Cakrawala

Jurnal Pendidikan

Volume 19 No 1 (2025)

http://cakrawala.upstegal.ac.id/index.php/Cakrawala

email: cakrawala.upstegal@gmail.com

Pengembangan Soal Esai Berbasis Model 4D: Analisis Validitas, Reliabilitas, Daya Pembeda, dan Tingkat Kesukaran pada Evaluasi Pembelajaran

[1]Laminah, [2] Muhammad Habibi.

[1]Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

[2]Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

Email: 22311023644@students.uin-suska.ac.id

Info Artikel

Diterima Januari

Disetujui Februari

Direvisi April

Dipublikasikan Mei

DOI:

ISSN:2549-9300 (Online) | ISSN: 1858-4497 (Print)

Abstract

This research aims to develop high quality essay questions using the Research and Development (R&D) method with the 4D model (Define, Design, Develop, and Disseminate). This methodology is combined with a descriptive quantitative approach to ensure the measurement of validity, reliability, distinguishing power and level of difficulty of the questions. The research method uses quantitative analysis of data from test results consisting of 15 questions. Validity was analyzed using the correlation value between the item scores and the total score, reliability was calculated using the Cronbach's Alpha formula, while the level of difficulty and differentiating power were analyzed based on the scores of students in the upper and lower groups. The research results showed that of the 15 questions, 12 questions were declared valid and 3 questions were invalid. The reliability of the questions is in the very good category with a value of 0.871. Most of the questions have a medium level of difficulty (10 questions), while 3 questions are in the easy category. Based on differentiating power, 3 questions are categorized as good, 8 questions are sufficient, and 1 question is bad. Overall, this evaluation question is quite good to use.

***Keywords**: Question development, question analysis, 4D Model*

Pengembangan Soal Esai Berbasis Model 4D: Analisis Validitas, Reliabilitas, Daya Pembeda, dan Tingkat Kesukaran pada Evaluasi Pembelajaran

Abstrak

Penelitian ini bertujuan untuk mengembangkan soal esai berkualitas tinggi melalui metode Research and Development (R&D) dengan model 4D (Define, Design, Develop, dan Disseminate). Metodologi ini dipadukan dengan pendekatan kuantitatif deskriptif untuk menjamin pengukuran validitas, reliabilitas, daya pembeda, dan tingkat kesukaran soal. Metode penelitian menggunakan analisis kuantitatif terhadap data hasil uji coba soal yang terdiri dari 15 butir soal. Validitas dianalisis menggunakan nilai korelasi antara skor butir dengan skor total, reliabilitas dihitung menggunakan rumus Alpha Cronbach, sementara tingkat kesukaran dan daya pembeda dianalisis berdasarkan skor siswa kelompok atas dan bawah. Hasil penelitian menunjukkan bahwa dari 15 butir soal, 12 soal dinyatakan valid dan 3 soal tidak valid. Reliabilitas soal berada pada kategori sangat baik dengan nilai 0,871. Sebagian besar soal memiliki tingkat kesukaran sedang (10 soal), sementara 3 soal masuk kategori mudah. Berdasarkan daya pembeda, 3 soal berkategori baik, 8 soal cukup, dan 1 soal jelek. Secara keseluruhan, soal evaluasi ini cukup baik untuk digunakan.

Kata Kunci:Pengembangan soal, analisis soal, Model 4D

PENDAHULUAN

Evaluasi pembelajaran adalah komponen penting dalam proses pendidikan, terutama untuk mengukur tingkat pencapaian hasil belajar siswa (Arikunto, 2018). Instrumen penilaian, seperti soal tes, perlu memenuhi standar validitas, reliabilitas, tingkat kesukaran, dan daya pembeda yang memadai untuk memberikan gambaran yang tepat mengenai kemampuan siswa (Brookhart & Mcmillan, 2020). Validitas memastikan bahwa soal benar-benar mengukur apa yang ingin diukur, sehingga hasil tes dapat diandalkan (Sugiyono, 2022). Reliabilitas menjamin konsistensi hasil tes jika dilakukan pengukuran ulang (Sudaryono, 2021). Tingkat kesukaran soal harus disesuaikan dengan kemampuan siswa agar tidak terlalu mudah atau terlalu sulit (Sumaryanta, 2021).

Sementara itu, daya pembeda menunjukkan kemampuan soal dalam membedakan siswa yang memiliki kemampuan tinggi dengan siswa yang memiliki kemampuan rendah (Idrus, 2019). Dengan demikian, keempat konsep ini saling melengkapi dan menjadi tolok ukur kualitas suatu instrumen penilaian. Daya pembeda membantu mengidentifikasi kemampuan siswa berdasarkan tingkat kesulitan yang sesuai. Selain itu, tingkat kesukaran menunjukkan sejauh mana soal dapat menjangkau beragam tingkat kemampuan siswa. Gabungan dari keempat elemen ini menghasilkan instrumen penilaian yang adil, terpercaya,

dan relevan. Oleh sebab itu, analisis menyeluruh terhadap kualitas soal menjadi kunci untuk menciptakan evaluasi yang efektif serta meningkatkan hasil belajar siswa.

Namun, dalam implementasinya, proses penyusunan soal esai sering menghadapi sejumlah tantangan, termasuk kurangnya kualitas soal yang tidak memenuhi standar validitas, reliabilitas, daya pembeda, dan tingkat kesukaran. Salah satu pendekatan yang dapat digunakan untuk mengatasi permasalahan ini adalah pengembangan instrumen secara sistematis melalui model *Four-Dimensional* (4D). Model ini terdiri atas empat tahap utama, yaitu *Define*, *Design*, *Develop*, dan *Disseminate*, yang dirancang untuk menghasilkan produk berkualitas secara terstruktur (Thiagarajan, Semmel, &Semmel, 1974). Model ini telah terbukti efektif dalam berbagai bidang pengembangan, termasuk dalam pembuatan instrumen evaluasi pembelajaran.

Meskipun demikian, penelitian mengenai pengembangan soal esai berbasis model 4D masih tergolong terbatas, terutama yang menggabungkan analisis kualitas soal dari dimensi validitas, reliabilitas, daya pembeda, dan tingkat kesukaran. Hal ini menjadi sangat penting, mengingat keempat dimensi tersebut merupakan indikator kunci dalam menilai efektivitas dan keandalan suatu instrumen evaluasi (Wess et al., 2021). Oleh karena itu, penelitian ini bertujuan untuk mengembangkan dan menganalisis soal esai yang berkualitas tinggi dengan mempertimbangkan empat dimensi utama, yaitu validitas, reliabilitas, daya pembeda, dan tingkat kesukaran.

Pengembangan soal esai dilakukan untuk memastikan bahwa instrumen evaluasi mampu mengukur kemampuan siswa secara komprehensif dan sesuai dengan tujuan pembelajaran. Analisis mendalam terhadap keempat dimensi ini bertujuan untuk mengidentifikasi kelemahan sekaligus memperbaiki aspek-aspek yang memengaruhi kualitas soal. Validitas memastikan soal relevan dengan kompetensi yang diukur, sedangkan reliabilitas menjamin konsistensi hasil penilaian. Daya pembeda membantu memisahkan siswa dengan kemampuan tinggi dan rendah, sementara tingkat kesukaran menentukan keseimbangan soal. Dengan memperhatikan semua dimensi tersebut, penelitian ini diharapkan dapat menghasilkan instrumen evaluasi yang adil, obyektif, dan efektif.

Hasil penelitian ini juga dapat memberikan rekomendasi bagi pendidik untuk meningkatkan kualitas evaluasi. Dengan demikian, pengembangan ini berkontribusi pada peningkatan mutu pembelajaran dan hasil belajar siswa secara keseluruhan. Lebih jauh, hasil penelitian ini diharapkan dapat menjadi acuan bagi pengembang soal dalam menciptakan instrumen evaluasi yang lebih efektif, relevan, dan sesuai dengan kebutuhan kurikulum. Dengan adanya pendekatan ini, diharapkan tercipta lingkungan pembelajaran yang mendukung pencapaian kompetensi siswa secara optimal. Penelitian ini pada akhirnya berupaya memberikan dampak positif bagi peningkatan kualitas pendidikan di berbagai jenjang.

METODE

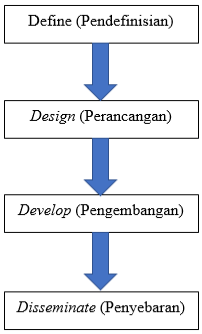
Penelitian ini menggunakan metode *Research* and *Development* (R&D) dengan model 4D yang meliputi tahapan *Define, Design, Develop,* dan *Disseminate*(Hariyanto et al., 2022) dan mengombinasikan pendekatan kuantitatif deskriptif. Model 4D dipilih karena pendekatannya yang terstruktur dalam menghasilkan produk yang sahih dan efisien dalam konteks pendidikan (Thiagarajan, Semmel, & Semmel, 1974). Metode ini memungkinkan peneliti untuk merancang, menguji, dan memperbaiki instrumen evaluasi secara sistematis, sehingga dapat mencapai standar kualitas yang diharapkan. Dengan pendekatan ini, diharapkan soal

ISSN:2549-9300 (Online) | ISSN: 1858-4497 (Print)

yang dikembangkan memenuhi standar kualitas yang diperlukan untuk mendukung pembelajaran yang efektif.

Tahap ***Define*** pada tahap ini, dilakukan analisis terhadap kebutuhan soal, yang mencakup identifikasi tujuan pembelajaran dan kompetensi yang akan diukur. Selain itu, dilakukan juga evaluasi terhadap soal-soal sebelumnya untuk mengidentifikasi kelemahan dan kekurangan yang perlu diperbaiki. Sedangkan tahap ***Design***, soal disusun berdasarkan indikator pembelajaran yang telah ditetapkan. Penyusunan soal dilakukan dengan memperhatikan keselarasan antara soal dan tujuan pembelajaran yang ingin dicapai, serta memastikan kesesuaian materi dengan tingkat kesulitan soal. Pada tahap ***Develop***, soal yang telah dirancang diuji coba. Data yang diperoleh dari uji coba akan dianalisis untuk mengevaluasi validitas, reliabilitas, daya pembeda, dan tingkat kesulitan soal.

Berdasarkan hasil analisis tersebut, soal yang tidak memenuhi kriteria kualitas akan direvisi. Tahap terakhir, ***Disseminate,*** pada tahap ini, soal yang telah diperbaiki disebarkan kepada guru atau sekolah untuk digunakan sebagai alat evaluasi dalam pembelajaran. Penyebaran dilakukan setelah soal dianggap memenuhi standar kualitas yang ditetapkan.mencakup diseminasi soal yang telah disempurnakan kepada guru atau sekolah. Dalam data analisis, penelitian ini memanfaatkan pendekatan kuantitatif deskriptif, yang digunakan untuk mengukur validitas, reliabilitas, daya pembeda, dan tingkat kesukaran soal.



Gambar 1. Tahap Pengembangan Model 4-D

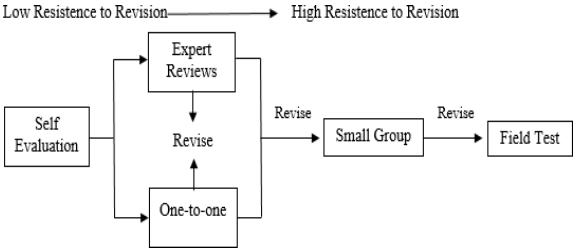
Kombinasi metode Research and Development (R&D) dan analisis kuantitatif deskriptif ini diharapkan dapat menghasilkan soal esai yang valid, reliabel, dan efektif. Melalui pendekatan ini, pengembangan soal dilakukan secara sistematis dengan mengacu pada kebutuhan kurikulum serta karakteristik siswa. Analisis kuantitatif deskriptif juga memungkinkan pengukuran kualitas soal berdasarkan data yang objektif, seperti daya pembeda dan tingkat kesukaran soal. Dengan validitas yang terjaga, soal-soal ini akan lebih tepat dalam menggambarkan kemampuan siswa. Reliabilitas yang tinggi memastikan hasil evaluasi konsisten, sementara daya pembeda yang baik membantu membedakan siswa dengan kemampuan berbeda. Oleh karena itu, soal esai yang dihasilkan tidak hanya meningkatkan efektivitas pembelajaran, tetapi juga memberikan gambaran yang lebih jelas mengenai pencapaian siswa.

HASIL DAN PEMBAHASAN

Prosedur Pengembangan Soal

Tahapan penelitian pengembangan menurut (Hermawan, 2019) difokuskan pada dua tahap yaitu tahap *preliminary* dan tahap *formative evaluation.* Tahap *preliminary* yaitu tahapmenentukan lokasi dan subjek penelitian, menghubungi kepala sekolah dan guru, mengatur jadwal penelitian dan prosedur kerja sama. Tahap *formative evaluation* terdiri dari

tiga langkah: *self evaluation, prototyping, dan field test. Self evaluation* melibatkan analisis siswa, kurikulum, dan perangkat, kemudian desain perangkat dibuat dan divalidasi melalui triangulasi data. *Prototyping* mencakup evaluasi oleh pakar *(Expert Review),* uji coba pada siswa/guru *(One-to-One),* dan uji coba pada kelompok kecil *(Small Group).* Hasil evaluasi dan uji coba digunakan untuk merevisi desain. Tahap akhir adalah *Field Test,* melibatkan 15 orang siswa Sekolah Dasar, diikuti dengan revisi final dan pengembangan produk akhir.



Gambar 2. Alur Prosedur Pengembangan Soal

Instrumen terdiri dari 15 butir soal yang dirancang untuk mengukur pemahaman konsep IPS siswa Kelas V SD pada materi Indonesiaku Kaya Raya. Analisis data dilakukan dengan langkah-langkah sebagai berikut : (1)Validitas: Menggunakan nilai korelasi antara skor butir dengan skor total. Soal dinyatakan valid jika nilai korelasi r_hitung > r_tabel (0,4409). (2) Reliabilitas: Dihitung menggunakan rumus *Alpha Cronbach* untuk mengukur konsistensi internal. (3) Tingkat Kesukaran: Ditentukan dengan membagi rata-rata skor siswa dengan skor maksimal. Kriteria: Mudah (TK ≥ 0,70), Sedang (0,30 ≤ TK < 0,70), dan Sukar (TK < 0,30). (4) Daya Pembeda: Menggunakan selisih rata-rata skor kelompok atas dan bawah dibagi skor maksimal. Kriteria: Baik (DP ≥ 0,40), Cukup (0,20 ≤ DP < 0,40), dan Jelek (DP < 0,20).

a. Jenis Soal Dan Penskoran

Soal yang dikembangkan menggunakan jenis soal esai, yang dirancang untuk mengukur kemampuan siswa dalam memahami, menganalisis, dan mengungkapkan gagasan secara terperinci. Soal esai dipilih karena jenis ini memberikan kesempatan kepada siswa untuk menjelaskan jawaban mereka secara mendalam, menunjukkan pemahaman konseptual, dan mengaitkan pengetahuan yang dimiliki dengan konteks yang relevan. Penskoran soal dilakukan dengan menggunakan pedoman rubrik penilaian yang mencakup aspek-aspek penting, seperti kelengkapan jawaban, keakuratan informasi, penggunaan argumen yang logis, dan relevansi dengan pertanyaan yang diajukan. Setiap jawaban siswa akan dinilai berdasarkan kriteria yang telah ditetapkan untuk memastikan penilaian dilakukan secara objektif dan konsisten.

Tabel 1. Rubrik Penskoran untuk Esai

Skor	Kriteria Penilaian
4	Menunjukkan pemahaman yang benar dengan mengembangkannya menggunakan kata kunci yang diinginkan.
3	Menyebutkan sebagian kata kunci, mengarah tapi tidak berkembang.
2	Ada upaya menyebutkan sebagian kata kunci dan mengarah pada jawaban yang benar.
1	Ada upaya menyebutkan kata kunci tetapi tidak benar.
0	Tidak ada jawaban benar/ mengarah kepada kebenaran.

b. Validitas

ISSN:2549-9300 (Online) | ISSN: 1858-4497 (Print)

Validitas adalah sejauh mana suatu instrumen pengukuran (misalnya, tes, kuesioner, atau survei) dapat mengukur apa yang seharusnya diukur(Haladyna, 2011)Sugiyono, 2022). Tujuan validitas adalah untuk memastikan kevalidan instrumen (Priyadi & Suryanti, 2017). Dengan kata lain, validitas menunjukkan keakuratan atau ketepatan dari alat ukur dalam merepresentasikan konsep atau konstruk yang dimaksud. Instrumen yang valid memastikan hasil yang dihasilkan akurat dan relevan dengan tujuan penelitian. Validitas tinggi meningkatkan kredibilitas dan keandalan hasil penelitian. Validitas menggunakan SPSS 22. Validitas menggunakan nilai korelasi antara skor butir dengan skor total. Soal dinyatakan valid jika nilai korelasi r_hitung > r_tabel (0,4409). Selanjutnya, hasil skor dianalisis dengan menggunakan teknik uji korelasi *product moment*. Hasil uji validitas disajikan pada Tabel 2.

Tabel 2. Hasil Uji Validitas Item Soal Pemahaman Konsep

No	Materi Pokok	Nomor Soal	
		Valid	Tidak Valid
1	Bagaimana bentuk Indonesiaku	4	3,10
2	Indonesiaku Kaya Hayatinya	1,2,5,11,12,13	
3	Indonesiaku kaya alamnya	6,7,8,9,14,	15

Berdasarkan tabel di atas, diketahui bahwa dari total 15 soal yang dianalisis, terdapat 3 butir soal yang dinyatakan tidak valid yaitu soal nomor 3,10, dan 15. Hal ini menunjukkan bahwa 12 soal lainnya memenuhi kriteria validitas yang ditentukan dan dapat digunakan dalam proses pembelajaran. Soal-soal yang tidak valid perlu direvisi, diperbaiki atau dibuang agar dapat memenuhi standar kelayakan. Secara keseluruhan, sebagian besar soal sudah memenuhi syarat validitas. Selanjutnya uji Reliabilitas menggunakan SPSS dengan jumlah 12 soal yang valid.

c. Reabilitas

Menurut (Stanley, 1967)reliabilitas sebagai tingkat konsistensi hasil pengukuran suatu instrumen ketika digunakan berulang kali dalam kondisi yang sama.Tinggi rendahnya reliabilitas, secara empirik ditunjukan oleh suatu angka yang disebut nilai koefisien reliabilitas. Reliabilitas yang tinggi ditunjukan dengan nilai rxx mendekati angka 1(Kuder, 1937). Kesepakatan secara umum reliabilitas yang dianggap sudah cukup memuaskan jika ≥ 0.700. Pengujian reliabilitas instrumen dengan menggunakan rumus *Alpha Cronbach*. Rumus *Alpha Cronbach*(Cronbach, 1943) sebagai berikut :

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_t^2}{\sigma_t^2}\right)$$

Keterangan :

r 11 = reliabilitas yang dicari

n = Jumlah item pertanyaan yang di uji

∑σ t2 = Jumlah varians skor tiap-tiap item

σ t2 = vrians total

Jika nilai alpha > 0.7 artinya reliabilitas mencukupi (*sufficient reliability*) sementara jika alpha > 0.80 ini mensugestikan seluruh item reliabel dan seluruh tes secara konsisten memiliki reliabilitas yang kuat.Jika alpha > 0.90 maka reliabilitas sempurna. Jika alpha antara 0.70 – 0.90 maka reliabilitas tinggi. Jika alpha 0.50 – 0.70 maka reliabilitas moderat. Jika alpha < 0.50 maka reliabilitas rendah. Jika alpha rendah, kemungkinan satu atau beberapa item tidak reliabel.

Tabel 3. Hasil Uji Reliabilitas Instrumen Tes Pemahaman Konsep

Reliability Statistics

Cronbach's Alpha	N of Items
,871	12

Hasil uji reliabilitas pada tabel menunjukkan bahwa nilai*Cronbach's Alpha* adalah 0.871 dengan jumlah item **(**N of Items**)** sebanyak 12. Jika alpha > 0.80 memiliki reliabel yang kuat. Instrumen dapat dikatakan reliabel dan dapat digunakan untuk penelitian karena menghasilkan hasil yang konsisten. Instrumen pengukuran memiliki tingkat reliabilitas kuat (α = 0.871) dan layak digunakan dalam penelitian.

d. Daya Pembeda

Daya pembeda soal adalah kemampuan sebuah soal untuk membedakan antara siswa yang sudah menguasai materi dengan siswa yang belum (Tobin, 2024). Semakin tinggi daya pembeda suatu soal, semakin baik soal tersebut dalam mengukur pemahaman siswa (Wright & Panchapakesan, 1969 ; Haladyna & Rodriguez, 2013). Cara Menghitung daya pembeda Soal kita membandingkan proporsi siswa kelompok atas (yang nilainya tinggi) yang menjawab benar dengan proporsi siswa kelompok bawah (yang nilainya rendah) yang menjawab benar. Indeks daya pembeda setiap butir dinyatakan dalam bentuk proporsi yang besarnya berkisar antara –1,00 sampai dengan +1,00. Penentuan daya beda penulisan ini menggunakan koefisien point biserial (rpbi), dimana semakin besar nilai rpbi maka butir tersebut semakin mampu membedakan peserta tes berkemampuan tinggi dengan yang berkemampuan rendah.

Rumus untuk menghitung daya beda soal esai adalah sebagai berikut:

$$\text{DP} = \frac{\sum A - \sum B}{SMI}$$

Keterangan:

DP: Daya Pembeda

$\sum A$: Rata-rata skor kelompok atas

$\sum B$: Rata-rata skor kelompok bawah

SMI: Skor maksimum ideal

Interpretasi Nilai Daya Beda:

DP ≥ 0,40: Soal memiliki daya beda baik.

0,30 ≤ DP < 0,40: Soal memiliki daya beda cukup.

0,20 ≤ DP< 0,30: Soal memiliki daya beda rendah.

DP < 0,20: Soal tidak baik untuk membedakan kemampuan siswa.

Catatan: cara menentukan banyaknya jumlah kelompok atas dan bawah analisis butir soal : (1) Untuk menentukan jumlah kelompok atas dan kelompok bawah, pertama-tama

ISSN:2549-9300 (Online) | ISSN: 1858-4497 (Print)

lihat dulu keseluruhan jumlah sampel (siswa) yang ikut serta. (2) Apabila sampel berjumlah lebih dari 30 orang, maka harus mengambil masing-masing siswa sebanyak 27% - 33% dari setiap kelompok atas maupun bawah. (3) Sedangkan jika sampel berjumlah setidaknya hanya 30 orang atau tidak lebih, maka harus mengambil masing-masing sebanyak 50% dari setiap kelompok atas maupun bawah. Hasil uji daya pembeda disajikan pada Tabel 4.

Tabel 4. Hasil Uji Daya Pembeda Instrumen Tes Pemahaman Konsep

	S	S	S	S	S	S	S	S	S	S	S	S
	1	2	4	5	6	7	8	9	11	12	13	14
Rata2 atas	1,86	3,14	3,29	2,86	3,57	3,29	3,43	2,43	2,14	2,29	4,00	3,86
Rata2 bawah	1,29	1,71	0,86	1,57	2,71	1,86	1,43	1,29	1,14	1,14	2,57	2,00
DP	0,14	0,36	0,61	0,32	0,21	0,36	0,5	0,29	0,25	0,29	0,36	0,46
Kriteria	Jelek	Cukup	Baik	Cukup	Cukup	Cukup	Baik	Cukup	Cukup	Cukup	Cukup	Baik

Berdasarkan tabel diatas dapat diketahui dari 12 butir soal yang valid, 3 butir soal berkategori Baik (DP ≥ 0,40), 8 butir soal Cukup (0,20 ≤ DP < 0,40), dan 1 butir soal Jelek (DP < 0,20). Soal dengan daya beda jelek perlu direvisi untuk meningkatkan efektivitasnya dalam membedakan kemampuan siswa.

e. Tingkat Kesukaran Soal

Tingkat kesukaran butiran soal adalah proporsi antara banyaknya peserta tes yang menjawab butir soal dengan benar dengan banyanya peserta. Hal ini berarti makin banyak peserta tes yang menjawab butir soal dengan benar maka makin besar indeks tingkat kesukaran yang berarti makin mudah butir soal itu. Sebaliknya makin sedikit peserta tes menjawab butir soal dengan benar maka soal tersebut makin sukar.

Rumus : TK = $\frac{Mean}{Skor\ Maksimum}$

Dimana

TK = Tingkat kesukaran soal uraian

Mean = Rata-rata Skor Siswa

Skor Maksimum = Skor maksimum yang ada pada pedoman penskoran

Interprestasi tingkat kesukaran soal

TK <0,30 = Sukar

0,30≤ TK ≤0,70 = Sedang

TK>0.70 = Mudah

Tabel 5. Hasil Uji Tingkat Kesukaran Instrumen Tes Pemahaman Konsep

	S	S	S	S	S	S	S	S	S	S	S	S
	1	2	4	5	6	7	8	9	11	12	13	14
Rata skor	1,53	2,33	2,00	2,10	3,20	2,60	2,50	1,80	1,60	1,67	3,33	2,87
SkorMaks	4	4	4	4	4	4	4	4	4	4	4	4
TK	0,38	0,58	0,50	0,50	0,80	0,65	0,60	0,45	0,4	0,42	0,83	0,72
Kriteria	Sedang	Sedang	Sedang	Sedang	Mudah	Sedang	Sedang	Sedang	Sedang	Sedang	Mudah	Mudah

Dapat dilihat pada Tabel 5, mayoritas soal memiliki tingkat kesukaran dengan kriteria sedang, yang mencakup soal nomor 1, 2, 3, 4, 6, 7, 8, 9, dan 10. Sementara itu, terdapat tiga soal yang memiliki kriteria mudah, yaitu soal nomor 5, 11, dan 12. Tidak ada soal yang termasuk dalam kriteria sukar, sehingga keseluruhan tingkat kesukaran soal berada pada kategori sedang hingga mudah. Hasil penelitian ini menunjukkan bahwa sebagian besar soal sudah memenuhi kriteria kualitas yang baik. Namun, beberapa soal perlu diperbaiki untuk meningkatkan daya beda dan menyeimbangkan tingkat kesukaran. Sebagai contoh, soal dengan daya beda jelek (nomor 1) perlu direvisi dengan mengubah redaksi atau tingkat kedalaman materi yang diuji.

KESIMPULAN

Penelitian ini menyimpulkan bahwa soal evaluasi yang dianalisis memiliki kualitas yang cukup baik secara keseluruhan. Dari 15 soal Sebanyak 12 soal layak digunakan, sementara 3 soal tidak valid dan harus dibuang. Reliabilitas soal sangat baik, tingkat kesukaran sebagian besar berada pada kategori sedang, dan daya pembeda didominasi oleh kategori cukup. Untuk meningkatkan kualitas, revisi diperlukan pada soal dengan daya beda rendah dan tingkat kesukaran yang terlalu mudah.

DAFTAR PUSTAKA

Arikunto, S. (2018). *Dasar-Dasar Evaluasi Pendidikan* (3rd ed.). Jakarta: Bumi Aksara.

Brookhart, S. M., & Mcmillan, J. H. (2020). *Classroom Assesment and Educational Measurement*. London and New York: Routlegde Taylor & Francis.

Cronbach, L. E. E. J. (1943). On Estimates of Test Reliability. *The Journal of Educational Psychology*, *34*, 485–494.

Haladyna, T. M. (2011). *Developing and Validating Multliple-Choice Test Items* (3rd ed.). London and New York: Routlegde Taylor & Francis.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test*. London and New York: Routlegde Taylor & Francis.

Hariyanto, B., Mz, I., & Su, W. (2022). 4D Model Learning Device Development Method of the Physical Geography Field Work Guidance Book. *MATEC Web of Conferences*, *05008*, 0–3.

Hermawan, I. (2019). *Metodologi Penelitian Pendidikan Kuantitatif, Kualitatif, Mixed Methode*. Jakarta: Hidayatul Quran Kuningan.

Idrus. (2019). Evaluasi Dalam Proses Pembelajaran. *Adaara: Jurnal Manajemen Pendidikan Islam*, *9*(2), 920–935.

Kuder, G. . & R. M. . (1937). T h e T h e o r y of t h e E s t i m a t i o n of T e s t R e l i a b i l i t y. *Psychometrika*, *2*(3), 151–160.

Priyadi, Rian.Suryanti, K. (2017). Pengembangan Instrumen Tes Pemahaman Konsep Hukum Gravitasi Universal. *View Metadata, Citation and Similar Papers at Core.Ac.Uk Brought to You by CORE Provided by Portal Jurnal Elektronik Universitas Negeri Malang JRPF (Jurnal Riset Pendidikan Fisika)*, *2*(2), 36–41.

Stanley, J. C. (1967). General and Special Formulas for Reliability of Differences. *Journal*

ISSN:2549-9300 (Online) | ISSN: 1858-4497 (Print)

of Educational Measurement, *4*(4), 249–252.

Sudaryono. (2021). *Metodologi Penelitian Kuantitatif,Kualitatif, dan Mix Method* (2nd ed.). Rajawali Pers.

Sugiyono. (2022). *Metode Penelitian Kuantitatif*. Bandung: Alfabeta.

Sumaryanta. (2021). *Model Pengembangan Tes*. Cirebon: Confident

Thiagarajan, Sivasailam. Semmel, Dorothy S. Semmel, M. I. (1974). Instructional development for training teachers of exceptional children: A sourcebook. *Bloomington,Indiana: University*. https://doi.org/10.1016/0022-4405(76)90066-2

Tobin, M. A. (2024). Guide to Item Analysis. *Schreyer Institute For Teaching Excellence*, 1–6.

Wess, R., Klock, H., Siller, H.-S., & Greefrath, G. (2021). *International Perspectives on the Teaching and Learning of Mathematical Modelling*. Springer Nature Switzerland. http://www.springer.com/series/10093

Wright, B., & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316446902900102