JKIP : Jurnal Kajian Ilmu Pendidikan
Volume 6 No 2 Halaman 723 - 735

Development of a Historical Thinking Skills Assessment Instrument Based on the Evalbee Application for Colonialism and Imperialism Material in High School Students

Pengembangan Instrumen Penilaian Keterampilan Berpikir Sejarah Berbasis Aplikasi Evalbee Untuk Materi Kolonialisme Dan Imperialisme Pada Siswa SMA

Anik Selviana Agustin[1], Wasino[2], Agus Yuwono[3]
Universitas Negeri Semarang
Email: anikaugust@gmail.com, wasino@mail.unnes.ac.id, agusyuwono@mail.unnes.ac.id

*Corresponding Author

Received : 18 June 2025, Revised : 25 July 2025, Accepted : 26 July 2025

ABSTRACT

*Historical thinking skills are a crucial element within the Merdeka Curriculum. These skills can be defined as the scientific procedures or steps involved in learning history. This research aimed to develop an assessment of historical thinking skills for high school students using the Evalbee application. The study employed a Research and Development (R&D) methodology, specifically the ADDIE model. The developed instrument was deemed feasible based on expert validation and empirical trial evidence. The feasibility achievement included an expert validity score of **0.91**, falling within the "very good" category. Expert validity was calculated using Aiken's V formula. Item analysis for both large-scale and small-scale trials utilized the Rasch model, and all items were declared valid. The Cronbach's Alpha reliability coefficient of **0.81** indicates a "good" range.*

Keywords: *Instrumen, Asessmen, Historical Thinking Skill, Evalbee.*

ABSTRAK

Kemampuan berpikir sejarah adalah elemen krusial dalam Kurikulum Merdeka, yang didefinisikan sebagai prosedur atau langkah ilmiah dalam mempelajari sejarah. Penelitian ini bertujuan untuk mengembangkan asesmen kemampuan berpikir sejarah bagi siswa SMA menggunakan aplikasi Evalbee. Penelitian ini menggunakan metodologi Riset dan Pengembangan (R&D), khususnya model ADDIE. Instrumen yang dikembangkan dinilai layak berdasarkan validasi ahli dan bukti uji coba empiris. Pencapaian kelayakan termasuk skor validitas ahli sebesar 0,91, yang termasuk dalam kategori "sangat baik". Validitas ahli dihitung menggunakan rumus Aiken's V. Analisis butir soal untuk uji coba skala besar dan skala kecil menggunakan model Rasch, dan semua butir soal dinyatakan valid. Koefisien reliabilitas Cronbach's Alpha sebesar 0,81 menunjukkan kategori "baik"

Kata Kunci: Pengembangan Instrumen, Penilaian, Ketrampilan Berfikir Sejarah, Evalbee

1. Pendahuluan

Penilaian atau evaluasi merupakan bagian tak terpisahkan dalam komponen penyelenggaraan Pendidikan. Evaluasi hasil belajar merupakan proses penilaian yang komprehensif terhadap seluruh aspek dalam dunia pendidikan, yang bertujuan untuk mengukur keberhasilan dalam mencapai tujuan pendidikan yang telah ditetapkan. (Qodir, 2017, p. 1) dengan demikian Penilaian dilakukan untuk memantau pendidikan dari waktu ke waktu.

Penilaian hasil belajar sejarah untuk melihat standar kecakapan yang harus dimiliki peserta didik menurut Kementrian Pendidikan Budaya Riset dan Teknologi (2022) terdiri dari 5 aspek yakni Keterampilan konsep Sejarah atau *Historical Conceptual Skill*,

Keterampilan berfikir sejarah atau *Historical Thinking skill*, Kesadaran Sejarah atau *Historical Consciousness*, Penelitian Sejarah atau *Historical Research* dan Keterampilan Praktis Sejarah *Historical Practice Skills* dalam melakukan penilaian di sekolah, guru hendaknya sesuai dengan standar kecakapan tersebut.

Dalam kurikulum merdeka keterampilan berfikir sejarah masuk kedalam Elemen keterampilan proses sejarah fase F. Penilaian Pembelajaran sejarah dalam kurikulum merdeka Menurut kementrian Pendidikan budaya riset dan teknologi, (2022) Secara progresif harus mampu mengkontekstualisasikan berbagai peristiwa yang terjadi di masa lalu dengan berbagai peristiwa yang dialami sekarang, untuk dapat saling merenungi, mengevaluasi, membandingkan, atau mengambil keputusan. keterampilan berfikir sejarah menjadi elemen yang sangat penting dalam kurikulum merdeka. Keterampilan berpikir sejarah dapat diartikan prosedur atau langkah – langkah ilmiah dalam belajar sejarah. Berfikir sejarah melibatkan kemampuan kognitif yang memungkinkan peserta didik untuk menganalisis peristiwa masa lalu, menginterpretasi sumber sejarah, dan menghubungkan masa lalu dengan masa kini. Kemampuan ini tidak hanya penting untuk memahami sejarah sebagai sebuah disiplin ilmu, tetapi juga memiliki implikasi yang luas bagi kehidupan sehari-hari.

Namun dalam pelaksanaannya dilapangan, penilaian pembelajaran sejarah seringkali menghadapi tantangan, seperti abstraksi konsep dan kurangnya sumber daya yang memadai,serta belum tersediannya perangkat penilaian yang valid dan reliabel untuk mengukur keterampilan berfikir sejarah, seperti halnya yang terjadi di Kabupaten Banyuasin dalam penyusunan soal para guru sejarah sebagian besar tidak didasarkan pada kisi – kisi melainkan hanya menyadur dari buku paket atau buku kumpulan soal yang diperoleh dari toko-toko buku, alhasil bentuk soal yang diujikan hanya sebatas hafalan belum sampai pada konsep pemahaman apalagi pemaknaaanya (Ofianto,2023). Dalam penelitiannya Ofianto (2023) juga mengungkapkan masih rendahnya keterampilan berfikir sejarah peserta didik SMA di Kabupaten Banyuasin Kondisi tersebut tentu saja bertentangan dengan konsep belajar sejarah dimana peserta didik tidak hanya sebatas menghafal fakta, tetapi juga dituntut untuk berpikir kritis yang memungkinkan peserta didik untuk menganalisis peristiwa sejarah secara mendalam, menghubungkan berbagai informasi, dan menarik kesimpulan yang bermakna. (Hudaidah, 2014, p. 7)

. Berdasarkan uraian tersebut peneliti terdorong untuk mengembangkan instrument peniliaan keterampilan berfikir dalam karya ilmiah tesis yang berjudul "Pengembangan Intrumen Penilaian Keterampilan Berfikir Sejarah Berbasis Aplikasi Evalbee Untuk Materi Kolonialisme dan Imperialisme Pada Siswa SMA".

2. Metodologi

Penelitian ini secara khusus difokuskan pada pengembangan instrument penilaian keterampilan berfikir sejarah yang terdiri dari lima aspek yakni (1). *Chronological thinking*, (2). *Historical comprehension*, (3). *Historical analysis and interpretation*,(4). *Historical research capabilities* (5). *Historical issues-analysis and decision-making.* Instrument berupa soal tes pilihan ganda dan uraian yang dilengkapi dengan kisi – kisi dan pedoman penskoran.

Pengembangan instrument penilaian keterampilan berfikir sejarah dalam penelitian ini menggunakan model ADDIE (*analysis, design, development*, *implementation, evaluation*) yang merupakan salah satu model pengembangan yang umum digunakan dalam desain instruksional.

3. Hasil dan Pembahasan

Pengembangan Instrumen keterampilan berfikir sejarah untuk materi kolonialisme dan imperialisme pada peserta didik tingkat SMA dilatar belakangi kendala – kendala yang

terjadi dilapangan seperti abstraksi konsep, kurangnya pemahaman guru akan aspek – aspek keterampilan berfikir sejarah dan belum adanya instrumen penilaian keterampilan berfikir sejarah yang valid dan reliabel. Keterampilan berfikir sejarah itu sendiri merupakan komponen yang penting dalam pembelajaran sejarah sesuai dengan standar kurikulum merdeka Menurut kementrian Pendidikan budaya riset dan teknologi, (2022) Secara progresif pembelajaran sejarah harus mampu mengkontekstualisasikan berbagai peristiwa yang terjadi di masa lalu dengan berbagai peristiwa yang dialami sekarang, untuk dapat saling merenungi, mengevaluasi, membandingkan, atau mengambil keputusan, sekaligus sebagai orientasi untuk kehidupan masa depan yang lebih baik. Muara dari pembelajaran sejarah yang berorientasi pada keterampilan berpikir secara alamiah akan mendorong pembentukan manusia merdeka yang memiliki kesadaran sejarah dan selaras dengan profil pelajar Pancasila.

Berdasarkan kurikulum merdeka elemen dalam pemebelajaran sejarah terdiri dari dua hal yakni keterampilan kosnep sejarah dan keterampilan proses sejarah. Dalam keterampilan konsep sejarah peserta didik tidak hanya sekedar tahu dan hafal tentang definisi konsep, tetapi juga harus tahu bagaimana menggunakan konsep sebagai bahan analisis untuk mengkaji sebuah peristiwa. Pemahaman konsep dibutuhkan untuk memperoleh penjelasan secara lebih luas dan bermakna tentang sebuah peristiwa, sedangkan dalam elemen keterampilan proses terutama untuk fase F terdapat beberapa aspek yakni keterampilan berfikir sejarah, kesadaran sejarah, penelitian sejarah dan ketrerampilan praktis sejarah yang harus dikuasai peserta didik fase F. Dalam kurikulum merdeka keterampilan berfikir sejarah menjadi salah satu hal yang ditekankan karena masuk dalam elemen keterampilan proses, keterampilan berfikir sejarah terdiri dari berpikir diakronis (kronologi); berpikir sinkronis; berpikir kausalitas; berpikir interpretasi; berpikir kritis; berpikir kreatif; berpikir kontekstual; berpikir imajinatif; berpikir multiperspektif; berpikir reflektif, kemendikbud ristek (2022).

Dalam observasi dan wawancara awal pada dua SMA Negeri dikota semarang menggambarkan bahwa keterampilan berfikir sejarah belum optimal diterapkan saat penilaian pembelajaran sejarah., padahal keterampilan berfikir sejarah dapat membantu guru menganalisa kesadaran sejarah peserta didik yang mana pada hakikatnya pembelajaran sejarah sendiri tidak hanya *transfer of knowledge* namun juga *transfer of velue* melalui keterampilan berfikir sejarah *velue* yang terdapat dalam pembelajaran sejarah dapat ter *transfer* kepada peserta didik, dengan demikian berdasarkan pengamatan dan wawancara awal pengembangan instrumen keterampilan berfikir sejarah menjadi hal yang urgen untuk segera dilakukan.

Pengembangan Instrumen keterampilan berpikir sejarah berfokus pada butir soal yang dikembangkan yakni soal pilihan ganda dengan lima pilihan jawaban, selain itu aspek – aspek keterampilan berfikir sejarah yang dikembangkanpun merujuk kepada 5 aspek yang meliputi (1). *Chronological thinking*, (2). *Historical comprehension*, (3). *Historical analysis and interpretation*,(4). *Historical research capabilities* (5). *Historical issues-analysis and decision-making.*(Nurjanah, 2020, pp. 99–100) atau pemahaman waktu, pemahaman peristiwa secara holistic dan analisis dan intepretasi sejarah, pemahaman sumber sejarah yang meyakinkan serta analisis terhadap isu – isu sejarah. aspek – aspek yang dipilih merupakan aspek yang lebih lengkap dibanding dengan penelitian – penelitian sebelumnya, dan aspek – aspek teresbut tentu saja melebihi standar keterampilan berfikir sejarah dalam kurikulum merdeka yang telah ditetapkan oleh kemndikbud.

Butir soal yang dikembangkan adalah pilihan ganda yang mana hal ini sangat berbeda dengan penelitian sebelumnya yang dilakukan oleh (Ofianto,2023) yang menggunakan soal -soal uraian, pemilihan bentuk soal pilihan ganda tentu saja mengacu kepada kepraktisan mengingat dilapangan tugas guru tidak hanya melakukan pembelajaran dikelas namun juga tugas tambahan lainnya serta beban adminstrasi, yang sedikit banyak menyita waktu di sekolah.

Selain pada butir soal pengembanagn juga ditekankan pada pemanfaatan apliaksi evalbee, yakni apliaksi scanner jawaban yang bekerja seperti pengorksi lembar jawab computer namun dapat diakses melalui handphone, hal ini tentu saja akan mempermudah guru dalam mengkoreksi jawaban. Nilai tambah yang lain adalah guru dapat membuat soal dalam bentu peper base yang praktis, peper base memiliki kelebihan dalam meminimalisisr kecurangan yang mungkin dilakukan oleh peserta didik.

Instrumen penilaian keterampilan berfikir sejarah yang telah dikembangkan sebelum dujikan kepada peserta didik diuji kelayakannya terlebih dahulu. Kelayakan instrumen adalah syarat mutlak agar sebuah alat ukur dapat menjalankan fungsinya secara akurat, andal, dan efektif. Instrumen tes dikatakan layak jika memenuhi kriteria validitas, yang mana dapat diartikan sebagai tingkat kepercayaan terhadap instrumen tes yang akan digunakan dalam mengukur apa yang ingin diukur (Ramadhani & Fitri, 2020). Proses pengembangan instrumen yang umum dan diakui validitasnya melibatkan tinjauan literatur, masukan ahli, dan uji coba lapangan (Akbar et al., 2023). Hal tersebut krusial karena menurut Gavora dan Kratochvilova (2021), pengembangan instrumen yang valid memerlukan proses validasi yang cermat untuk memastikan akurasi pengukuran. Sebagaimana diuraikan dalam metodologi penelitian, kelayakan ini ditinjau dari dua perspektif utama yakni penilaian para ahli dan uji coba empiris yang meliputi (validitas, reliabilitas, tingkat kesukaran dan daya beda soal).

Tahap awal dalam menentukan kelayakan instrumen adalah melalui proses validasi oleh para ahli di bidangnya, yang dalam penelitian ini terdiri dari 5 ahli dengan rincian 3 ahli sejarah, 2 ahli instrumen dan Bahasa. Validasi ahli memberikan justifikasi awal mengenai kesesuaian instrumen dengan tujuan pengukurannya, sejalan dengan prinsip validitas yang menyatakan bahwa sebuah tes harus mengukur apa yang seharusnya diukur. Berdasarkan Hasil analisis kuantitatif menggunakan formula Aiken's V menunjukkan bahwa instrumen yang dikembangkan secara umum dinyatakan layak untuk digunakan. Rata-rata keseluruhan nilai validitas dari ahli adalah 0.91 melebihi V tabel 0.87 yang menjadi standar minimal penerimaan.hasil validitas dapat dilihat pada tabel berikut :

Tabel Rekapitulasi Analisis Validasi Ahli.

Aspek	**Ahli**					**V**	**Ket**
	1	**2**	**3**	**4**	**5**		
Materi 1	4	3	3	4	4	0.87	Tinggi
Materi 2	4	4	3	4	4	0.93	Tinggi
Materi 3	4	3	4	4	3	0.87	Tinggi
Materi 4	4	4	4	3	4	0.93	Tinggi
Materi 5	4	4	4	4	4	1.00	Tinggi
Konstruksi 1	3	4	3	4	4	0.87	Tinggi
Konstruksi 2	4	4	4	3	4	0.93	Tinggi
Konstruksi 3	4	4	3	3	4	0.87	Tinggi
Konstruksi 4	4	4	3	4	4	0.93	Tinggi
Bahasa 1	4	3	3	4	4	0.87	Tinggi
Bahasa 2	3	4	4	4	3	0.87	Tinggi
Bahasa 3	4	4	4	4	4	1.00	Tinggi

Temuan ini memberikan justifikasi awal yang kuat bahwa instrumen ini, dari segi desain dan konten, telah memenuhi standar kelayakan minimum sejalan Jika nilai validasi yang diperoleh sama dengan pernyataan Rohman et al., (2024) nilai minimal yang dibutuhkan dalam tabel Aiken (Nilai koefisisen Aiken). Produk tersebut dianggap valid.

Uji kelayakan selanjutnya adalah uji empiris yang terdiri dari uji skala kecil 31 peserta didik dan uji skala besar 111 peserta didik. Analisis data empiris ini memanfaatkan pemodelan Rasch untuk memberikan bukti kuantitatif yang kuat tentang kelayakan setiap butir soal (Zhang et al., 2023). Validitas butir soal dianalisis memakai perangkat lunak Winstep, berpatokan pada tiga kriteria utama: Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), dan Point Measure Correlation (Pt. Mean Corr). Dalam uji skala kecil hasil analisis menunjukan dua puluh butir soal valid dengan ZSTD berada pada rentang (-2 sampai dengan +2) namun demikian butir soal nomor 14 dan 19 memiliki nilai MNSQ 2,58 dan 1.53 diatas ambang batas penerimaan (0.5-1.5) nilai MNSQ yang tinggi mengindikasikan adanya pola tak terduga dan ketidak konsistenan dari peserta didik yang

dimungkian soal tersebut ambigu atau membingungkan. Sementara untuk PT Measure CORR berada dibawah ambang penerimaan 0.4 dan 0.85 terdapat 3 soal yakni item nomor 14 (0.30), nomor 15 (0.36) dan nomor 5 (0.38).

Fenomena tersebut juga terjadi pada hasil uji skala besar yang masih menunjukan kedua puluh soal valid dengan nilai ZSTD berada pada ambang batas kriteria penerimaan (-2 sampai dengan +2) dan hasil MNSQ yang berada diatas batas maksimal yakni (0.5 – 1.5) pada butir soal nomor 11 (2.13) dan 12 (1.57) hal ini menunjukan konsistensi meskipun jumlah peserta didik sebagai sampel semakin banyak dan heterogen. Peningkatan terjadi pada PT – Measure CORR dimana item soal yang berada dibawah ambang batas penerimaan menjadi dua item yakni item nomor 16 (0.38) dan item nomor 9 (0.36) Meskipun demikian, karena kriteria utama ZSTD terpenuhi, validitas konstruk instrumen secara keseluruhan dapat diterima dalam penelitian ini karena Pengembangan instrumen penilaian yang valid untuk kemampuan berpikir sejarah merupakan tantangan signifikan yang memerlukan perhatian terhadap nuansa disipliner **(**Monte-Sano & Wineburg, 2017). Rekapitulasi hasil hasil Validitas butir soal untuk uji skala kecil dan besar tersaji pada tabel berikut:

Tabel 1.2 Analisis Validitas Butir Soal Uji Skala Kecil

Item	**Outfit MNSQ**	**Outfit ZSTD**	**Pt-Mean Corr**	**Keputusan**
Q1	0.84	0.40	0.53	Valid
Q2	0.95	0.00	0.48	Valid
Q3	0.92	-0.30	0.44	Valid

Q4	1.00	0.10	0.45	Valid
Q5	0,85	-0.40	0.38	Valid
Q6	0.72	-1.2	0.50	Valid
Q7	0.84	0.1	0.59	Valid
Q8	1.43	0.7	0.43	Valid
Q9	1.12	0.4	0.47	Valid
Q10	0.93	-0.2	0.40	Valid
Q11	0.92	-0.3	0.42	Valid
Q12	0.89	-0.3	0.46	Valid
Q13	0.66	-0.4	0.58	Valid
Q14	2.58	0.30	0.30	Valid
Q15	1.09	0.5	0.36	Valid
Q16	0.77	0.00	0.55	Valid
Q17	0.82	-0.6	0.44	Valid
Q18	1.08	0.40	0.42	Valid
Q19	1.53	0.9	0.41	Valid
Q20	1.00	0.3	0.52	Valid

Tabel 1.3 Analisis Validitas Butir Soal Uji Skala Besar

Item	Outfit MNSQ	Outfit ZSTD	Pt-Mean Corr	Keputusan
Q1	1.09	0.46	0.51	Valid
Q2	0.99	0.01	0.49	Valid
Q3	1.35	0.95	0.46	Valid
Q4	0.69	0.71	0.42	Valid
Q5	0.87	-0.57	0.47	Valid
Q6	1.03	0.21	0.52	Valid
Q7	1.04	0.24	0.46	Valid
Q8	1.05	0.27	0.41	Valid
Q9	0.59	0.61	0.36	Valid
Q10	1.16	0.93	0.41	Valid
Q11	2.13	1.48	0.42	Valid
Q12	1.57	1.33	0.48	Valid
Q13	0.82	0.57	0.47	Valid
Q14	0.54	-1.40	0.54	Valid
Q15	0.99	0.03	0.50	Valid
Q16	1.17	0.88	0.38	Valid
Q17	1.24	1.32	0.41	Valid
Q18	0.76	0.69	0.43	Valid
Q19	1.21	1.13	0.48	Valid
Q20	1.16	0.87	0.48	Valid

Reliabilitas instrumen menggambarkan konsistensi kepercayaan. hasil pengukuran reliabilitas pada uji skala kecil nilai reliabilitas item yang diperoleh adalah 0.84 dan barada pada kategori “Bagus” (Hasanah & Aini, 2025), nilai reliabilitas mengalami kenaikan pada uji skala besar yakni sebesar 0.97 dengan kategori “istimewa” (Hasanah & Aini, 2025). Sementara untuk nilai Cronbach Alpha dari uji sekala kecil adalah 0.80 dan mengalami kenaikan pada uji sekala besar yakni 0,81 keduanya berada pada kategori “cukup” dan “bagus”. Hal ini menunjukkan semua item/butir dalam instrumen dapat diandalkan untuk mengukur konstruk secara tepat dan konsisten (Astuti et al., 2022). Butir – butir soal yang telah dikembangkan memiliki tingkat konsistensi dan kepercayaan internal yang sangat baik, artinya butir – butir soal didalamya mengukur konstruk yang sama yakni keterampilan berfikir sejarah pada materi kolonialisme dan imperialisme ditambah analisis pemodelan Raschd digunakan meningkatkan akurasi instrumen dan konsistensi hasil pengukuran. Model ini juga memungkinkan kalibrasi item, sehingga menghasilkan interpretasi temuan penelitian yang lebih valid (Sujatmika et al., 2025). Hasil Reliabilitas butir soal tersaji pada tabel berikut :

Tabel 1.4 Reliabilitas Uji Skala Kecil

Reliabilitas	Reliabilitas	Reliabilitas

Person	Item	Alpha Cronbach
0.76	0.84	0.80
Cukup	**Bagus**	**Bagus**

Tabel 1.5 Reliabilitas Uji Skala Besar

Reliabilitas Person	Reliabilitas Item	Reliabilitas Alpha Cronbach
0.78	0.97	0.81
Cukup	**Istimewa**	**Bagus**

Analisa selanjutnya adalah Tingkat kesukaran soal, soal yang baik harusnya memiliki tingkat persebaran kesukaran yang seimbang, Ini penting karena menurut penelitian Utaminingsih et al. (2024), soal yang baik itu tidak boleh terlalu mudah atau terlalu sulit. Soal yang terlampau sulit bisa membuat siswa patah semangat, sementara soal yang lebih mudah diperlukan untuk memacu semanat mereka dalam mengerjakan. Berdasarkan analisis tingkat kesukaran menggunakan nilai logit dari model Rasch, instrumen ini telah memenuhi kriteria tersebut.pada uji skala kecil persebaran soalnya sebagai berikut 10 butir dengan kriteria " mudah", 3 butir dengan kriteria "sedang", 7 butir dengan kriteria "sukar". Dalam uji skala besar persebaran soal menjadi semakin imbang dan variatif , 1 butir dengan kriteria "sangat mudah", 7 butir dengan kriteria "mudah", 5 butir dengan kriteria "sedang", 4 butir dengan kriteria "sulit" dan 3 butir dengan kriteria "sangat sulit". Tingkat kesukaran soal memegang peranan esensial dalam pengembangan dan evaluasi instrumen pengukuran. Hal ini krusial untuk memastikan bahwa alat ukur mampu mengidentifikasi dan membedakan kemampuan individu secara akurat dan valid. Butir soal dengan distribusi tingkat kesukaran yang proporsional—mencakup kategori mudah, sedang, dan sulit akan mempermudah dalam memetakan rentang kemampuan responden secara komprehensif (Lucky et al., 2025). Persebaran Tingkat kesukaran Butir soal dapat dilihat pada Tabel berikut:

Butir Soal	**Nilai Logit**	**Keterangan**

Tabel 1.5 Persebaran Tingkat Kesukaran Butir Soal Pada Uji Skala Kecil.

Butir Soal	Nilai Logit	Keterangan
1	0.13	Sedang
2	-0.27	Sedang
3	1.67	Sukar
4	-0.6	Mudah
5	1.99	Sukar
6	1.64	Sukar
7	-1.44	Mudah
8	-1.44	Mudah
9	-1.07	Mudah
10	1.48	Sukar
11	1.48	Sukar
12	0.50	Sedang
13	-1.07	Mudah
14	-1.44	Mudah
15	1.48	Sukar
16	-1.44	Mudah
17	1.81	Sukar
18	-1.44	Mudah
19	-1.07	Mudah
20	-1.44	Mudah

Tabel 1.5 Persebaran Tingkat Kesukaran Butir Soal Pada Uji Skala Besar.

Butir Soal	Nilai Logit	Keterangan
1	-0.51	Mudah
2	-0.10	Sedang
3	2.41	Sangat Sukar
4	-1.65	Mudah
5	-0.35	Sedang
6	0.74	Sukar
7	-1.57	Mudah
8	-1.18	Mudah
9	-2.59	Sangat Mudah
10	0.64	Sukar
11	3.49	Sangat Sukar
12	2.58	Sangat Sukar
13	-0.94	Mudah
14	-1.43	Mudah
15	-0.05	Sedang
16	-0.30	Sedang
17	0.19	Sedang
18	-1.20	Mudah
19	0.89	Sukar
20	0.94	Sukar

Uji selanutnya adalah daya beda soal. Daya beda soal adalah kemampuan sebuah butir soal untuk memisahkan siswa berkemampuan tinggi dari yang rendah. Ini diukur menggunakan indeks separasi (H) dari analisis Winstep. Pada uji coba skala kecil, nilai H untuk *Separation Person* adalah 3 ("Baik ") artinya instrumen sudah bisa membagi peserta didik menjadi tiga kelompok (rendah, sedang, tinggi) sedangakan untuk nilai H *Separation item* Adalah 4 yang artinya terdapat 4 kelompok butir soal. Peningkatan signifikan terlihat di uji coba skala besar, di mana nilai H Separation item meningkat menjadi 8 yang artinya terdapat delapan kelompok butir soal yang dapat diidentifikasi. Sedangkan nilai H Separation Person menunjukan konsistensi dari uji sekala kecil yakni sebesar 3 yang menunjukan bahwa kelompok responden dapat dibedakan menjadi tiga kelompok (rendah, sedang, tinggi). Peningkatan nilai H Ini menunjukkan butir-butir soal di dalamnya sangat efektif dalam membedakan siswa dengan level kemampuan yang berbeda secara sangat halus. Peningkatan daya beda ini menjadi salah satu indikator keberhasilan kunci karena

daya beda yang "Baik" memungkinkan guru memberikan umpan balik yang lebih tepat dan intervensi pembelajaran yang sesuai dengan kebutuhan unik setiap kelompok peserta didik.

Setelah melewati serangkaian proses validasi dan uji coba menyeluruh, instrumen keterampilan berpikir sejarah ini dinyatakan layak untuk digunakan. Kelayakan ini tidak hanya karena memenuhi berbagai kriteria, tetapi juga karena instrumen ini selaras dengan landasan teoretis yang telah dipaparkan sebelumnya. Instrumen ketrampilan berfikir sejarah dinyatakan layak berdasarkan validasi ahli yang komprehensif dan data uji coba empiris yang kuat. Hasil uji coba skala besar menunjukkan bahwa instrumen ini memiliki validitas konstruk yang kuat, reliabilitas dalam kategori "Bagus" (0,80), sebaran tingkat kesukaran yang ideal, dan daya beda yang baik. Karakteristik ini membuktikan bahwa instrumen yang dikembangkan bukan sekadar kumpulan soal, melainkan alat ukur yang teruji secara ilmiah, mampu mengukur keterampilan berpikir sejarah materi kolonialisme dan imperialisme pada peserta didik kelas XI secara akurat.

Analisis Profil Keterampilan berfikir sejarah peserta didik merupakan tahapan yang penting dalam penelitian ini untuk mengethau tingkat presentase setiap aspek ketrampilan berfikir sejarah, dari data yang diperoleh pada uji sklala kecil menunjukan kriteria “sedang” secara terperinci hasilnya aspek *Historical Comperhension* dengan 74.73% berada pada kategori Tinggi, aspek Historical analysis and interpretation 74.19% berada pada kateogri Tinggi, aspek Chronological thinking 47.31% berada pada kategori Sedang, aspek Historical issues-analysis and decision-making 65.59% berada pada ketegori Sedang dan aspek Historical research capabilities 53.76% berada pada kategori Sedang, rerata untuk keseluruahn aspek adalah 63.12 dan berada pada rentang sedang. Hasil yang berbeda terjadi pada uji skala besar dengan rincian sebagai berikut hasilnya aspek *Historical Comperhension* dengan 71.52% berada pada kategori Tinggi, aspek *Historical analysis and interpretation* 56.91% berada pada kateogri sedang, aspek Chronological thinking 12.42% berada pada kategori Rendah, aspek Historical issues-analysis and decision-making 60.10% berada pada ketegori sedang dan aspek Historical research capabilities 60.91% berada pada kategori sedang, rerata untuk keseluruahn aspek adalah 52.35% dan berada pada rentang sedang

Dari data yang telah diperoleh terdapat perbedaan pada semua aspek keterampilan berfikir sejarah meskipun beberapa masih pada kategori yang sama penurunan ekstrim terjadi pada aspek *chronological thinking* dari 47.31% menjadi 12.42% hal ini menjadikan yang semula berkategori sedang menjadi berkategori rendah. Penuruanan ekstrim lainnya terjadi pada aspek *Historical analysis and interpretation* dengan presentase awal 74.19% menjadi 56.91% sehingga kriterianyapun ikut turun dari tinggi menjadi sedang. Perbedaan hasil ini disebabkan perbedaan jumlah populasi dimana populasi dalam uji skala besar lebih banyak dan bersifat heterogen. Walaupun demikian secara menyeluruh baik dalam uji skala kecil maupun besar rerata keduanya berada pada kategori sedang sehingga instrumen ini konsisten dan dapat disimpulkan keterampilan berfikir sejarah peserta didik berada pada level sedang

KESIMPULAN

Pengembangan Instrumen Ktrampilan Berfikir Sejarah Untuk Materi Kolonialisme dan Imperlaisme Pada Peserta Didik Kelas XI bersifat urgen untuk dilakukan.

Instrumen yang dikembangkan dinyatakan layak untuk digunakan berdasarkan validasi ahli dan bukti uji coba empiris. Capaian kelayakan ini diuraikan sebagai berikut nilai validitas dari ahli /expert Djugsment adalah 0,91 pada renatng kategori sangat baik. Hasil ini melampaui kriteria minimal yaitu 0,87. 20 butir soal dinyatakan valid berdasarkan analisis model Rasch. Instrumen memiliki reliabilitas dengan kategori “Istimewa” dengan nilai Reliablitas Item sebesar 0,97 dan nilai Cronbach Alpha 0.81 dengan kategori “Baik”. Instrumen memiliki sebaran tingkat kesukaran yang ideal, dari sangat mudah hingga sangat

sulit. Instrumen memiliki daya beda denga kategori “Baik” dengan nilai H=3 untuk sparasi person dan nilai H = 8 Untuk Sprasi item yang mampu membedakan siswa ke dalam tingakatan dengan akura. Profil keterampilan berfikir sejarah peserta didik kelas XI SMA pada materi kolonialisme dan imperialisme secara umum berada pada kategori “Sedang”. Hasil ini konsisten pada uji coba skala kecil maupun besar.

Refrence

Abuhassna, H., Yahaya, N., & Zakaria, M. A. Z. M. (2024). Strategies for Successful Blended Learning—A Bibliometric Analysis and Reviews. International Journal of Interactive Mobile Technologies (iJIM), 16(13).

Akbar, W., Jaafar, N. I., & Mohezar, S. (2023). Technological work environment: instrument development and measurement. *Behaviour and Information Technology*, *42*(1), 25–45. https://doi.org/10.1080/0144929X.2021.2013536

Aricindy, A., Wasino , W., Atmaja, H. T., & Wijaya, A. (2023). Mainstreaming Peace Education in Multicultural Schools Yayasan Perguruan Sultan Iskandar Muda Medan and Karang Turi Semarang. *Proceedings of International Conference on Science, Education, and Technology*, *9*(1), 91–94. Retrieved from https://proceeding.unnes.ac.id/ISET/article/view/2399

Astuti, Budi, Universitas Negeri Yogyakarta, Edi Purwanta, Universitas Negeri Yogyakarta, Yulia Ayriza, Universitas Negeri Yogyakarta, Caraka Putra Bhakti, Universitas Ahmad Dahlan, Rizqi Lestari, Universitas Negeri Yogyakarta, Herwin Herwin, and Universitas Negeri Yogyakarta. 2022. “Cypriot Journal of Educational High School Students during the COVID-19 Pandemic.” *Cypriot Journal of Educational Sciences* 17(2):410–21.

Cahyadi, R. A. H. (2019). Pengembangan Bahan Ajar Berbasis Addie Model. *Halaqa: Islamic Education Journal*, *3*(1), 35–42. https://doi.org/10.21070/halaqa.v3i1.2124

Catterall, A. (2020). Supporting delivery of diagnostic assessments for apprentices with undiagnosed additional needs. *Higher Education, Skills and Work-Based Learning*, *10*(4), 623–641. https://doi.org/10.1108/HESWBL-02-2019-0032

Gavora, P., & Kratochvilova, S. (2021). Developing and validating a survey instrument for measuring teachers’ self-efficacy in implementing differentiated instruction. Educational Measurement: Issues and Practice, 40(4), 74-84. https://doi.org/1111/emip.12465

Hamari, J., Shernoff, D. J., Rowe, J., Wang, C., Price, S., & Sihvonen, J. (2016). The impact of gamification on student engagement and academic performance in higher education: A systematic review. Computers & Education, 98, 301-318. https://doi.org/10.1016

Hasanah, H., & Aini, F. Q. (2025). Pengembangan Instrumen Tes Keterampilan Proses Sains pada Topik Kesetimbangan Kimia: Analisis dengan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika Dan IPA*, *5*(1), 68–82. https://doi.org/10.53299/jagomipa.v5i1.1043

Hudaidah. (2014). jurnal keterampilan berfikir sejarah. *Historical Thinking, Keterampilan Berpikir Utama Bagi Mahasiswa Sejarah*, *3*(Historical Thinking, Keterampilan Berpikir Utama Bagi Mahasiswa Sejarah), 6–12.

Ibrahim Maulana Syahid, Nur Annisa Istiqomah, & Azwary, K. (2024). Model Addie Dan Assure Dalam Pengembangan Media Pembelajaran. Journal of International Multidisciplinary Research, 2(5), 258–268. https://doi.org/10.62504/jimr469

Ison, M. J. B., & Oñate, C. C. T. (2021). Effectiveness of automation in evaluating test results using EvalBee as an alternative optical mark recognition (OMR): A quantitative-evaluative approach from a Philippine public school. International

Journal of Theory and Application in Elementary and Secondary School Education, 3(2), 61-75. https://doi.org/10.31098/ijtaese.v3i2.661

Liu, Y., Wang, J., Zhang, Z., Wang, J., Luo, T., Lin, S., Li, J., & Xu, S. (2024). Development and validation of an instrument for measuring civic scientific literacy. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1). https://doi.org/10.1186/s43031-023-00092-3

Monte-Sano, L. M., & Wineburg, S. S. (2017). Measuring historical thinking: The development and validation of the historical thinking assessment. American Educational Research Journal, 54(4), 750-785. https://doi.org/10.3102/0002831217702758

Muslihin, Heri Yusuf, Suryana, and Dodi. 2022. "Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model." *International Journal of Instruction* 15(2):207–22.

Nuha, F. H., Wasino, W., & Widiarti, N. . (2025). The Effectiveness of the Wordwall Media-Assisted Problem Based Learning Model on the Ability to Solve Problems in Science and Technology. *Eduvest - Journal of Universal Studies*, *5*(6), 6745–6755. https://doi.org/10.59188/eduvest.v5i6.50252

Nurjanah, W. (2020). *Historical Thinking Skils dan Critical Thinking Skills Historical Thinking Skils And Critical Thinking Skills* (Vol. 23, Issue 1).

Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia Sma N 1 Kutowinangun Tahun Pelajaran 2019/2020 Menggunakan Model Iteman Dan Rasch. Jurnal Pendidikan Kimia, 10(2), 205–214

Putra, B., Kadek, I., Nuryana, D., Augusta, R., & Firdaus, J. (2019). *Rancang Bangun Aplikasi Koreksi Lembar Jawaban Komputer Menggunakan Metode Deteksi Tepi Canny*.

Qodir, Abdul. 2017. *Evaluasi Dan Penilaian Pembelajaran*. 1st ed. edited by M. Pd. , M. I. Kom. Ngalimun. Yogyakartta: K Media.

Rohman, Fatchur, Agus Dharmawan, and Murni Sapta Sari. 2024. "Development of a 4C Skills Evaluation Instrument for Biology : A Validity and Reliability Study on Indonesian High School Students Learning." *International Journal of Innovative Research and Scientific Studies* 7(31):701–17. doi:10.53894/ijirss.v7i2.2873.

Rusilowati, A., Astuti, B., & Rahman, N. A. (2019). How to improve student's scientific literacy. *Journal of Physics: Conference Series*, *1170*(1). https://doi.org/10.1088/1742-6596/1170/1/012028

Rusilowatil, A., Sundari, & Marwoto, P. (2021). Development of integrated teaching materials vibration, wave and sound with ethnoscience of bundengan for optimization of students' scientific literation. *Journal of Physics: Conference Series*, *1918*(5). https://doi.org/10.1088/1742-6596/1918/5/052057

Saa, S. (2024). Merdeka Curriculum: Adaptation of Indonesian Education Policy in The Digital Era and Global Challenges. *Revista de Gestao Social e Ambiental*, *18*(3). https://doi.org/10.24857/rgsa.v18n3-168

Sadler, T. D., Zangori, L., & Friedrichsen, P. J. (2021). Developing and Using Multiple Models to Promote Scientific Literacy in the Context of Socio-Scientific Issues. *Science and Education*, *30*(3), 589–607. https://doi.org/10.1007/s11191-021-00206-1

Sudjana, N. (2017). *Asesmen: Teori dan Praktik*. Pustaka Pelajar: Yogyakarta.

Sugiyono. (2013). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta: Bandung.

Sujatmika, Sigit, Sutarno, Mohammad Maskuri, and Baskoro Adi Prayitno. 2025. "Applying the Rasch Model to Measure Students ' Critical Thinking Skills on the Science Topic of the Human Circulatory System." *EURASIA Journal of Mathematics, Science and Technology Education* 21(4):1–12.

Waruwu, M. (2024). Metode Penelitian dan Pengembangan (R&D): Konsep, Jenis, Tahapan dan Kelebihan. *Jurnal Ilmiah Profesi Pendidikan*, *9*(2), 1220–1230. https://doi.org/10.29303/jipp.v9i2.2141

Zhang, L., Liu, X., & Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). https://doi.org/10.1186/s43031-023-00093-2