Edu Society: Jurnal Pendidikan, Ilmu Sosial, dan Pengabdian Kepada Masyarakat
Vol 5 No 3 2025, hal 1383-1390
Avaliable online at: https://jurnal.permapendis-sumut.org/index.php/edusociety

Pemberian Skor dan Sistem Penilaian

Lesti Almaida[1], Arpan[2], Athala Rania Insyra[3], Alpina Pionita[4], Ayu Elfera[5]
[1,2,3,4,5] Institut Syekh Abdul Halim Hasan Binjai, Indonesia
Email: lestialmaida04@gmail.com [1], arpancuyy625@gmail.com [2], raniainsyra70@gmail.com [3], alpinapionita661@gmail.com [4], ayuelfera16@gmail.com [5]

Abstrak

Pemberian skor dan sistem penilaian merupakan inti dari asesmen pendidikan dan pelatihan karena menentukan kualitas keputusan akademik: kelulusan, pemeringkatan, umpan balik, hingga perbaikan pembelajaran. Tantangan utama dalam penilaian modern adalah menjaga validitas (apakah skor benar-benar merepresentasikan kompetensi yang dituju), reliabilitas (konsistensi skor), keadilan/fairness (minim bias), serta kegunaan (mendukung pembelajaran). Artikel ini membahas konsep, desain, dan implementasi sistem penilaian dari pendekatan klasik (rubrik analitik/holistik, skala penilaian, pembobotan) hingga pendekatan berbasis pengukuran (generalizability theory, many-facet Rasch) dan tren terkini seperti penilaian otomatis berbasis kecerdasan buatan (AI) untuk esai dan tugas kompleks. Metode yang digunakan adalah studi literatur terarah (2021–2025) dan sintesis konseptual. Hasil pembahasan merumuskan kerangka desain sistem penilaian yang dapat diterapkan di sekolah/perguruan tinggi: (1) penyelarasan capaian pembelajaran–tugas–kriteria, (2) pemilihan model skor dan pembobotan yang transparan, (3) kalibrasi rater dan pengendalian variasi penilai, (4) verifikasi reliabilitas dan bukti validitas, (5) tata kelola data dan audit fairness, serta (6) integrasi umpan balik formatif. Artikel menutup dengan rekomendasi praktis: penggunaan rubrik yang "dapat diaudit", analisis reliabilitas multi-facet untuk tugas ber-rater, dan kehati-hatian pada AI-grading melalui evaluasi bias, keamanan data, serta validasi berkelanjutan.

Kata Kunci: ***AI Grading, Fairness, Pemberian Skor, Penilaian Otomatis, Reliabilitas, Rubrik, Sistem Penilaian, Validitas.***

Score Giving and Assessment System

Abstract

Scoring and grading systems are central to educational assessment because they shape academic decisions, certification, ranking, feedback, and instructional improvement. Contemporary assessment faces recurring challenges: maintaining validity (scores represent the intended construct), reliability (score consistency), fairness (minimizing bias), and utility (supporting learning). This article reviews the principles, design choices, and implementation strategies of scoring systems, ranging from traditional approaches (analytic/holistic rubrics, rating scales, weighting schemes) to measurement-oriented approaches (generalizability theory, many-facet Rasch measurement) and recent developments such as AI-based automated scoring for essays and complex responses. Using a focused literature review (2021–2025) and conceptual synthesis, the paper proposes a practical design framework: (1) alignment of learning outcomes–tasks–criteria, (2) transparent scoring models and weighting, (3) rater calibration and control of rater effects, (4) verification of reliability and validity evidence, (5) data governance and fairness auditing, and (6) integration of formative feedback. The

article concludes with actionable recommendations: deploy "auditable" rubrics, apply multi-facet reliability analyses for rater-mediated tasks, and adopt AI-grading cautiously through bias evaluation, privacy safeguards, and continuous validation.
Keywords: ***AI Grading, Fairness, Scoring, Automated Assessment, Reliability, Rubrics, Assessment Systems, Validity.***

PENDAHULUAN

Penilaian (assessment) merupakan komponen fundamental dalam sistem pendidikan karena berfungsi sebagai dasar pengambilan keputusan akademik, baik pada tingkat individu peserta didik maupun institusi pendidikan secara keseluruhan. Keputusan seperti kelulusan, kenaikan tingkat, pemeringkatan, pemberian sertifikasi, hingga evaluasi mutu pembelajaran sangat bergantung pada hasil penilaian yang diwujudkan dalam bentuk skor dan nilai. Oleh karena itu, pemberian skor dan sistem penilaian tidak dapat dipandang sebagai proses teknis semata, melainkan sebagai aktivitas ilmiah yang menuntut ketepatan konseptual, ketelitian metodologis, serta pertimbangan etis dan keadilan (AERA, APA, & NCME, 2024).

Dalam praktik pendidikan, skor sering kali dipersepsikan sebagai representasi langsung dari kemampuan peserta didik. Padahal, skor hanyalah indikator kuantitatif yang dihasilkan melalui seperangkat keputusan desain penilaian, seperti pemilihan jenis tugas, kriteria penilaian, rubrik, skala penskoran, pembobotan komponen, serta prosedur koreksi. Kesalahan atau ketidaktepatan dalam salah satu komponen tersebut dapat menghasilkan skor yang tidak akurat dan menyesatkan, sehingga berpotensi menimbulkan ketidakadilan akademik (Brookhart, 2023). Oleh sebab itu, sistem penilaian yang baik harus mampu menghasilkan skor yang valid, reliabel, dan fair, serta memiliki kegunaan nyata dalam mendukung proses pembelajaran.

Isu validitas menjadi perhatian utama dalam pemberian skor. Validitas tidak hanya berkaitan dengan apakah suatu tes mengukur apa yang seharusnya diukur, tetapi juga mencakup kesesuaian interpretasi dan penggunaan skor untuk tujuan tertentu (AERA et al., 2024). Dalam konteks sistem penilaian berbasis kinerja dan tugas autentik, validitas sangat dipengaruhi oleh kualitas rubrik dan keselarasan antara capaian pembelajaran, aktivitas belajar, dan kriteria penilaian (Vlachopoulos & Makri, 2024). Tanpa keselarasan tersebut, skor yang dihasilkan berisiko tidak merepresentasikan kompetensi sebenarnya, meskipun prosedur penskoran dilakukan secara konsisten.

Selain validitas, reliabilitas juga merupakan aspek krusial dalam sistem penilaian. Reliabilitas mengacu pada tingkat konsistensi skor apabila penilaian diulang dalam kondisi yang sebanding. Pada penilaian objektif seperti tes pilihan ganda, reliabilitas relatif mudah dikendalikan. Namun, pada penilaian kinerja, esai, dan proyek, reliabilitas sering terpengaruh oleh variasi penilai (rater), variasi tugas, serta interaksi antara penilai dan peserta didik (Shavelson & Webb, 2022). Penelitian terkini menunjukkan bahwa penggunaan rubrik analitik dan pelatihan rater dapat meningkatkan konsistensi skor, meskipun tidak sepenuhnya menghilangkan bias penilai (Khamboonruang, 2023).

Permasalahan fairness atau keadilan dalam penilaian juga semakin mendapat perhatian, terutama dalam konteks pendidikan yang berorientasi pada inklusivitas dan kesetaraan. Sistem penilaian yang tidak adil dapat merugikan kelompok tertentu

berdasarkan latar belakang bahasa, budaya, atau karakteristik individu lainnya. Dalam penilaian berbasis rater, bias dapat muncul dalam bentuk perbedaan tingkat keketatan (severity) atau kelonggaran (leniency) antarpenilai (WIDA, 2024). Oleh karena itu, pengembangan sistem penilaian modern menuntut adanya mekanisme pengendalian bias dan evaluasi keadilan secara berkelanjutan.

Seiring dengan perkembangan teknologi, sistem penilaian juga mengalami transformasi signifikan melalui pemanfaatan teknologi digital dan kecerdasan buatan. Penilaian otomatis (automated scoring), khususnya pada tugas esai dan jawaban terbuka, mulai banyak digunakan untuk meningkatkan efisiensi dan konsistensi penskoran (Xie et al., 2024). Model pembelajaran mesin dan large language models (LLM) menunjukkan kemampuan yang menjanjikan dalam meniru pola penskoran manusia. Namun demikian, sejumlah penelitian menegaskan bahwa sistem penilaian berbasis AI masih menghadapi tantangan serius terkait validitas konstruk, transparansi algoritma, serta potensi bias terhadap kelompok tertentu (Williamson et al., 2024; Loukina et al., 2024).

Di sisi lain, paradigma penilaian dalam pendidikan juga bergeser dari penilaian yang semata-mata bersifat sumatif menuju penilaian formatif yang berorientasi pada pembelajaran. Penilaian formatif menekankan pemberian umpan balik yang bermakna untuk membantu peserta didik memahami kekuatan dan kelemahan mereka, serta memperbaiki strategi belajar (Panadero & Jonsson, 2023). Dalam konteks ini, sistem pemberian skor tidak hanya berfungsi sebagai alat seleksi, tetapi juga sebagai sarana pembelajaran itu sendiri. Rubrik, apabila dirancang dan digunakan secara tepat, dapat berperan sebagai alat scaffolding yang meningkatkan pemahaman peserta didik terhadap kriteria keberhasilan (Panadero et al., 2025).

Meskipun demikian, berbagai penelitian menunjukkan bahwa penggunaan rubrik dan sistem penilaian belum selalu menghasilkan dampak positif yang konsisten. Efektivitas sistem penilaian sangat bergantung pada kualitas desain, konteks penerapan, serta literasi asesmen baik dari pendidik maupun peserta didik (Brookhart, 2023). Hal ini menunjukkan perlunya pendekatan yang komprehensif dan berbasis bukti dalam merancang sistem pemberian skor.

Berdasarkan latar belakang tersebut, kajian mengenai pemberian skor dan sistem penilaian menjadi sangat relevan untuk menjawab tantangan pendidikan kontemporer. Artikel ini bertujuan untuk membahas secara mendalam konsep, prinsip, dan praktik pemberian skor serta sistem penilaian, dengan menyoroti aspek validitas, reliabilitas, fairness, dan pemanfaatan teknologi terkini. Diharapkan, kajian ini dapat memberikan kontribusi teoretis dan praktis bagi pendidik, peneliti, dan pengambil kebijakan dalam mengembangkan sistem penilaian yang akurat, adil, dan berorientasi pada peningkatan kualitas pembelajaran.

METODE

Penelitian ini menggunakan pendekatan kajian pustaka (literature review) dengan desain studi literatur terarah (focused literature review). Metode ini dipilih karena tujuan utama penelitian adalah untuk menganalisis, mensintesis, dan merumuskan pemahaman konseptual mengenai pemberian skor dan sistem penilaian berdasarkan temuan ilmiah terkini, bukan untuk menguji hipotesis melalui pengumpulan data lapangan. Studi literatur

terarah memungkinkan peneliti untuk memusatkan kajian pada tema spesifik, yaitu validitas, reliabilitas, fairness, serta perkembangan teknologi dalam sistem penilaian pendidikan (Snyder, 2019; Vlachopoulos & Makri, 2024).

Penelitian ini termasuk dalam kategori penelitian kualitatif deskriptif berbasis dokumen. Pendekatan kualitatif digunakan untuk memahami konsep, prinsip, dan praktik pemberian skor serta sistem penilaian secara mendalam melalui interpretasi sumber-sumber tertulis yang relevan (Creswell & Poth, 2018). Analisis dilakukan secara naratif dan tematik untuk mengidentifikasi pola, kesamaan, dan perbedaan pandangan antar peneliti mengenai desain dan implementasi sistem penilaian.

Keabsahan data dalam kajian pustaka ini dijaga melalui beberapa strategi, yaitu: (1) Triangulasi sumber, dengan membandingkan temuan dari berbagai jenis publikasi; (2) Kredibilitas sumber, dengan memprioritaskan jurnal bereputasi dan standar resmi; (3) Konsistensi analisis, dengan menggunakan kerangka teoretis yang diakui secara luas dalam pengukuran pendidikan (AERA et al., 2024). Melalui prosedur tersebut, hasil kajian diharapkan memiliki tingkat kepercayaan yang tinggi dan dapat dijadikan rujukan dalam pengembangan sistem pemberian skor dan penilaian di berbagai konteks pendidikan.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil sintesis literatur mengenai pemberian skor dan sistem penilaian, kemudian membahasnya secara kritis dengan mengaitkan teori, temuan empiris, serta implikasi praktis dalam konteks pendidikan. Pembahasan difokuskan pada empat aspek utama, yaitu desain sistem pemberian skor, reliabilitas dan validitas penilaian, fairness dan bias penilaian, serta pemanfaatan teknologi dan kecerdasan buatan dalam sistem penilaian modern.

Desain Sistem Pemberian Skor dan Rubrik Penilaian

Hasil kajian literatur menunjukkan bahwa desain sistem pemberian skor yang efektif sangat bergantung pada kejelasan kriteria penilaian dan keselarasan antara capaian pembelajaran, tugas, serta rubrik yang digunakan. Rubrik penilaian terbukti menjadi alat yang paling banyak direkomendasikan dalam penilaian kinerja, proyek, dan tugas autentik karena mampu meningkatkan transparansi dan konsistensi skor (Panadero & Jonsson, 2023).

Namun demikian, berbagai penelitian menegaskan bahwa keberadaan rubrik saja tidak secara otomatis menjamin kualitas penilaian. Rubrik yang dirancang secara umum, dengan deskriptor level yang ambigu, justru berpotensi menimbulkan interpretasi yang berbeda antarpenilai (Brookhart, 2023). Oleh karena itu, rubrik analitik dengan kriteria yang spesifik dan deskriptor yang dapat diamati secara empiris lebih disarankan dibandingkan rubrik holistik, terutama untuk tujuan formatif dan pengembangan kompetensi peserta didik (Panadero et al., 2025).

Temuan lain menunjukkan bahwa pembobotan kriteria dalam rubrik harus mencerminkan prioritas kompetensi yang ingin dicapai. Ketidakseimbangan bobot dapat menyebabkan skor akhir lebih merepresentasikan aspek tertentu saja, misalnya kemampuan teknis, dan mengabaikan aspek lain seperti komunikasi atau pemikiran kritis (Vlachopoulos & Makri, 2024). Oleh sebab itu, desain sistem pemberian skor harus dilakukan secara sadar dan berbasis pada analisis tujuan pembelajaran.

Reliabilitas Penilaian dan Variasi Penilai

Hasil kajian literatur mengungkap bahwa reliabilitas masih menjadi tantangan utama dalam sistem penilaian berbasis kinerja dan esai. Variasi penilai (rater variability) muncul sebagai salah satu sumber kesalahan pengukuran terbesar, terutama ketika penilaian melibatkan lebih dari satu penilai dengan latar belakang dan pengalaman yang berbeda (Shavelson & Webb, 2022).

Pendekatan Generalizability Theory banyak direkomendasikan dalam literatur terkini karena mampu mengidentifikasi sumber-sumber error pengukuran secara lebih komprehensif dibandingkan pendekatan reliabilitas klasik. Melalui pendekatan ini, peneliti dan praktisi dapat menentukan apakah reliabilitas lebih efektif ditingkatkan dengan menambah jumlah penilai, jumlah tugas, atau memperbaiki desain rubrik (AERA et al., 2024).

Selain itu, penggunaan Many-Facet Rasch Measurement (MFRM) terbukti efektif dalam mendeteksi perbedaan tingkat keketatan (severity) dan kelonggaran (leniency) antarpenilai. Studi oleh Khamboonruang (2023) menunjukkan bahwa penilai dapat secara signifikan berbeda dalam memberikan skor, meskipun menggunakan rubrik yang sama. Temuan ini memperkuat pentingnya pelatihan penilai, penggunaan contoh jangkar (anchor samples), serta moderasi penilaian secara berkala untuk meningkatkan konsistensi skor.

Dalam konteks praktis, hasil kajian ini menunjukkan bahwa institusi pendidikan tidak dapat sepenuhnya bergantung pada rubrik tertulis, tetapi perlu mengintegrasikan prosedur pengendalian reliabilitas sebagai bagian dari sistem penilaian yang utuh.

Validitas Skor dan Interpretasi Nilai

Dari sisi validitas, hasil kajian menunjukkan bahwa validitas skor tidak hanya ditentukan oleh kualitas instrumen penilaian, tetapi juga oleh cara skor tersebut diinterpretasikan dan digunakan. Standar pengukuran pendidikan menekankan bahwa validitas merupakan argumen berbasis bukti yang harus didukung oleh data empiris dan rasional teoretis (AERA et al., 2024).

Penilaian yang tidak selaras dengan tujuan pembelajaran berisiko menghasilkan skor yang tidak valid, meskipun reliabel. Misalnya, penggunaan tes tertulis untuk mengukur keterampilan praktik atau kolaborasi sering kali tidak mampu merepresentasikan kompetensi yang sebenarnya (Vlachopoulos & Makri, 2024). Oleh karena itu, tugas autentik dan penilaian kinerja dipandang lebih sesuai untuk mengukur kompetensi kompleks, meskipun menuntut sistem pemberian skor yang lebih cermat.

Selain itu, literatur juga menekankan pentingnya mempertimbangkan konsekuensi penggunaan skor. Sistem penilaian yang terlalu berorientasi pada nilai akhir dapat menurunkan motivasi intrinsik peserta didik dan mendorong perilaku belajar yang dangkal (Brookhart, 2023). Dengan demikian, hasil kajian mendukung pergeseran paradigma menuju penilaian formatif yang menempatkan skor sebagai sarana umpan balik, bukan sekadar alat seleksi.

Fairness dan Bias dalam Sistem Penilaian

Isu fairness menjadi salah satu tema dominan dalam literatur lima tahun terakhir. Hasil kajian menunjukkan bahwa ketidakadilan dalam penilaian dapat muncul baik pada penilaian manual maupun otomatis. Dalam penilaian manual, bias penilai dapat

dipengaruhi oleh faktor non-akademik seperti bahasa, gaya komunikasi, atau persepsi subjektif terhadap peserta didik (WIDA, 2024).

Pada penilaian otomatis, fairness menjadi isu yang lebih kompleks. Beberapa penelitian menunjukkan bahwa sistem Automated Essay Scoring (AES) cenderung memberikan skor yang kurang akurat pada kelompok tertentu, terutama peserta didik dengan latar belakang bahasa non-dominan (Loukina et al., 2024). Hal ini disebabkan oleh ketergantungan model pada data latih yang tidak selalu representatif.

Oleh karena itu, literatur merekomendasikan evaluasi fairness secara eksplisit melalui analisis perbedaan performa model lintas kelompok, serta penerapan prinsip human-in-the-loop untuk keputusan bernilai tinggi seperti kelulusan atau sertifikasi (Williamson et al., 2024). Temuan ini menegaskan bahwa keadilan harus menjadi pertimbangan utama dalam setiap desain sistem pemberian skor, baik manual maupun berbasis teknologi.

Pemanfaatan Teknologi dan Kecerdasan Buatan dalam Penilaian

Hasil kajian menunjukkan bahwa pemanfaatan teknologi dan kecerdasan buatan dalam sistem penilaian memberikan peluang besar dalam meningkatkan efisiensi dan konsistensi penskoran. Model berbasis pembelajaran mendalam dan large language models (LLM) menunjukkan performa yang semakin mendekati penilai manusia dalam tugas penilaian esai (Xie et al., 2024).

Namun demikian, berbagai studi juga menegaskan bahwa AI grading belum mampu sepenuhnya menggantikan peran penilai manusia, terutama dalam menilai aspek kompleks seperti orisinalitas ide, kreativitas, dan konteks budaya (Williamson et al., 2024). Oleh karena itu, pendekatan hibrida yang mengombinasikan penilaian otomatis dan penilaian manusia dipandang sebagai solusi paling realistis dalam jangka menengah.

Selain itu, penggunaan AI dalam penilaian menuntut adanya tata kelola data yang ketat, transparansi algoritma, serta validasi berkelanjutan. Tanpa mekanisme tersebut, sistem penilaian berbasis AI berisiko menurunkan kepercayaan publik terhadap hasil penilaian dan institusi pendidikan secara keseluruhan (Loukina et al., 2024).

Implikasi Praktis bagi Sistem Penilaian Pendidikan

Berdasarkan hasil kajian dan pembahasan di atas, dapat disimpulkan bahwa sistem pemberian skor yang berkualitas harus dirancang secara komprehensif dengan mempertimbangkan aspek pedagogis, psikometris, dan etis. Institusi pendidikan perlu mengembangkan kebijakan penilaian yang menekankan transparansi, pelatihan penilai, serta evaluasi sistem penilaian secara berkelanjutan (AERA et al., 2024).

Hasil kajian ini juga menegaskan pentingnya peningkatan literasi asesmen bagi pendidik agar mereka mampu merancang dan menerapkan sistem penilaian yang tidak hanya akurat secara teknis, tetapi juga mendukung pembelajaran bermakna. Dengan demikian, pemberian skor tidak lagi dipandang sebagai akhir dari proses pembelajaran, melainkan sebagai bagian integral dari strategi peningkatan kualitas pendidikan.

Pemberian skor dan sistem penilaian merupakan komponen strategis dalam pendidikan yang memiliki implikasi langsung terhadap kualitas pembelajaran, keadilan akademik, serta pengambilan keputusan institusional. Berdasarkan hasil kajian literatur yang telah dibahas, dapat disimpulkan bahwa sistem penilaian yang efektif tidak hanya

menuntut kejelasan prosedur penskoran, tetapi juga harus didukung oleh bukti validitas, reliabilitas, dan fairness yang memadai (AERA, APA, & NCME, 2024) sesuai dengan hasil diskusi dan analisis dari Dini selfani, Liza Akmalia Lubis, Maya Nursyafitri.

Hasil kajian menunjukkan bahwa desain sistem pemberian skor yang selaras dengan capaian pembelajaran dan didukung oleh rubrik yang jelas mampu meningkatkan transparansi serta konsistensi penilaian. Namun demikian, rubrik tidak dapat berdiri sendiri tanpa pelatihan penilai dan mekanisme moderasi yang sistematis. Variasi penilai masih menjadi sumber utama kesalahan pengukuran dalam penilaian kinerja dan esai, sehingga pendekatan seperti Generalizability Theory dan Many-Facet Rasch Measurement sangat direkomendasikan untuk mengendalikan dan memantau reliabilitas skor (Shavelson & Webb, 2022; Khamboonruang, 2023).

Dari sisi validitas, kajian ini menegaskan bahwa skor harus dipahami sebagai hasil interpretasi berbasis bukti, bukan sekadar angka akhir. Penggunaan instrumen penilaian yang tidak selaras dengan tujuan pembelajaran berisiko menghasilkan skor yang menyesatkan dan berdampak negatif terhadap proses belajar. Oleh karena itu, penilaian autentik dan penilaian formatif perlu diintegrasikan secara sistematis agar skor berfungsi sebagai sarana umpan balik yang mendorong pembelajaran bermakna (Vlachopoulos & Makri, 2024; Panadero et al., 2025).

Aspek fairness juga muncul sebagai isu krusial dalam sistem penilaian modern. Baik penilaian manual maupun penilaian berbasis teknologi memiliki potensi bias yang dapat merugikan kelompok tertentu. Kajian ini menunjukkan bahwa evaluasi fairness harus menjadi bagian integral dari desain sistem penilaian, terutama dengan meningkatnya penggunaan penilaian otomatis dan kecerdasan buatan. Pendekatan human-in-the-loop dan audit bias secara berkala dipandang sebagai langkah penting untuk menjaga keadilan dan akuntabilitas penilaian (Loukina et al., 2024; Williamson et al., 2024).

Seiring dengan berkembangnya teknologi, pemanfaatan kecerdasan buatan dalam sistem penilaian menawarkan peluang besar dalam meningkatkan efisiensi dan konsistensi penskoran. Namun, kajian ini menegaskan bahwa teknologi tidak dapat menggantikan sepenuhnya peran pendidik sebagai pengambil keputusan pedagogis. Sistem penilaian berbasis AI harus diposisikan sebagai alat bantu yang melengkapi, bukan menggantikan, penilaian manusia, serta harus didukung oleh tata kelola data dan validasi berkelanjutan (Xie et al., 2024).

SIMPULAN

Secara keseluruhan, kajian ini menyimpulkan bahwa sistem pemberian skor yang berkualitas adalah sistem yang transparan, dapat diaudit, adil, dan berorientasi pada pembelajaran. Pengembangan sistem penilaian ke depan perlu mengintegrasikan pendekatan pedagogis dan psikometris secara seimbang, serta meningkatkan literasi asesmen pendidik agar penilaian benar-benar berkontribusi pada peningkatan mutu pendidikan. Temuan ini diharapkan dapat menjadi rujukan bagi pendidik, peneliti, dan pengambil kebijakan dalam merancang dan mengimplementasikan sistem penilaian yang lebih akurat dan berkeadilan.

DAFTAR PUSTAKA

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2024). *Standards for educational and psychological testing*. American Educational Research Association.

Andrade, H. L. (2022). Rubrics as formative assessment tools: Student perspectives and implications for practice. *Assessment & Evaluation in Higher Education, 47*(4), 543–556.

Andrade, H. L., Brookhart, S. M., & Yu, E. (2023). Classroom grading practices: Theory, evidence, and future directions. *Educational Assessment, 28*(2), 67–86.

Assingkily, M. S. (2021). *Metode Penelitian Pendidikan: Panduan Menulis Artikel Ilmiah dan Tugas Akhir.* Yogyakarta: K-Media.

Brookhart, S. M. (2021). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brookhart, S. M. (2023). *Grading and learning: Practices that support student achievement* (2nd ed.). ASCD.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.

de la Torre, J., & Minchen, N. (2021). Cognitive diagnostic assessment in educational measurement. *Educational Measurement: Issues and Practice, 40*(2), 40–53.

Fulcher, G. (2022). *Practical language testing*. Routledge.

Jonsson, A., & Panadero, E. (2021). The use and design of scoring rubrics in higher education. *Educational Research Review, 34*, 100398.

Khamboonruang, A. (2023). Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A many-facets Rasch measurement approach. *PASAA: Journal of Language Teaching and Learning in Thailand, 66*, 5–36.

Lane, S., Raymond, M. R., & Haladyna, T. M. (2022). *Handbook of test development* (2nd ed.). Routledge.

Loukina, A., Madnani, N., & Cahill, A. (2024). Fairness considerations in automated essay scoring. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* (pp. 1–11). Association for Computational Linguistics.

Messick, S. (2021). Validity of educational assessment: Historical foundations and current perspectives. *Educational Psychologist, 56*(3), 145–158.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th ed.). SAGE Publications.

Nitko, A. J., & Brookhart, S. M. (2022). *Educational assessment of students* (8th ed.). Pearson.

Panadero, E., & Jonsson, A. (2023). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 38*, 100479. https://doi.org/10.1016/j.edurev.2022.100479

Panadero, E., Andrade, H., & Brookhart, S. M. (2025). Using rubrics for formative purposes: Factors that influence their effectiveness. *Educational Assessment, 30*(1), 1–20.

Popham, W. J. (2021). *Classroom assessment: What teachers need to know* (9th ed.). Pearson.

Rasch, G. (2021). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.