moodle quizzes for online assessments. *International Journal of STEM Education*, *6*(1), 27. https://doi.org/10.1186/s40594-019-0181-4

Gómez-Soberón, J. M., Gómez-Soberón, M. C., Corral-Higuera, R., Arredondo-Rea, S. P., Almaral-Sánchez, J. L., & Cabrera-Covarrubias, F. G. (2013). Calibrating questionnaires by psychometric analysis to evaluate knowledge. *SAGE Open*, *3*(3), 215824401349915. https://doi.org/10.1177/2158244013499159

Hayat, B. (2021). Klasika: Program analisis item dan tes dengan pendekatan klasik. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia (JP3I)*, *10*(1), 1–11. https://doi.org/10.15408/jp3i.v10i1.20551

Hunt, T. (2013, September 25). *Quiz statistics calculations*. MoodleDocs. https://docs.moodle.org/dev/Quiz_statistics_calculations

Ikawati, V. (2015). Desain dan implementasi model pembelajaran e-learning di Program Studi Teknik Elektro Universitas 17 Agustus 1945 Cirebon dengan Modular Object Oriented Dynamic Learning Environment. *Emitor: Jurnal Teknik Elektro*, *15*(1), 15–21. https://doi.org/10.23917/emitor.v15i1.1754

Jacobs, P., & Viechtbauer, W. (2017). Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Research Synthesis Methods*, *8*(2), 161–180. https://doi.org/10.1002/jrsm.1218

López-Tocón, I. (2021). Moodle quizzes as a continuous assessment in higher education: An exploratory approach in physical chemistry. *Education Sciences*, *11*(9), 500. https://doi.org/10.3390/educsci11090500

Moodle. (2024). *Statistics*. Moodle Statistics. https://stats.moodle.org/

Pramita, M., Sukmawati, R. A., Purba, H. S., Wiranda, N., Kusnendar, J., & Sajat, M. S. (2021). Student acceptance of e-learning to improve learning independence in the Department of Computer Education. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *4*(1), 34–44. https://doi.org/10.23917/ijolae.v4i1.9265

Ramdani, Z., Widyastuti, T., & Ferdian, F. R. (2019). Penerapan analisis teori klasik, model Rasch, dan computer based test Moodle: Sebuah pilot studi. *Indonesian Journal of Educational Assesment*, *1*(2), 21. https://doi.org/10.26499/ijea.v1i2.9

Rice, W. (2015). *Moodle e-learning course development: A complete guide to create and develop engaging e-learning courses with Moodle* (3 ed.). Packt Publishing.

Setiawan, A., & Rahayu, D. N. F. A. (2022). SISENSI: QR code-based academic attendace system. *Urecol Journal. Part E: Engineering*, *2*(1), 29–36. https://doi.org/10.53017/uje.141

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology*, *9*, 1585. https://doi.org/10.3389/fpsyg.2018.01585

Tim Pusdiklat Pegawai Kemendikbud. (2016). *Penilaian hasil belajar*. Pusdiklat Pegawai Kemendikbud. https://pusdiklat.kemdikbud.go.id/file/e-publikasi/02.%20BAHAN%20AJAR/Modul%20Pelatihan%20Teknis/03.15%20Pelatihan%20Teknik%20Fasilitasi%20Melatih%20bagi%20Pamong%20Belajar/03.15%20Modul%20Pelatihan%20TFM%20bagi%20Pamong%20Belajar%2005.%20Penilaian%20Hasil%20Belajar.pdf

Waspada, I., Bahtiar, N., & Wibowo, A. (2019). Clustering student behavior based on quiz activities on moodle LMS to discover the relation with a final exam score. *Journal of Physics: Conference Series*, *1217*(1), 012118. https://doi.org/10.1088/1742-6596/1217/1/012118

Wibowo, S. (2016). Misuses cronbach alpha on achievement tests. *Assessment for Improving Students' Performance*, 355–359. https://www.researchgate.net/publication/304478310_Misuses_Cronbach_Alpha_On_Achievement_Tests