



ANALISIS BUTIR TES KETERAMPILAN BERPIKIR KRITIS DAN PENGETAHUAN PROSEDURAL GRAMMAR BAHASA INGGRIS

Herlia Alfiana, Hari Karyono, Wawan Gunawan

Universitas PGRI Adi Buana Surabaya, Kampus Menanggal, Jl. Dukuh Menanggal XII, Surabaya 60234 Jawa Timur, Indonesia

Article History

Received: 07-12-2020

Accepted: 09-12-2020

Published: 01-03-2021

Available online: 15-01-2021

Keywords

keterampilan berpikir kritis; pengetahuan prosedural; analisis butir; google form; bahasa Inggris

Abstrak

Penelitian ini berfokus pada analisis butir tes pilihan ganda untuk mengukur tes keterampilan berpikir kritis dan pengetahuan prosedural materi *grammar* untuk tingkat pemula (*beginner*) pada mata kuliah Bahasa Inggris. Analisis butir tes mencakup: (1) tingkat kesukaran butir tes, (2) tingkat daya pembeda butir tes, (3) validitas butir tes dan (4) reliabilitas tes. Untuk mengukur tingkat kesukaran digunakan rumus perbandingan jumlah *responden* yang menjawab benar dengan jumlah *responden*. Untuk mengukur validitas dan daya pembeda soal digunakan uji *Pearson product moment*. Kemudian, uji *Cronbach alpha* digunakan untuk menentukan reliabilitas tes. Hasil uji tingkat kesukaran menempatkan butir tes pada kategori mudah, sedang dan sukar. Hasil uji daya pembeda mengklasifikasikan butir tes pada kategori baik dan unggul. Uji validitas membuktikan seluruh butir tes valid. Uji reliabilitas menyatakan bahwa tes yang dirancang reliabel. Dengan demikian butir tes yang dirancang tidak membutuhkan perbaikan dan dapat digunakan untuk mengukur keterampilan berpikir kritis dan pengetahuan prosedural pada *grammar* Bahasa Inggris.

Abstract

This study focuses on evaluating the quality of multiple-choice tests to measure critical thinking skills and procedural knowledge tests of English grammar for beginners in higher education. The analysis of test items includes (1) item difficulty, (2) discrimination index, (3) validity, and (4) reliability of the tests. Item difficulty is obtained from the percentage comparison of the number of respondents who answer correctly with the total number of respondents. The validity and discrimination index of the test items are analyzed using the Pearson product-moment test. The Cronbach alpha test is used to evaluate the reliability of the test. The item difficulty test classified into easy, medium, and difficult categories. The results of the discrimination index test classify the test items into good and superior categories. The validity test proves that all test items are valid and the reliability test proves that the test is reliable. Finally, the test items have good quality and capable to measure critical thinking skills and procedural knowledge in English grammar.

Corresponding author: Herlia Alfiana
Adress: Universitas PGRI Adi Buana Surabaya, Kampus Menanggal,
Jl. Dukuh Menanggal XII, Surabaya 60234 Jawa Timur, Indonesia,
Instansi: Universitas PGRI Adi Buana Surabaya,
E-mail: herlia.albana@gmail.com

2021 Universitas Negeri Malang
p-ISSN 2406-8780
e-ISSN 2654-7953



PENDAHULUAN

Di tengah tantangan pandemi *Covid-19* ini, dosen perlu beradaptasi dengan cara baru mengevaluasi pembelajaran (Seifert, 2020) yang sesuai untuk *e-learning*. Hal ini dikarenakan, *e-learning* merupakan satu-satunya upaya untuk tetap mampu menyelenggarakan pembelajaran pada masa pandemi ini. Sehingga bentuk assessmen dalam pembelajaran perlu disesuaikan dengan pelaksanaan *e-learning* (Popchev & Orozova, 2020).

Pada penelitian ini dikembangkan alat ukur dalam bentuk *google form* untuk memudahkan akses bagi *responden* (Castro, 2018). *Google form* memudahkan akses bagi *responden* untuk menemukan dan mengerjakan tes karena lebih sederhana dan sering digunakan. Sarana ini dipilih karena memungkinkan untuk menerima dan merekap respon dalam jumlah besar.

Kemampuan berpikir kritis menjadi salah satu target capaian pembelajaran utama pada Pendidikan di perguruan tinggi (Liu et al., 2016). Selain itu, dalam berpikir kritis juga diperlukan pengetahuan prosedural mengenai bagaimana berpikir kritis dilakukan. Sehingga dua keterampilan ini saling berhubungan dan perlu dikembangkan oleh mahasiswa melalui pembelajaran (Sadhu et al., 2019) termasuk dalam mata kuliah Bahasa Inggris di perguruan tinggi (Karapetian, 2020). Kedua keterampilan ini saling berhubungan karena keduanya diperlukan untuk menyelesaikan permasalahan-permasalahan (Wechsler et al., 2018; Wuryaningrum et al., 2020). Selain itu kemampuan berbahasa Inggris dapat ditinjau dari segi keterampilan berpikir kritis (Haerazi et al., 2020) dan pengetahuan proseduralnya (Gunawan et al., 2019).

Dalam pembelajaran tidak banyak dosen yang mengevaluasi pengetahuan prosedural dan keterampilan berpikir kritis mahasiswa. Sehingga kesadaran dosen akan pengembangan keterampilan ini perlu ditingkatkan (Amin & Adiansyah, 2018). Untuk dapat mengevaluasi perkembangan keterampilan berpikir kritis dalam pembelajaran Bahasa Inggris, dibutuhkan kemampuan menyusun tes untuk mengukur keterampilan berpikir kritis. Akan tetapi, kemampuan pengajar dalam menyusun tes keterampilan berpikir kritis dalam pembelajaran

Bahasa Inggris masih kurang dan sangat membutuhkan pelatihan untuk menyusun tes tersebut (Laili et al., 2020).

Pada pembelajaran Bahasa Inggris, perkembangan keterampilan berpikir kritis dan pengetahuan prosedural tidak banyak diukur. Hal ini dikarenakan dalam pembelajaran Bahasa Inggris di perguruan tinggi lebih banyak berfokus pada perkembangan keempat keterampilan Bahasa yaitu *speaking*, *writing*, *reading* dan *listening* semata (Karyono, 2019; Khunaifi et al., 2020; Wahyuni, 2017). Pada studi sebelumnya pengembangan untuk tes berpikir kritis lebih banyak dilakukan untuk bidang studi IPA (Ningsih et al., 2018; Rosidah et al., 2018). Kemudian, dalam studi ini dikembangkan tes untuk mengukur keterampilan berpikir kritis dan pengetahuan prosedural untuk *grammar* pada mata kuliah Bahasa Inggris.

Tes keterampilan berpikir kritis dan pengetahuan prosedural dari kompetensi *grammar* dapat dikonstruksikan dalam bentuk pilihan ganda (Agust, 2019; Sahebkhair, 2020; Szabó, 2008, p. 77). Akan tetapi, kecacatan butir tes pilihan ganda dapat terjadi pada tahap merancang tes (Haladyna, 2004, p. 183; Quaigrain & Arhin, 2017). Sehingga wajib dilakukan analisis butir tes pada soal pilihan ganda sebelum soal digunakan (Hofer et al., 2017; Kusumawati & Hadi, 2018). Hasil dari analisis butir ini dapat digunakan untuk memperbaiki butir tes (Alamoudi et al., 2017). Selain itu, hasil analisis butir bermanfaat untuk memberikan umpan balik terhadap peningkatan kualitas pembelajaran dan hasil pembelajaran (Musa et al., 2018).

Tujuan dari penelitian ini adalah melakukan analisis butir tes terhadap tes yang telah dirancang untuk mengukur keterampilan berpikir kritis dan pengetahuan prosedural *grammar* Bahasa Inggris dengan pemanfaatan *google form*. Dalam analisis butir mencakup di dalamnya (1) analisis tingkat kesukaran butir tes, (2) analisis daya pembeda butir tes, (3) analisis validitas butir tes dan (4) analisis reliabilitas tes.

METODE

Jenis penelitian ini termasuk dalam penelitian deskriptif yang tujuannya adalah mendapatkan informasi mengenai tingkat

kesukaran, daya pembeda, reliabilitas dan validitas butir tes untuk mengukur kemampuan berpikir kritis dan pengetahuan prosedural pada *grammar* Bahasa Inggris yang disampaikan melalui *google form*. Butir soal yang telah dirancang diujikan pada 52 *responden* pada November 2020. Seluruh *responden* yang dilibatkan telah memiliki pengetahuan yang diperlukan yaitu materi *grammar* dalam Bahasa Inggris yang diujikan. Jumlah butir soal yang diujikan untuk dua keterampilan yang diukur masing-masing 10 butir, sehingga totalnya 20 butir.

Tes untuk mengukur kemampuan berpikir kritis pada *grammar* Bahasa Inggris disusun berdasarkan indikator pada tabel 1. Adapun indikator yang digunakan untuk mengukur pengetahuan prosedural pada *grammar* Bahasa Inggris sebagaimana pada tabel 2.

Untuk mengukur tingkat kesukaran butir tes dilakukan dengan membandingkan total *responden* yang menjawab dengan benar dengan jumlah seluruh *responden*. Kemudian dari hasil yang diperoleh dibandingkan dengan kriteria Indeks Kesukaran (IK) sebagaimana pada tabel 3.

Daya pembeda pada soal pilihan ganda merupakan kapasitas dari soal pilihan ganda untuk membedakan *responden* yang memperoleh skor tinggi dan yang rendah

(Mahjabeen et al., 2017). Untuk mengetahui daya pembeda soal dilakukan dengan mengklasifikasikan nilai r_{hitung} yang diperoleh dari uji *Pearson product moment* dari setiap butir tes terhadap indeks daya pembeda (IDP). Terdapat enam kategori yang digunakan untuk mengklasifikasikan daya pembeda soal. Berikut ini tabel 4 kategori indeks daya pembeda soal.

Uji validitas dapat dilakukan dengan melihat sebaran nilai yang diperoleh ketika soal diuji-cobakan. Untuk menentukan uji validitas digunakan uji *Pearson product moment*. Butir soal yang diujikan dapat dikatakan valid apabila nilai *pearson correlation* atau r_{hitung} yang diperoleh lebih besar daripada r_{tabel} . Selain itu, untuk menentukan validitas dapat juga disimpulkan dari kombinasi nilai *pearson correlation* dan signifikansi yang diperoleh. Apabila nilai *pearson correlation* bernilai positif serta didukung dengan nilai signifikansi yang kurang dari 0,05, maka butir tes dinyatakan valid.

Namun, apabila nilai signifikansinya sudah memenuhi syarat, yaitu kurang dari 0,05 tetapi *pearson correlation* bernilai *negative*, maka butir tes dikatakan tidak valid. Terlebih apabila nilai signifikansi dari hasil uji tersebut bernilai lebih dari 0,05 maka, butir tes dinyatakan tidak valid. Untuk mengukur reliabilitas instrumen digunakan uji *Cronbach alpha*.

Tabel 1 Indikator Keterampilan Berpikir Kritis

No	Indikator
1	Menganalisis ketepatan penggunaan <i>grammar</i> dalam kalimat Bahasa Inggris
2	Menentukan <i>part of speech</i> yang tepat dalam kalimat Bahasa Inggris
3	Menentukan kata kerja (<i>verb</i>) yang tepat untuk kalimat Bahasa Inggris

Tabel 2 Indikator Pengetahuan Prosedural

No	Indikator
1	Mengkonstruksi kalimat Bahasa Inggris yang tepat secara <i>grammar</i>
2	Mengevaluasi susunan kalimat Bahasa Inggris yang tepat secara <i>grammar</i>
3	Menentukan konstruksi kalimat Bahasa Inggris yang tepat secara <i>grammar</i>

Tabel 3 Kategori Indeks Kesukaran Butir Tes

IK	Kriteria
$\leq 0,20$	Sukar Sekali
0,21-0,30	Sukar
0,31-0,69	Sedang
0,70-0,80	Mudah
$\geq 0,81$	Mudah Sekali

Sumber: (Ahmed & Moalwi, 2017)

Tabel 4 Kategori Indeks Daya Pembeda Butir Soal

IDP	Kriteria
Negatif	Rusak
0- 0,19	Buruk
0,20-0,29	Sedang
0,30-0,39	Baik
=0,4	Baik sekali
>0,4	Unggul

Sumber: (Ahmed & Moalwi, 2017)

Untuk menentukan apakah hasil dari uji ini menghasilkan kesimpulan reliabel atau tidaknya instrumen dibutuhkan dasar pedoman. Suatu instrumen dinyatakan reliabel apabila nilai *Alpha Cronbach* yang diperoleh bernilai lebih besar dari 0,06. Sebaliknya, apabila nilai *Alpha Cronbach* yang diperoleh kurang dari 0,06, maka instrumen dinyatakan tidak reliabel.

HASIL

Hasil pengukuran tingkat kesukaran dari tiap-tiap butir soal ditampilkan dalam tabel 5. Dari hasil perhitungan diperoleh butir tes yang tergolong dalam kriteria mudah, sedang dan sukar. Dari pengukuran tingkat kesukaran tidak diperoleh adanya butir tes yang berada pada kategori mudah sekali dan sukar sekali. Butir tes yang termasuk dalam kategori sedang sebesar 85%. Untuk kategori sukar diperoleh persentase 10%. Sedangkan pada kategori mudah persentasenya lebih sedikit dibanding kategori yang lain yaitu sebesar 5%. Untuk menggambarkan proporsi butir soal pada masing-masing kategori dijelaskan pada gambar 1.

Dari hasil pengukuran daya pembeda soal yang menggambarkan kemampuan soal dalam membedakan *responden* dengan kemampuan tinggi dan rendah diperoleh dua kategori hasil yaitu baik dan unggul. Jumlah butir tes pada kategori baik adalah dua butir. Sedangkan pada

kategori unggul diperoleh 18 butir. Untuk menggambarkan hasil yang lebih rinci dipaparkan pada tabel sebagaimana tabel 6.

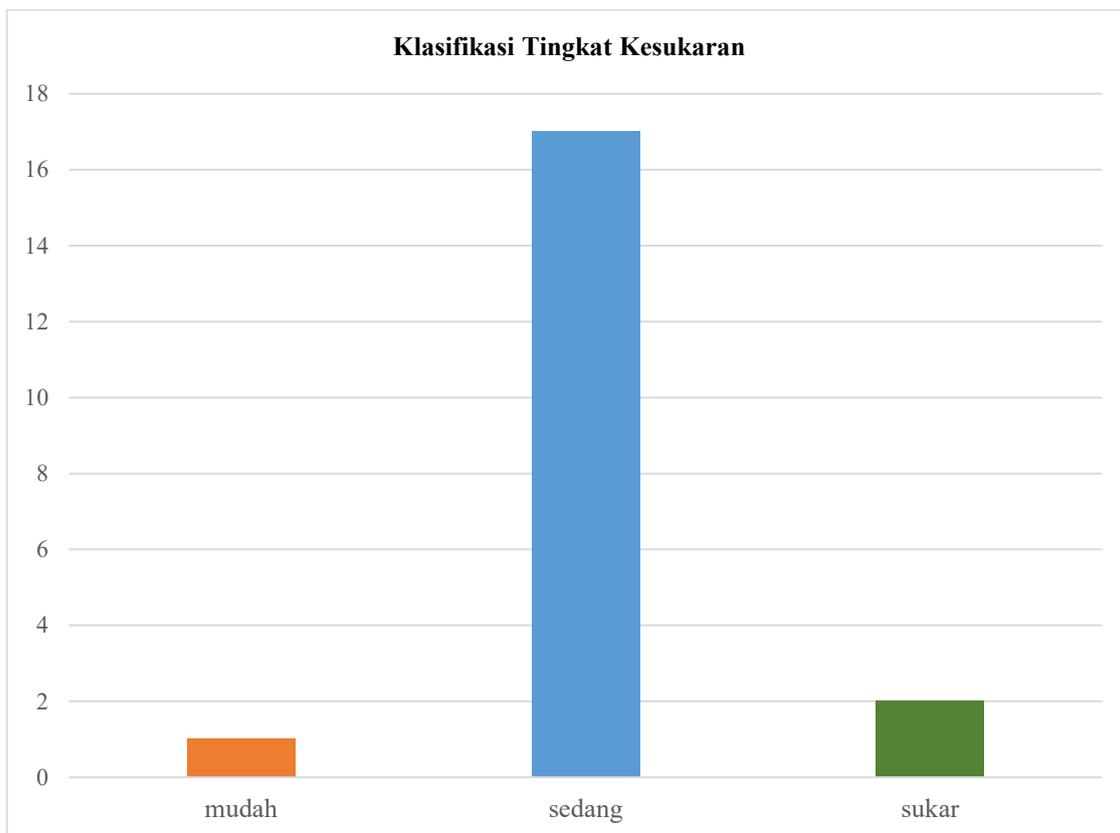
Dari tabel hasil pengukuran daya pembeda diatas diperoleh kesimpulan bahwa persentase butir yang termasuk dalam kategori unggul adalah 90%. Sedangkan butir soal dengan kategori baik sejumlah 10%. Perbandingan persentase keduanya dapat dilihat pada gambar 2.

Uji reliabilitas menggunakan uji *Cronbach alpha* menghasilkan nilai *Cronbach alpha* sebesar 0,868. Dari nilai yang dihasilkan tersebut dapat disimpulkan bahwa instrumen dinyatakan reliabel. Hal ini dirumuskan dengan dasar nilai *Cronbach alpha* yang didapat dari hasil pengujian lebih besar dari 0,6. Hasil uji *Cronbach alpha* disajikan pada tabel 7.

Untuk menyimpulkan valid dan tidaknya suatu butir tes, terdapat dua cara. *Pertama*, dengan membandingkan nilai *pearson correlation* atau r_{hitung} dengan r_{tabel} , apabila nilai r_{hitung} yang diperoleh dari pengujian lebih besar dari r_{tabel} maka butir tes dinyatakan valid ($r_{hitung} > r_{tabel}$). *Kedua*, validitas suatu butir tes dapat dilihat dari kombinasi nilai *pearson correlation* dan signifikansi yang diperoleh. Bila nilai *pearson correlation* dari hasil pengujian bernilai positif dan di dukung dengan pemerolehan nilai signifikansi yang bernilai kurang dari 0,05, maka butir tes dikatakan valid.

Tabel 5 Hasil Uji tingkat kesukaran

Butir Soal	Indeks Kesukaran	Kriteria
1	0,62	Sedang
2	0,38	Sedang
3	0,77	Mudah
4	0,62	Sedang
5	0,52	Sedang
6	0,63	Sedang
7	0,56	Sedang
8	0,48	Sedang
9	0,44	Sedang
10	0,58	Sedang
11	0,40	Sedang
12	0,56	Sedang
13	0,52	Sedang
14	0,44	Sedang
15	0,27	Sukar
16	0,50	Sedang
17	0,27	Sukar
18	0,40	Sedang
19	0,40	Sedang
20	0,62	Sedang



Gambar 1. Grafik Hasil Pengukuran Tingkat Kesukaran Butir Tes

Tabel 6 Hasil Uji Daya Pembeda

Butir Soal	IDP	Kriteria
1	0,53	Unggul
2	0,73	Unggul
3	0,46	Unggul
4	0,32	Baik
5	0,51	Unggul
6	0,61	Unggul
7	0,50	Unggul
8	0,54	Unggul
9	0,57	Unggul
10	0,47	Unggul
11	0,54	Unggul
12	0,65	Unggul
13	0,36	Baik
14	0,54	Unggul
15	0,55	Unggul
16	0,66	Unggul
17	0,63	Unggul
18	0,48	Unggul
19	0,61	Unggul
20	0,42	Unggul

Dari data pada tabel 8 tersebut dapat disimpulkan bahwa seluruh butir tes valid. Masing-masing butir tes memiliki nilai *Pearson correlation* atau r_{hitung} yang positif. Pada masing-masing butir tes didapatkan nilai signifikansi yang bernilai kurang dari 0,05.

Namun, apabila nilai signifikansinya sudah memenuhi syarat, yaitu kurang dari 0,05 tetapi *pearson correlation* bernilai negatif, maka butir soal dikatakan tidak valid. Terlebih lagi jika nilai signifikansi yang didapat lebih dari 0,05 maka, butir tes dinyatakan tidak valid. Uji korelasi *product moment Pearson* menunjukkan hasil pada tabel 8.



Gambar 2 Grafik Hasil Analisis Daya Pembeda Butir Tes

Tabel 7 Hasil Uji *Cronbach Alpha*

Nilai <i>Cronbach Alpha</i>	Jumlah Butir Soal
0,868	20

Tabel 8 Hasil Uji Validitas

Butir Soal	Nilai R _{hitung}	Nilai Signifikansi	Validitas
1	0,53	0,00	Valid
2	0,73	0,00	Valid
3	0,46	0,001	Valid
4	0,32	0,022	Valid
5	0,51	0,00	Valid
6	0,61	0,00	Valid
7	0,50	0,00	Valid
8	0,54	0,00	Valid
9	0,57	0,00	Valid
10	0,47	0,00	Valid
11	0,54	0,00	Valid
12	0,65	0,00	Valid
13	0,36	0,009	Valid
14	0,54	0,00	Valid
15	0,55	0,00	Valid
16	0,66	0,00	Valid
17	0,63	0,00	Valid
18	0,48	0,00	Valid
19	0,61	0,00	Valid
20	0,42	0,002	Valid

PEMBAHASAN

Assesmen pada pembelajaran Bahasa memiliki tiga fungsi (Cheng & Fox, 2017, p. 7). Pertama, fungsi instruksional assesmen yaitu memperoleh informasi mengenai kemampuan dan pemahaman mahasiswa. Kedua, fungsi *student-centered* yaitu mengidentifikasi kelebihan dan kekurangan yang dimiliki mahasiswa. Ketiga, fungsi *administrative* yaitu memberikan penilaian berbentuk skala numerik dan ordinal untuk pencapaian belajar mahasiswa.

Pada penelitian ini dikembangkan tes untuk mengukur keterampilan berpikir kritis dan pengetahuan prosedural dari materi *grammar* Bahasa Inggris dalam bentuk pilihan ganda. Tes yang telah dirancang disampaikan melalui *Google form*. Responden yang terlibat dalam uji coba tes ini tidak mengalami kesulitan dalam mengakses dan mengerjakan tes.

Dalam pembelajaran Bahasa Inggris keterampilan berpikir kritis dan pengetahuan prosedural perlu diakomodasi untuk dapat berkembang (Gunawan et al., 2019; Karapetian, 2020). Untuk dapat mengembangkan pembelajaran dalam rangka menunjang kedua keterampilan ini di dalam pembelajaran, dibutuhkan alat ukur untuk mengukur tingkat

keterampilan sebelum dan sesudah mengikuti proses pembelajaran (Fauzi et al., 2019). Sehingga perkembangan dari keterampilan ini dapat diukur. Kemudian instrumen pengukuran ini dikembangkan dengan tujuan untuk menentukan kemampuan awal dan kemampuan akhir setelah mengikuti proses pembelajaran.

Data mengenai kemampuan awal diperlukan untuk merancang pembelajaran yang sesuai dengan mahasiswa. Selain itu data ini juga dibutuhkan untuk mengetahui kebutuhan belajar masing-masing mahasiswa. Pemenuhan kebutuhan belajar mahasiswa dapat memungkinkan adanya pembelajaran yang efektif (Rufii & Rochmawati, 2019).

Data kemampuan akhir dibutuhkan untuk mengevaluasi apakah pembelajaran Bahasa Inggris yang telah dirancang efektif (Nalendra et al., 2020; Nurvina Sularso, 2020; Staub, 2017) untuk meningkatkan keterampilan berpikir kritis (Jumaya & Evyanto, 2018; Nastiti, 2020) dan pengetahuan prosedural (Degeng et al., 2017; Gunawan et al., 2019). Hasil evaluasi menjadi dasar untuk memperbaiki program pembelajaran yang telah dirancang (Rahayu et al., 2018) sehingga proses pembelajaran dapat ditingkatkan kualitas dan efektivitasnya, khususnya dalam peningkatan kualitas *e-learning* Bahasa Inggris (As'at et al., 2017; Wills & Baker, 2017, 2018).

Dalam menyusun butir tes untuk pembelajaran Bahasa, nilai total mampu merepresentasikan bagian dari kompetensi Bahasa yang diukur (Szabó, 2008, p. 65). Oleh sebab itu, tes guna mengukur keterampilan berpikir kritis dan pengetahuan prosedural yang dikembangkan untuk pembelajaran Bahasa Inggris secara spesifik dilihat dari segi kompetensi *grammar*. Tes untuk mengevaluasi keterampilan berpikir kritis dan pengetahuan prosedural dari kompetensi *grammar* dapat dikonstruksikan dalam bentuk pilihan ganda (Agust, 2019; Sahebkheir, 2020; Szabó, 2008, p. 77).

Tes dalam bentuk pilihan ganda adalah salah satu bentuk tes yang paling efektif (Gierl et al., 2017) dan memudahkan dalam penghitungan nilai. Selain itu, bentuk tes pilihan ganda

memberikan ruang lingkup berpikir untuk *responden* dengan menyajikan beberapa kemungkinan jawaban. Manfaat lain dari tes pilihan ganda yaitu: (1) kelebihanannya dalam mengumpulkan data dengan jumlah sampel yang besar, (2) memiliki penilaian yang obyektif yang akurat dan mudah dihitung, (3) soal pilihan ganda terbukti mempunyai tingkat reliabilitas dan validitas yang lebih tinggi dibanding model tes yang lain (Khairani & Shamsuddin, 2016).

Suatu instrumen pilihan ganda harus memiliki konsistensi yang sama ketika digunakan berkali-kali (Creswell & Creswell, 2018, p. 215). Konsistensi hasil tes ini disebut *reliability* (Leavy, 2017, p. 113). Namun, instrumen yang reliabel belum tentu valid (Garson, 2013, p. 100). Artinya, reliabilitas ini penting tetapi tidak cukup untuk menggambarkan validitas. Validitas menggambarkan tingkat ketelitian suatu instrumen dalam mengukur capaian yang hendak diukur. Baik tidaknya suatu tes ditunjukkan dari validitasnya (Creswell & Creswell, 2018, p. 215). Oleh sebab itu, sebelum digunakan suatu instrumen harus diuji validitasnya sebelum digunakan untuk mengumpulkan data (Sukestini et al., 2020).

Terdapat dua *independent parameter* yang harus diukur pada tes, yaitu tingkat kesukaran dan kemampuan individu (Alamoudi et al., 2017; Szabó, 2008, p. 52). Kemampuan individu digambarkan dalam daya beda soal. Daya beda soal menunjukkan kemampuan yang dimiliki suatu butir tes dalam membedakan *responden* yang berkemampuan tinggi dan sebaliknya (Burud et al., 2019; Son, 2019). Butir tes dengan daya pembeda rendah mengandung kata yang ambigu sehingga perlu diperbaiki (Quaigrain & Arhin, 2017).

Tingkat kesukaran pada butir tes menggambarkan besaran probabilitas *item* tersebut dapat dijawab dengan benar oleh *responden*. Tingkat kesukaran diperoleh dengan membandingkan jumlah *responden* yang menjawab dengan benar dengan jumlah total *responden*. Tingkat kesukaran pada soal pilihan ganda dipengaruhi oleh letak *distractor* yang paling menarik pada pilihan jawaban (Shin et al., 2020).

Penelitian ini berfokus pada analisis butir tes yang telah dikembangkan. Analisis butir tes pilihan ganda dilakukan dengan cara mengumpulkan, menyimpulkan dan menggunakan informasi dari para *responden* yang bertujuan untuk mengukur kualitas butir tes serta kualitas suatu tes secara keseluruhan (Ahmed & Moalwi, 2017; Pawade & Diwase, 2016). Dari hasil uji coba diperoleh data reliabilitas tes, validitas, tingkat kesukaran, dan daya pembeda butir tes yang digunakan sebagai data untuk melakukan analisis butir tes.

Hasil analisis tingkat kesukaran butir tes menunjukkan tingkatan kesulitan butir tes berada pada tiga klasifikasi yaitu mudah, sedang dan sukar. Jumlah butir tes terbanyak ada pada kategori sedang dengan persentase 85%. Jadi butir tes memiliki kemungkinan yang cukup tinggi untuk dapat diselesaikan dengan benar.

Kemudian hasil pengujian daya pembeda soal menggolongkan butir tes ke dalam dua kategori yaitu baik dan unggul. Jumlah butir tes kategori unggul lebih banyak dari pada kategori baik yaitu 90% unggul dan 10% baik. Artinya butir tes memiliki kemampuan pembeda yang sangat baik dalam membedakan *responden* berkemampuan tinggi dan rendah. Dengan demikian butir tes yang telah dirancang tidak memerlukan perbaikan.

Dalam menyusun butir tes penting untuk memastikan reliabilitasnya karena menentukan homogenitas tes pada konstruk yang diukur (Adhi & Aly, 2018). Dari hasil analisis tes uji reliabilitas dengan uji *Cronbach alpha* diperoleh nilai *Cronbach alpha* sebesar 0,868. Maka tes yang dirancang dinyatakan reliabel. Dengan demikian, butir tes tersebut bersifat homogen. Artinya, jika digunakan pada *responden* yang lain akan memperoleh hasil yang sama.

Butir tes yang valid dibutuhkan untuk meningkatkan kualitas evaluasi dan mengembangkan bank soal yang sesuai (Ahmed & Moalwi, 2017; Musa et al., 2018). Berdasarkan uji validitas *Pearson product moment*, seluruh butir tes terbukti valid. Dengan demikian seluruh butir tes dapat diterima dan menjadi bagian dari bank soal untuk dapat digunakan seterusnya (Haladyna, 2004, p. 202). Artinya, tes yang telah teruji ini dapat digunakan

dalam pengukuran keterampilan berpikir kritis dan pengetahuan prosedural pada pembelajaran Bahasa Inggris untuk materi *grammar* pada tingkat *beginner*.

SIMPULAN

Tes untuk mengukur keterampilan berpikir kritis dan pengetahuan prosedural dari kompetensi *grammar* Bahasa Inggris dapat dikonstruksikan dalam bentuk pilihan ganda. Akan tetapi, pada umumnya butir tes yang dirancang memiliki kecacatan sehingga wajib dilakukan analisis butir sebelum tes tersebut digunakan. Penelitian ini berfokus pada analisis butir tes yang mencakup tingkat kesukaran, daya pembeda, reliabilitas dan validitas.

Untuk memperoleh data analisis butir tes soal diujikan pada 52 *responden*. Tingkat kesukaran diperoleh dengan membandingkan jumlah *responden* yang menjawab benar dengan jumlah seluruh *responden*. Hasil uji tingkat kesukaran menunjukkan butir tes berada pada kategori sukar, sedang dan mudah. Uji daya beda soal yang dilakukan dengan mengklasifikasikan nilai t_{hitung} yang diperoleh dari uji *Pearson product moment* dari setiap butir tes terhadap indeks daya pembeda, menyatakan butir tes termasuk kategori baik dan unggul.

Uji reliabilitas tes dilakukan dengan uji *Cronbach alpha*. Dari uji tersebut, tes pilihan ganda yang dirancang dinyatakan reliabel. Kemudian uji validitas butir tes yang dilakukan dengan uji validitas *Pearson product moment*, memperoleh hasil valid untuk semua butir tes yang dirancang.

Penelitian ini dapat dikembangkan dalam penelitian lanjutan. Tes *grammar* Bahasa Inggris yang telah dikembangkan untuk keterampilan berpikir kritis dan pengetahuan prosedural yang telah teruji tingkat kesukaran, daya beda, validitas serta reliabilitasnya dapat digunakan untuk mengukur kedua keterampilan tersebut pada studi eksperimen. Pada studi eksperimen, tes ini dapat digunakan sebagai instrumen pengukuran untuk mengkaji efektivitas atau pengaruh suatu strategi pembelajaran.

DAFTAR RUJUKAN

Adhi, M. I., & Aly, S. M. (2018). Student Perception and Post-exam Analysis of One Best MCQs

and One Correct MCQs: A Comparative Study. *Journal of the Pakistan Medical Association*, 68(4), 570–575.

- Agust, S. (2019). How Does Rasch Model Reveal Dishonesty between Coastal Students and Easy Grammar Test? *Jurnal Iqra': Kajian Ilmu Pendidikan*, 4(2), 214–230.
- Ahmed, I. A. M., & Moalwi, A. A. (2017). Correlation Between Difficulty and Discrimination Indices of MCQs Type A in Formative Exam in Anatomy. *Journal of Research & Method in Education*, 7(5), 28–43.
- Alamoudi, A. A., El-Deek, B. S., Park, Y. S., Al-Shawwa, L. A., & Tekian, A. (2017). Evaluating The Long-term Impact of Faculty Development Programs on MCQ Item Analysis. *Journal Medical Teacher*, 39(sup1), S45--S49.
- Amin, A. M., & Adiansyah, R. (2018). Lecturers' Perception on Students' Critical Thinking Skills Development and Problems Faced by Students in Developing Their Critical Thinking Skills. *Jurnal Pendidikan Biologi Indonesia*, 4(1), 1.
- As'at, M., Setyosari, P., & Ulfa, S. (2017). The Development of English Instructional Design Social Learning Network (SLN)-Based Using Auction (Interested-based Negotiation) Strategy. *JINOTEP (Jurnal Inovasi Dan Teknologi Pembelajaran)*, 4(1), 2–7.
- Burud, I., Nagandla, K., & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences*, 7(4), 1136.
- Castro, S. (2018). Google Forms Quizzes and Substitution, Augmentation, Modification, and Redefinition (SAMR) Model Integration. *Issues and Trends in Educational Technology*, 6(2), 4–14.
- Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning*. Palgrave.
- Creswell, J. W., & Creswell, J. D. (2018). Research Design Qualitative, Quantitative, and Mixed Methods Approaches. In *Fast Facts to Loving your Research Project*. Sage Publications.
- Degeng, M. D. K., Setyosari, P., Degeng, I. N. S., & Kuswandi, D. (2017). Pengaruh Learning Control dalam Pembelajaran menggunakan Media Web Terhadap Hasil Belajar Pengetahuan Prosedural. *Jurnal Pendidikan Dan Pembelajaran (JPP)*, 23(2), 90–95.
- Fauzi, M. (2019). Pengaruh Strategi Pembelajaran Swa-Atur Dengan Discovery Learning dan Gaya Kognitif Terhadap Hasil Belajar Kimia. *Edcomtech Jurnal Kajian Teknologi Pendidikan*, 4(1), 56-66
- Garson, G. D. (2013). *Validity and Reliability*. Statistical Associates Publishing. www.statisticalassociates.com

- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87(6), 1082–1116.
- Gunawan, W., Degeng, I. N. S., Utaya, S., & Sulton. (2019). The Improvement of Conceptual and Procedural Understanding by Scaffolding with Responsiveness. *International Journal of Innovation, Creativity and Change*, 7(5), 162–179.
- Haerazi, Utama, I. M. P., & Hidayatullah, H. (2020). Mobile Applications to Improve English Writing Skills Viewed from Critical Thinking Ability for Pre-service Teachers. *International Journal of Interactive Mobile Technologies*, 14(7), 58–72.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.
- Hofer, S. I., Schumacher, R., & Rubin, H. (2017). The Test of Basic Mechanics Conceptual Understanding (bMCU): Using Rasch Analysis to Develop and Evaluate an Efficient Multiple Choice Test on Newton's Mechanics. *International Journal of STEM Education*, 4(1), 1–20.
- Jumaya, A., & Evyanto, W. (2018). Efektivitas Project Based Learning Berbasis ICT Terhadap Kemampuan Critical Thinking Mahasiswa EFL. *Seminar Nasional Ilmu Sosial Dan Teknologi*, 1, 19–24.
- Karapetian, A. O. (2020). Creating ESP-based Language Learning Environment to Foster Critical Thinking Capabilities in Students' Papers. *European Journal of Educational Research*, 9(2), 717–728.
- Karyono, H. (2019). Pengaruh Penggunaan Metode Kooperatif Tipe Jigsaw dan KWL Terhadap Kemampuan Membaca Teks Report Siswa Kelas IX SMP Negeri 2 Balongpanggang. *Jurnal Education and Development*, 7(3), 7.
- Khairani, A. Z., & Shamsuddin, H. (2016). Assessment for Learning Within and Beyond The Classroom. In *Assessment for Learning Within and Beyond the Classroom* (pp. 417–426). Springer.
- Khunaifi, A. R., Ilham, & Suratmoko, B. (2020). Pelatihan Bahasa Inggris Berequivalent TOEFL Pada Reading Test CBT Terhadap Mahasiswa Semester Akhir di Universitas Muhammadiyah. 5(2), 192–198.
- Kusumawati, M., & Hadi, S. (2018). An Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from Mathematics Assessments in Senior High School. *REiD (Research and Evaluation in Education)*, 4(1), 70–78.
- Laili, M., Aini, N., & Christanti, A. (2020). Higher Order Thinking Skills (HOTS) Dalam Penilaian Bahasa Inggris Siswa SMA. *Lintang Songo: Jurnal Pendidikan*, 3(1), 18–25.
- Leavy, P. (2017). *Research Design Quantitative, Qualitative, Art-Based, and Community-Based Participatory Research Approaches*. The Guilford Press.
- Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing Critical Thinking in Higher Education: The HEIghten™ Approach and Preliminary Validity Evidence. *Assessment and Evaluation in Higher Education*, 41(5), 677–694.
- Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Konain, S., & Rizvi, M. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, 9(3), 109.
- Musa, A., Shaheen, S., Elmardi, A., & Ahmed, A. (2018). Item Difficulty & Item Discrimination as Quality Indicators of Physiology MCQ Examinations at The Faculty of Medicine Khartoum University. *Khartoum Medical Journal*, 11(2), 1477–1468.
- Nalendra, R. W., Wiyokusumo, I., & Leksono, I. P. (2020). The Effect of Learning Model with Youtube Versus Media Flash Card Media and Learning Motivation on Learning. *Journal of English Educational Study*, 3(2), 129–139.
- Nastiti, A. A. (2020). An Analysis of English Textbook Focusing on Reading Material to Enhance Students' Critical Thinking in Recount Text and Narrative Text. *RETAIN*, 8(3), 69–79.
- Ningsih, D. R., Ramalis, T. R., & Purwana, U. (2018). Pengembangan Tes Keterampilan Berpikir Kritis Berdasarkan Analisis Teori Respon Butir. *WaPFi (Wahana Pendidikan Fisika)*, 3(2), 45.
- Nurvina Sularso, A. (2020). Evaluation of Online Learning for English Practice Course in Akademi Telkom Jakarta. *Journal of Informatics and Communications Technology*, 21(1), 59–067.
- Pawade, Y. R., & Diwase, D. S. (2016). Can Item Analysis of MCQs Accomplish The Need of a Proper Assessment Strategy for Curriculum Improvement in Medical Education? *Journal of Educational Technology*, 13(1), 44–53.
- Popchev, I. P., & Orozova, D. A. (2020). Towards a Multistep Method for Assessment in e-Learning of Emerging Technologies. *Cybernetics and Information Technologies*, 20(3), 116–129.
- Quaigrain, K., & Arhin, A. K. (2017). Using Reliability and Item Analysis to Evaluate a Teacher-Developed Test in Educational Measurement and Evaluation. *Cogent Education*, 4(1), 1–11.
- Rahayu, A., Abidin, Z., & Susilaningsih, S. (2018).

- Evaluasi Program Pembelajaran Full Day School Di SDN Bunulrejo 2 Malang. *JINOTEP (Jurnal Inovasi dan Teknologi Pembelajaran): Kajian dan Riset Dalam Teknologi Pembelajaran*, 4(2), 82-87.
- Rosidah, N. A., Ramalis, T. R., & Suyana, I. (2018). Karakteristik Tes Keterampilan Berpikir Kritis Berdasarkan Teori Respon Butir. *Jurnal Inovasi Dan Pembelajaran Fisika*, 5(1), 54–63.
- Rufii, R., & Rochmawati, D. (2019). Evaluation of Universal Design for Constructivist-based Statistics Learning Module for Students' Increased Motivation. *Journal of Education and Learning (EduLearn)*, 13(3), 431–440.
- Sadhu, S., Ad'hiya, E., & Laksono, E. W. (2019). Exploring and Comparing Content Validity and Assumptions of Modern Theory of An Integrated Assessment: Critical Thinking-Chemical Literacy Studies. *Jurnal Pendidikan IPA Indonesia*, 8(4), 570–581.
- Sahebkheir, F. (2020). Improving Grammar Achievement through Using Metatalk Considering Iranian EFL Learners with Advanced Language Proficiency. *South Asian Research Journal of Arts, Language and Literature*, 2(4), 56–59.
- Seifert, T. (2020). Student Assessment in Online Learning : Challenges and Effective Practices During Covid-19. *Proceedings of EdMedia + Innovate Learning*, pp. 106-10(2001), 1–4.
- Shin, J., Bulut, O., & Gierl, M. J. (2020). The Effect of The Most-Attractive-Distractor Location on Multiple-Choice Item Difficulty. *Journal of Experimental Education*, 88(4), 643–659. <https://doi.org/10.1080/00220973.2019.1629577>
- Son, A. L. (2019). Instrumentasi Kemampuan Pemecahan Masalah Matematis: Analisis Reliabilitas, Validitas, Tingkat Kesukaran Dan Daya Beda Butir Soal. *Gema Wiralodra*, 10(1), 41–52.
- Staub, D. F. (2017). Quality Assurance and Foreign Language Programme Evaluation. In *Evaluation in Foreign Language Education in the Middle East and North Africa* (pp. 273–292). Springer.
- Sukestini, E., Fatirul, A. N., & Hartono, H. (2020). Problem Based Learning with ICT Based with Learning Creativity to Improve History Learning Achievement. *Jurnal Pendidikan Dan Pengajaran*, 53(3), 227.
- Szabó, G. (2008). *Applying Item Response Theory in Language Test Item Bank Building* (Vol. 10). Peter Lang.
- Wahyuni, F. A. (2018). Blended Learning: Dua Metode (Synchronous and Asynchronous) Untuk Matakuliah Writing Materi Argumentative Essay. *JINOTEP (Jurnal Inovasi dan Teknologi Pembelajaran): Kajian dan Riset Dalam Teknologi Pembelajaran*, 3(2), 137-143.
- Wechsler, S. M., Saiz, C., Rivas, S. F., Vendramini, C. M. M., Almeida, L. S., Mundim, M. C., & Franco, A. (2018). Creative and Critical Thinking: Independent or Overlapping Components? *Thinking Skills and Creativity*, 27(December 2017), 114–122.
- Wills, T., & Baker, C. (2018). Engaging Students With Synchronous Online Learning Activities. *Innovations in Teaching & Learning Conference Proceedings*, 10.
- Wills, T., & Baker, C. (2017). Strategies to Make Online Teaching Better Than Face-To-Face. *Innovations in Teaching & Learning Conference Proceedings*, 9.
- Wuryaningrum, R., Bektiarso, S., & Suyitno, I. (2020). The Effects of Knowledge-Transforming Text on Elementary Students' Declarative, Procedural Knowledge, and Motivation in Environmental Learning. *International Journal of Instruction*, 13(1), 567–586.