



Penggunaan Penskor Jawaban Esai Otomatis dalam Pengukuran Pengetahuan Guru

Raden Ahmad Hadian Adhy Permana*, Ari Widodo, Wawan Setiawan, Siti Sriyati

Sekolah Pascasarjana Universitas Pendidikan Indonesia, Bandung 40154, Indonesia

*Email: radenahmadhadian@upi.edu.

DOI: 10.24815/jupi.v5i4.22724

Article History:

Received: September 17, 2021
Accepted: December 1, 2021

Revised: November 23, 2021
Published: December 6, 2021

Abstract. Measuring the ability of teachers is part of the evaluation in the national education system. The development of the instrument in the form of essay questions has become an alternative because the instrument is considered to have better authenticity than the form of multiple choice questions. The implications of using the essay instrument were scored consistency and efficient resources. The purpose of this study was to examine the inter-rater reliability between the automated essay scoring system and human raters in measuring teachers' knowledge using the essay instrument. The research was conducted randomly with 200 junior high school science teachers. The quantitative method design was applied to investigate the intra-class correlation coefficient (ICC) and Pearson's correlation (r) as indicators of the ability of automated essay scoring (UKARA) that had been used. The main data in this study were test answers from participants in the form of limited essay answers that have been distributed online. The inter-rater reliability coefficient between the UKARA and the human rater in this study was in the high category (more than 0.7) for all items or means that the score given by UKARA has a strong correlation with the score given by human rater. These results mean that UKARA has adequate capability as an automated essay scoring system on the measurement of science teacher knowledge.

Keywords: Automated essay scoring, inter-rater reliability, competency measurement

Pendahuluan

Asesemen terhadap guru perlu dilakukan secara berkesinambungan dan kontekstual sebagai bagian dari evaluasi terhadap profesionalismenya. Pengetahuan guru menjadi salah satu faktor utama yang mempengaruhi berbagai aspek hasil belajar siswa (Akturk & Ozturk, 2019; Sukardi, dkk., 2017), sehingga perlu menjadi perhatian para pemangku kebijakan. Kemampuan guru adalah kemampuan yang bersifat dinamis, berkembang sesuai dengan perkembangan jaman, sehingga asesmen yang dilakukan juga seharusnya menyesuaikan dengan perkembangan kemampuan yang dibutuhkan atau standar pendidikan yang ditetapkan. Perubahan yang dapat dilakukan untuk asesmen terhadap guru antara lain adalah kerangka kerja yang menjadi dasar pengembangan instrumen yang sesuai dengan perkembangan zaman, dan penggunaan teknologi dalam sistem evaluasi (Kuntarto, dkk., 2019; Luckin, 2017; Zlatkin-Troitschanskaia, dkk., 2019).

Salah satu perkembangan dalam pengukuran adalah penggunaan instrumen dalam bentuk isian singkat, uraian terbatas, ataupun bentuk lain yang dikenal dengan jawaban yang dikonstruksi. Bentuk instrumen tersebut menjadi alternatif yang lebih baik bagi soal

dalam berupa pilihan ganda yang dianggap lebih tradisional dalam pengukuran pengetahuan guru (Liu, dkk., 2016). Instrumen soal uraian dapat menjadi pilihan karena beberapa kelebihanannya, antara lain menunjukkan persepsi asli subjek sebagai jawaban, mengurangi kemungkinan jawaban tebakan, penyusunan lebih efisien, perangkat soal lebih sederhana (Khan & Ishfaq, 2016; Rios & Wang, 2018). Bentuk-bentuk instrumen yang memiliki karakter lebih autentik digunakan untuk mendapatkan hasil asesmen yang lebih akurat dan sesuai tujuan evaluasi yang dilakukan, walaupun umumnya menunjukkan nilai yang terlihat lebih rendah dibandingkan hasil tes lain yang lebih tradisional (Aalaei, dkk., 2016; Darling-Hammond, dkk., 2012; Nkhoma, dkk., 2020; van Helvoort, dkk., 2017).

Penilaian dengan instrumen soal uraian yang jawabannya dikonstruksi sendiri oleh subjek tes memiliki beberapa keunggulan dibandingkan penilaian pilihan ganda tradisional (Rios & Wang, 2018). Instrumen esai yang dimaksud sebagai alternatif tersebut adalah instrumen berbentuk soal dengan jawaban uraian singkat atau uraian terbatas, bukan esai terbuka atau suatu paragraf. Selama ini peran untuk soal jawaban pendek umumnya sebagai pelengkap, tidak hanya dari esai yang lebih panjang tetapi juga berupa respon tradisional yang dipilih (Brew & Leacock, 2013). Instrumen esai tersebut pada dasarnya memerlukan sumber daya besar dalam penskorannya, tetapi dengan perkembangan teknologi saat ini hambatan tersebut sudah dapat diatasi (Sevcikova, 2018; Latifi & Gierl, 2021). Pengembangan sistem yang secara otomatis menilai atau menskor jawaban uraian tersebut dapat membantu mengurangi biaya, personel, dan waktu yang diperlukan secara signifikan.

Pada dua dekade terakhir ini telah banyak sistem penskoran otomatis yang dikembangkan, antara lain Intelligent Essay Assessor, Intellimetric, Project Essay Grade, dan e-rater (Valenti, dkk., 2003; Shin & Gierl, 2020). Masing-masing AES menggunakan sistem kerja yang berbeda-beda, tetapi persamaannya adalah sistem tersebut tidak benar-benar membaca dan memahami esai seperti halnya manusia. Penilai manusia dapat secara langsung mengevaluasi berbagai variabel intrinsik yang menarik, seperti diksi, kefasihan, atau tata bahasa, sementara sistem AES menggunakan kesamaan, perkiraan, atau kemungkinan korelasi antara jawaban dan basis data yang dimiliki sistem (Attali & Burstein, 2006). AES dengan basis kecerdasan artifisial dapat menentukan skor antara lain berdasarkan perbandingan kesamaan dengan kunci jawaban yang diinput ataupun dilatih dengan menginput jawaban dan hasil penskoran oleh manusia.

Agar hasil penskoran bermanfaat dan dapat dipercaya, hasil tersebut harus memenuhi ekspektasi kualitas tertentu (Chung & Baker, 2003). Kriteria kualitas untuk penilaian tersebut secara umum adalah validitas dan reliabilitas dari penilaian yang dilakukan. Selain itu dalam penentuan skor, perlu diyakinkan pula standarisasi dari proses yang dilakukan. Kualitas yang tinggi dalam penskoran oleh manusia dapat diperoleh dari penggunaan rubrik penskoran yang valid dan penskor yang kompeten. Sementara validitas penggunaan sistem penskoran otomatis adalah validasi yang diberikan oleh penskor manusia atau kunci jawaban yang diinput ke dalam sistem serta kapabilitas sistem yang digunakan secara digital, misalnya algoritma yang digunakan. Reliabilitas untuk sistem penskoran otomatis yang digunakan dapat diukur secara statistik melalui korelasi dengan hasil penskoran oleh penskor manusia. Proses mengukur reliabilitas antar penskor tersebut adalah dilakukan dengan cara memperkirakan konsistensi penskoran antar penskor dan kesepakatan di antara penskor-penskor tersebut (Keith, 2003).

Penilaian otomatis dapat berkontribusi pada keandalan, kelayakan, dan keefektifan untuk instrumen jawaban singkat atau jawaban terbatas, seperti juga untuk jawaban esai yang lebih panjang (Brew & Leacock, 2013). Soal jawaban singkat menekankan ketepatan konten sehingga tidak sama dengan esai yang umumnya berfungsi menekankan tata bahasa dan mekanika. *Automated Essay Scoring* (AES) atau penskor esai secara otomatis secara umum adalah kemampuan teknologi komputer untuk mengevaluasi dan menilai tulisan (Shermis & Burstein, 2003). Perkembangan teknologi informasi memberikan

dampak positif dalam wilayah pendidikan, antara lain pemanfaatan teknologi untuk penilaian secara elektronik yang memungkinkan penilaian dilakukan secara otomatis, dengan menggunakan sistem penilaian otomatis (Hasanah, dkk., 2019).

Sistem penilaian otomatis memiliki beberapa keunggulan, seperti mampu menilai jawaban dengan lebih objektif, konsisten, dan lebih cepat, selain itu manusia bisa salah saat melakukan penilaian dan memungkinkannya adanya kesepakatan antar penilai yang tidak sempurna karena kelelahan, bias, atau efek pengurutan (Cohen, 2017; Haley, dkk., 2007). Alat penilaian otomatis dengan keakuratan yang terbukti akan dapat membantu meningkatkan penggunaan penilaian terbuka (Liu, dkk., 2016). Implikasi penting lainnya dari penilaian otomatis adalah umpan balik yang diberikan kepada pihak berkepentingan dalam pengukuran dapat mengarahkan program tindak lanjut yang lebih akurat. Hal ini serupa dengan penjelasan Linn, dkk. (2014), bahwa penggunaan penskor otomatis dapat mencapai tingkat untuk memberikan bimbingan yang sama efektifnya dengan bimbingan guru dalam mendorong siswa untuk meningkatkan hasil belajar. Sistem penskoran esai otomatis kebanyakan didasarkan pada kecerdasan buatan (*artificial intelligence/AI*), linguistik komputasi, dan dasar ilmu kognitif, bukan pada psikometri tradisional dan pengukuran pendidikan (Landauer, dkk., 2003). Sistem penilaian otomatis memiliki potensi untuk menghasilkan tingkat efisiensi ini untuk berbagai jenis soal yang kompleks (Clauser, dkk., 2002). Akurasi penskor otomatis untuk jawaban esai telah diuji dan menunjukkan hasil yang beragam, tetapi umumnya telah memiliki kemampuan memadai untuk memberikan skor terhadap data-data penelitian (Ahmad, 2019; Aji, dkk., 2011; Attali & Burstein, 2006; Byrne, dkk., 2010; Liu, dkk., 2016; Valenti, dkk., 2003).

Reliabilitas yang tinggi merupakan salah satu indikator performa yang optimal dari penggunaan penskor otomatis (Chung & Baker, 2003). Pendapat lain menurut Liu, dkk. (2016), bahwa kesalahan dalam penelitian mengenai penskoran otomatis kemungkinan berasal dari pengaruh batasan ukuran sampel, rendahnya persentase kesepakatan antar penskor manusia, dan kendala respon secara kebahasaan (leksikal dan sintaksis). Hal tersebut dipengaruhi oleh varian respon dan skor yang diberikan oleh penskor manusia maupun komputer. Frekuensi rendah dari respons tertentu berkontribusi pada penilaian yang kurang memuaskan oleh mesin otomatis (Ha, dkk., 2011).

Permasalahan yang dibahas dalam penelitian ini didasari oleh pertanyaan penelitian: "Bagaimana reliabilitas antara hasil penskoran otomatis oleh program komputer UKARA dan hasil penskoran oleh penilai manusia?". Tujuan penelitian yang telah dilakukan ini adalah untuk mengetahui reliabilitas antar penskor otomatis dengan penskor manusia dalam pengukuran pengetahuan guru IPA menggunakan instrumen soal esai. Hasil penelitian ini bermanfaat untuk meningkatkan efisiensi pengukuran tingkat pengetahuan guru IPA, dalam aspek bentuk instrumen dan fungsi asesmen pada skala luas.

Metode

Pendekatan dan metode penelitian yang diterapkan adalah penelitian kuantitatif dengan analisis reliabilitas secara statistik. Data dalam penelitian ini adalah jawaban yang diberikan oleh 200 orang guru IPA SMP di Provinsi Banten sebagai subjek dalam tes pengetahuan mengenai pembelajaran dan penilaian. Partisipan ditetapkan secara acak dan bersifat *convenience sampling* (Etikan, 2016). Penjaringan partisipan melalui kelompok-kelompok kerja guru yang ada di setiap wilayah kabupaten/kota. Tes dilakukan menggunakan 15 butir soal yang disampaikan kepada partisipan melalui internet dalam bentuk Google Form. Partisipan menjawab soal-soal tersebut secara langsung dalam moda internet (*online*). Hasil jawaban secara otomatis terekam berbentuk file Excel dalam akun Google Drive. Seluruh butir soal tes berbentuk pertanyaan yang meminta jawaban esai

atau uraian terbatas. Instrumen esai tersebut telah diuji pada penelitian sebelumnya dan diketahui memiliki validitas dan reliabilitas yang tinggi. Jawaban dari partisipan kemudian diberikan skor oleh seorang penilai manusia sebagai penskor 1 dan program komputer UKARA sebagai penskor 2. Proses penilaian jawaban oleh penskor 1 dipandu dengan penggunaan rubrik penskoran yang menjadi kelengkapan instrumen tes. Penskor 2 adalah suatu program penskoran jawaban uraian otomatis yang saat ini dikelola oleh Pusat Asesmen Pendidikan, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. Proses penskoran dengan komputer tersebut melalui proses input jawaban ke dalam program dan kemudian sistem penskoran bekerja berbasis kecerdasan buatan untuk menghasilkan skor. Data yang dihasilkan oleh kedua penskor berupa jawaban dan skor untuk masing-masing subjek dan butir soal.

Pengolahan awal data untuk bahan analisis meliputi identifikasi perbedaan jawaban secara kualitatif, rekapitulasi hasil penskoran secara keseluruhan, dan rekapitulasi hasil penskoran berdasarkan jumlah jawaban tiap skor untuk mengetahui komposisinya. Jenis jawaban dikelompokkan berdasarkan kompleksitas jawaban yang sesuai dengan rubrik penskoran yang digunakan oleh penskor manusia, yaitu jenis jawaban satu kalimat sederhana, jenis jawaban dalam kalimat majemuk, dan jenis jawaban dalam bentuk beberapa poin atau kalimat terpisah. Kategori jawaban terkait dengan komposisi ditetapkan secara kontekstual, yaitu terbagi menjadi jawaban dengan skor proporsional (25-50% tiap skor) dan tidak proporsional (kurang dari 25% atau lebih dari 50% tiap skor). Rekapitulasi penskoran keseluruhan menjadi bahan untuk penghitungan reliabilitas antar penskor untuk setiap butir soal. Analisis data dilakukan secara kuantitatif berdasarkan nilai reliabilitas antar penskor berupa *intraclass correlation coefficient* (ICC) dan nilai korelasi Pearson (Finch & French, 2018). Identifikasi perbedaan jawaban dan rekapitulasi menurut komposisi jumlah jawaban tiap skor jawaban menjadi bahan untuk analisis perbandingan hasil perhitungan reliabilitas antar penskor. Pengolahan data menggunakan perangkat lunak SPSS 2.0 dan *microsoft office excel*. Interpretasi dan analisis dilakukan berdasarkan hasil pengolahan data kuantitatif tersebut.

Hasil dan Pembahasan

Hasil kajian bentuk jawaban pada rubrik penskoran dan komposisi skor untuk tiap butir soal disajikan pada Tabel 1. Informasi pertama yang ditunjukkan pada Tabel 1 adalah mengenai jenis jawaban yang telah terbagi menjadi tiga kelompok, yaitu kalimat sederhana (5 soal), kalimat majemuk (7 soal), dan kalimat berupa daftar jawaban (3 soal). Berdasarkan penentuan tersebut, maka kelompok I (kalimat sederhana) dianggap jawaban yang paling mudah untuk dipahami, sehingga kemungkinan perbedaan hasil penskoran antara dua penskor menjadi lebih kecil. Kelompok kedua adalah jawaban yang terdiri dari dua pokok jawaban, masih mudah dipahami, tetapi ada kemungkinan perbedaan penskoran yang lebih tinggi daripada jenis jawaban pertama. Soal yang termasuk pada kelompok ketiga menjadi jawaban yang lebih sulit ditentukan skornya karena memiliki kompleksitas yang paling tinggi sehingga kemungkinan perbedaan hasil penskoran juga menjadi paling tinggi. Konsistensi hasil dari dua penskor berbeda akan menentukan nilai reliabilitas suatu instrumen, sehingga pemahaman atau persepsi terhadap jawaban menjadi faktor penentu (van Helvoort, dkk., 2017). Pengelompokan ini tidak dirancang

secara khusus, karena kisi-kisi instrumen soal hanya didasarkan kepada tingkat kesulitan secara konseptual, bukan berdasarkan kompleksitas kalimat jawaban yang diberikan.

Tabel 1. Hasil Kajian Jawaban

No. Soal	Jenis jawaban	Komposi Skor Jawaban
1	II	Tidak proporsional
2	I	Tidak proporsional
3	III	Tidak proporsional
4	II	Proporsional
5	II	Tidak proporsional
6	I	Tidak proporsional
7	II	Tidak proporsional
8	II	Tidak proporsional
9	III	Tidak proporsional
10	III	Tidak proporsional
11	I	Proporsional
12	I	Tidak proporsional
13	II	Proporsional
14	II	Tidak proporsional
15	I	Tidak proporsional

Keterangan: I: kalimat sederhana; II: kalimat majemuk; III: daftar jawaban

Informasi kedua yang dapat dilihat pada Tabel 1 adalah bahwa sebagian besar jawaban soal tidak memiliki komposisi skor yang proporsional dari hasil penskoran yang diperoleh. Hanya butir soal nomor 4, 11, dan 13 yang mendapatkan skor secara proporsional atau dalam 25 - 50% untuk masing-masing skor. Sementara 12 butir soal lainnya tidak mendapatkan skor dengan komposisi yang proporsional atau adanya kecenderungan yang tinggi (lebih dari 50%) pada skor tertentu. Komposisi yang ideal dalam proses kalibrasi merupakan faktor yang diharapkan karena kecerdasan buatan pada sistem penskoran akan mendapat input yang lebih baik dibandingkan jika komposisi skor jawaban yang diinput tidak proporsional (Vacha-Haase, dkk., 2000). Instrumen penskoran pada penelitian ini menggunakan skala 0 – 2 atau tiga tingkat penskoran, sehingga proporsi jawaban secara kuantitatif yang ideal adalah sekitar 30% untuk masing-masing skor.

Hasil kalkulasi secara statistik untuk nilai reliabilitas dan persentase kesepakatan seperti terlihat pada Tabel 2. Hasil pengujian reliabilitas ICC yang ditunjukkan pada Tabel 2 untuk semua soal memiliki reliabilitas antar penskor yang memadai. Seluruh soal telah memiliki koefisien minimal 0,7 atau berkategori reliabilitas baik. Secara rata-rata juga terlihat bahwa ICC telah mencapai nilai 0,77. Hasil uji coba yang ditunjukkan pada Tabel 2 tersebut adalah uji coba menggunakan data penskoran secara kontekstual atau berbasis data penskoran sebenarnya dari 200 subjek penelitian. Hasil penskoran oleh komputer juga tidak sama persis dengan hasil penskoran oleh manusia sesuai dengan nilai reliabilitas yang diperoleh, jika penskoran oleh komputer sama persis dengan hasil penskoran oleh manusia, maka nilai reliabilitas yang diperoleh adalah 1. Selanjutnya hasil penskoran oleh penskor manusia menjadi basis untuk penskoran yang dilakukan menggunakan program komputer.

Tabel 2. Koefisien Reliabilitas Antar Penskor

No. Soal	ICC	r
1	0,87	0,87
2	0,76	0,77
3	0,68	0,70
4	0,76	0,76
5	0,75	0,75
6	0,70	0,70
7	0,76	0,76
8	0,75	0,75
9	0,90	0,90
10	0,67	0,70
11	0,82	0,83
12	0,74	0,75
13	0,79	0,79
14	0,83	0,83
15	0,75	0,75
Rerata	0,77	0,80

Hasil uji coba yang ditunjukkan pada Tabel 2 tersebut juga adalah korelasi r (Pearson) dengan kategori kuat (≥ 7) sampai sangat kuat (≥ 9). Hasil pengujian terkait korelasi ini sesuai dengan penjelasan Keith (2003), bahwa korelasi antara penskoran otomatis dan penskoran oleh manusia akan bervariasi tergantung pada faktor dalam percobaan atau proses yang dilakukan. Faktor-faktor yang mempengaruhi tersebut antara lain jenis fitur pada sistem penskoran otomatis (Ke & Ng, 2019), data dasar ke dalam kecerdasan buatan (Liu, dkk., 2016; Taghipour & Ng, 2016), serta adanya hubungan antara reliabilitas dan jumlah data (Khan & Ishfaq, 2016). Untuk hasil pada Tabel 2, seperti telah dijelaskan sebelumnya, adalah data sebenarnya, bukan data khusus untuk pelatihan/kalibrasi.

Keterkaitan antara nilai ICC dan korelasi seperti dijelaskan oleh Fleenor, dkk. (1996) dan Liu, dkk. (2016), bahwa ada hubungan antara reliabilitas antar penskor dan koefisien korelasi (Pearson). Dalam artikel yang disusunnya, Fleenor dkk. juga menyebutkan istilah korelasi Pearson sebagai "*interrater reliability* Pearson". Berdasarkan hasil pengujian pada penelitian ini terlihat bahwa nilai ICC dengan kategori baik secara rata-rata ditunjukkan dengan korelasi yang tinggi pula. Berdasarkan hasil analisis korelasi dan deskriptif tidak terlihat adanya hubungan yang signifikan antara jenis kalimat jawaban dan koefisien reliabilitas. Hasil analisis korelasi secara statistik antara jenis kalimat jawaban dan koefisien reliabilitas (ICC) diperoleh nilai korelasi kurang dari 0,1 atau dalam kategori sangat rendah. Sementara pada analisis secara deskriptif tidak terlihat adanya pola yang konsisten antara jenis kalimat dan koefisien reliabilitas. Untuk hasil ini terlihat berbeda jika dibandingkan beberapa hasil penelitian lain, misalnya yang telah dijelaskan oleh van Helvoort, dkk. (2017). Informasi tersebut secara spesifik ditunjukkan pada butir soal nomor 9 yang memiliki nilai ICC tertinggi, tetapi butir soal nomor 10 merupakan mendapat nilai ICC terendah dari 15 butir soal yang telah diuji coba. Hasil analisis statistik menunjukkan bahwa koefisien korelasi 0,17 atau termasuk kategori hubungan antara nilai ICC dan komposisi skor pada penelitian ini tidak signifikan atau sangat lemah. Hasil analisis statistik tersebut juga didukung dari kajian secara deskriptif, yaitu 3 butir soal yang memiliki komposisi jawaban proporsional tidak mendapatkan ICC paling tinggi ataupun paling rendah. Hasil tersebut menjadi indikasi kuantitatif bahwa proporsi skor jawaban tidak menentukan secara nyata nilai reliabilitas antar penskor pada penelitian ini. Walaupun demikian semua butir soal dengan komposisi yang proporsional mendapatkan ICC lebih dari 0,7 – 0,8 secara konsisten atau sesuai dengan hasil-hasil penelitian sebelumnya.

Hasil uji coba ini secara keseluruhan menunjukkan skor hasil penilaian otomatis berpotensi baik untuk soal yang cukup kompleks. Studi sebelumnya juga telah menunjukkan adanya potensi penilaian otomatis untuk soal sains dalam skala besar dan penilaian kelas (Kuo & Wu, 2013; Linn, dkk., 2014; Liu, dkk., 2016; Nehm & Haertig, 2012). Konsisten dengan penelitian sebelumnya, UKARA secara keseluruhan dianggap menunjukkan hasil yang memadai dan berpotensi untuk ditingkatkan dalam penilaian soal menggunakan 3 jenjang skor. Hasil pengujian yang ditunjukkan oleh nilai ICC dan korelasi dengan penskor manusia pada kategori yang baik dan sangat baik.

Reliabilitas didefinisikan oleh para ahli sebagai pengukuran konsistensi antara penilai dalam urutan atau posisi relatif dari nilai performa, terlepas dari nilai absolut yang diberikan setiap penilai (Graham, dkk., 2012). Reliabilitas antar penskor yang tinggi tentunya diharapkan dalam hasil kalibrasi terhadap UKARA ini, tetapi berdasarkan penjelasan Attali & Burstein (2006) dalam hasil penelitiannya mengenai E-rater, reliabilitas penskoran antara manusia dan mesin tidak akan pernah sekuat reliabilitas antara penskor manusia dan manusia. Oleh karena itu untuk mencapai nilai reliabilitas yang sangat tinggi dalam penskoran soal uraian merupakan suatu permasalahan yang cukup sulit (Khan & Ishfaq, 2016).

UKARA adalah suatu sistem penskoran yang bekerja dengan adanya proses komputasi oleh kecerdasan artifisial yang tertanam dalam perangkat lunak tersebut. Penggunaan algoritma dalam sistem kecerdasan artifisial berbeda dengan aplikasi komputer yang berkembang lebih awal untuk penskoran, yaitu memberikan skor berdasarkan kunci jawaban yang diberikan oleh programmer (Zimmerman, dkk., 2018). E-rater adalah sistem penskor pertama yang digunakan dalam penilaian skala besar dan berisiko tinggi (Shermis, dkk., 2013). Pada sebagian besar aplikasi, program penskoran esai otomatis pertama-tama "dilatih" pada sampel esai yang telah dinilai oleh penilai manusia dan kemudian diuji secara statistik menggunakan multipel regresi yang melibatkan pemilihan sekumpulan variabel prediktor dan bobot regresi optimal (Keith, 2003). Begitu pula sebaliknya, diharapkan semakin banyak contoh jawaban dalam kalibrasi atau pelatihan maka semakin kaya "kamus" yang dibuat oleh program komputer tersebut. Penggunaan kata atau struktur tata bahasa yang berbeda untuk mengekspresikan ide yang serupa dari jawaban subjek dapat menjadi sumber kesalahan atau perbedaan dalam penskoran oleh komputer (Ha, dkk., 2011; Liu, dkk., 2016). Konsistensi menjadi suatu aspek esensial yang dimiliki oleh sistem penskoran otomatis dan menjadi kelemahan yang dimiliki oleh penskor manusia, khususnya untuk penskoran terhadap hasil tes esai dalam skala besar.

Model penskor esai otomatis umumnya divalidasi dengan cara membandingkan kesamaan hasil penskorannya dengan hasil penskoran manusia untuk data yang sama (Attali & Burstein, 2006; Chung & Baker, 2003; Graham, dkk., 2012; Williamson, dkk., 2010). *Rater* atau penskor manusia dianggap sebagai standar untuk penskoran yang akan dilakukan lebih luas oleh penskor otomatis, artinya hasil perbandingan kesamaan menjadi penentu kualitas penskor otomatis (Powers, dkk., 2015; Shermis, dkk., 2010; Toranj & Ansari, 2012; Wahlen, dkk., 2020; Williamson, dkk., 2010; Zhang, 2013). Korelasi antara program penskoran esai otomatis dan penilai manusia akan bervariasi tergantung pada kalibrasi yang dilakukan, sampel yang digunakan untuk menghitung korelasi, atau validasi silang yang digunakan (Keith, 2003). Ukuran perbandingan antara kedua penskor berupa koefisien hasil analisis statistik antara lain menggunakan skor kappa, skor kappa kuadrat (QWK), korelasi Pearson, korelasi Spearman, dan Kendall's Tau (Taghipour & Ng, 2016). Graham, dkk. (2012) membandingkan hasil komputasi untuk kesepakatan antar penskor dan reliabilitas antar penskor antara hasil manusia dan mesin, yaitu berupa persentase absolut, koefisien Kappa Cohen, dan koefisien korelasi *intra-class* (ICC).

Faktor-faktor dalam suatu uji reliabilitas dapat menjadi pembatas atau kendala dalam mencapai nilai reliabilitas yang memadai. Untuk mencapai nilai reliabilitas yang sangat

tinggi dalam penskoran soal uraian merupakan suatu permasalahan yang cukup sulit, keterbatasan jumlah subjek dalam penskoran yang dilakukan dapat menjadi kendala (Khan & Ishfaq, 2016). Komposisi sampel dapat mempengaruhi reliabilitas penskoran (Vacha-Haase, dkk., 2000). Faktor lain dalam percobaan yang dapat mempengaruhi reliabilitas berupa proses pelatihan atau kalibrasi, sampel yang digunakan untuk menghitung korelasi, atau jika sampel validasi lain yang digunakan (silang) (Keith, 2003). Jika penelitian validasi hanya dilakukan dengan sampel kalibrasi, perkiraan korelasi tersebut akan meningkat atau lebih tinggi. Sementara jika menggunakan sampel data lain akan lebih bervariasi tergantung data yang digunakan. Penelitian ini memiliki keterbatasan dalam jumlah sampel, dan data yang alami dari jawaban partisipan sehingga variabilitas jawaban partisipan cenderung tinggi tetapi memiliki komposisi yang tidak seimbang.

Penggunaan program komputer berbasis kecerdasan buatan untuk penskoran tentunya berbeda dengan program komputer untuk analisis butir soal atau penskoran yang lebih terdahulu dikembangkan. Algoritma yang digunakan dalam sistem komputer berbasis kecerdasan buatan tidak menskor menggunakan kata kunci atau kunci jawaban secara persis, tetapi mengolahnya seperti otak manusia bekerja berdasarkan "pengetahuan" yang diperoleh sebelumnya (Shin & Gierl, 2020; Zimmerman, dkk., 2018). Penjelasan lain dari Attali & Burstein (2006), penskor otomatis dapat bekerja secara dependen atau tidak terikat kepada penskoran yang diberikan oleh penskor manusia dalam sistem komputasi yang diterapkan. Proses ini ditunjukkan dalam data tersebut, yaitu reliabilitas tidak selalu mencapai nilai sangat kuat pada uji coba secara kontekstual atau kondisi sebenarnya. Reliabilitas terlihat bervariasi tanpa ada pola tertentu pada jawaban-jawaban dengan tipe yang berbeda dan komposisi skor yang tidak proporsional.

Penelitian ini menunjukkan bahwa tipe jawaban yang diberikan skor dan komposisi bahan untuk kalibrasi sistem penskoran tidak berkorelasi secara nyata dengan koefisien reliabilitas antar penskor. Dalam salah satu hasil penelitian sebelumnya telah dijelaskan bahwa hasil penskoran oleh komputer dapat dipengaruhi oleh beberapa faktor seperti panjang kalimat jawaban atau jumlah kata yang digunakan, gaya bahasa, kesalahan grammar, penyusunan kalimat, variasi kata yang digunakan, dan jumlah contoh jawaban untuk kalibrasi kecerdasan buatan yang digunakan (Ramalingam, dkk., 2018; Horbach & Zesch, 2019). Penjelasan tersebut menjadi dasar bahwa faktor tipe jawaban yang diharapkan sebenarnya juga dapat mempengaruhi kualitas penskoran otomatis, tetapi pada penelitian ini hal tersebut tidak ditunjukkan secara jelas. Sementara untuk komposisi skor pada jawaban untuk kalibrasi berawal dari prediksi bahwa komposisi sampel dapat mempengaruhi reliabilitas penskoran atau adanya keterkaitan antara komposisi dan reliabilitas (Vacha-Haase, dkk., 2000). Tetapi hal tersebut tidak mudah untuk diperoleh ketika jumlah subjek pengukuran relatif sedikit dan uji coba dilakukan menggunakan data asli jawaban dari partisipan tersebut.

Potensi yang memadai dari suatu penskor esai otomatis, seperti yang dijelaskan sebelumnya, tidak terlepas dari adanya kesalahan atau *error* dalam proses penskoran dan pengujian yang dilakukan. Menurut Liu, dkk. (2016), kesalahan dari penelitian seperti ini kemungkinan berasal dari pengaruh tiga sumber kesalahan, yaitu: (1) batasan ukuran sampel; (2) kurangnya statistik kesepakatan manusia-manusia; dan (3) kemungkinan respons atau jawaban secara kebahasaan (leksikal dan sintaksis). Hasil pengujian terhadap sampel yang lebih besar dapat menunjukkan reliabilitas dan korelasi yang tinggi, tetapi di sisi lain sampel dalam jumlah yang lebih kecil dapat menunjukkan persentase kesepakatan yang lebih tinggi dan reliabilitas lebih rendah. Hal tersebut dipengaruhi oleh varian respon dan skor yang diberikan oleh penskor manusia maupun komputer. Ha, dkk. (2011) menjelaskan dalam laporan hasil penelitiannya bahwa frekuensi rendah dari respons tertentu berkontribusi pada penilaian yang kurang memuaskan oleh mesin otomatis. Hasil yang diberikan oleh dua macam penskor otomatis yang berbeda untuk jumlah data yang sama dapat menunjukkan perbedaan (Shin & Gierl, 2020). Oleh karena itu meningkatkan

ukuran sampel atau jumlah jawaban subjek dapat memungkinkan evaluasi yang lebih baik penskor UKARA tersebut.

Faktor kendala yang muncul dari jawaban subjek adalah faktor kebahasaan dalam jawaban. Variasi leksikal dan sintaksis dari respons, keragaman bahasa, sinonim, dan salah eja adalah jenis variasi lain yang mungkin tidak dimodelkan secara memadai oleh mesin penilaian (Ha, dkk., 2011). Kendala tersebut juga terjadi pada jawaban subjek dalam penelitian ini, dalam konteks yang berbeda. Penelitian di negara lain mungkin memperhatikan adanya keragaman walaupun dalam bahasa nasional yang sama, tetapi dalam konteks Indonesia mungkin sedikit berbeda. Seperti yang dijelaskan Crossley dkk. (2016), untuk soal uraian nilai penting tidak hanya dari konten yang ditanyakan tetapi juga penggunaan bahasa oleh subjek dalam jawaban yang diberikan memiliki variabilitas dan dapat mempengaruhi penskoran. Penelitian ini dilakukan dalam lingkup satu provinsi, sehingga keragaman bahasa atau sintaksis tidak terlalu tinggi. Satu hal yang perlu diperhatikan dalam jawaban subjek adalah kebiasaan menyingkat kata, seperti ketika melakukan obrolan di media sosial. Banyak sekali subjek yang menjawab tidak menulis secara lengkap, tetapi menyingkat kata-kata yang digunakan. Permasalahan ini mungkin tidak akan terjadi atau dapat diminimalisir jika proses pengisian soal didampingi secara intensif atau tes yang dilakukan adalah tes yang menentukan untuk para subjek. Sementara pada proses dalam penelitian ini, subjek tidak terbebani dengan hasil yang akan mereka peroleh, sehingga kalimat jawaban banyak dibuat secara tidak formal.

Sumber kesalahan yang potensial menurunkan hasil uji coba dalam tahap penyiapan dapat terkait dengan akurasi penskoran manusia. Konsistensi penskoran manusia umumnya dibatasi oleh jumlah respons yang harus diberi skor. Keandalan skor yang dihasilkan oleh penskor manusia dapat secara langsung mempengaruhi kualitas dari hasil penskoran otomatis (Liu, dkk., 2014, 2016). Penskoran yang diberikan oleh beberapa penskor manusia dan kemudian dijadikan satu set hasil merupakan salah satu aspek yang mempengaruhi dalam penggunaan komputer secara otomatis. Penskoran oleh manusia tersebut umumnya juga menunjukkan adanya perbedaan, walaupun mungkin kesepakatan dan reliabilitas antar penskor sudah tinggi (Ramalingam, dkk., 2018). Selain itu, proses awal dari penyusunan soal dan penyusunan rubrik juga telah dioptimalkan, tetapi faktor subjektivitas dan konsistensi penskor manusia tetap menjadi faktor yang dapat menjadi kendala. Sementara penskor otomatis memiliki kemampuan yang dipengaruhi oleh penyiapan atau kalibrasi yang dilakukan berdasarkan penskoran oleh manusia. Penggunaan data hasil penskoran manusia adalah tahapan yang tidak dapat dihindari saat ini, termasuk untuk UKARA.

Hasil penelitian ini tidak dapat menunjukkan validitas instrumen, karena validitas sistem penskoran otomatis tidak dapat dicerminkan dari hasil uji reliabilitas. Reliabilitas antar penskor manusia dan mesin otomatis menjadi indikator tingkat validitas skor yang diberikan oleh penskor otomatis (Powers, dkk., 2015). Sementara menurut Chung & Baker, (2003), keandalan atau reliabilitas yang tinggi adalah kondisi yang diperlukan tetapi tidak cukup untuk memenuhi validasi penilaian yang dilakukan. Validasi penskoran menggunakan program komputer umumnya tidak dilakukan secara terhadap skor yang dihasilkan, tetapi melalui proses pada perangkat yang lain, yaitu rubrik. Proses penyiapan dalam penggunaan UKARA adalah bentuk validasi secara teknis, yaitu menginput jawaban-jawaban yang telah diberikan oleh penskor manusia ke dalam sistem komputer. Validitas berada dalam skor yang diberikan oleh penskor manusia tersebut. Chung dan Baker, (2003) juga menjelaskan bahwa validasi terhadap mesin adalah validasi terhadap produk yang dihasilkan sesuai dengan tujuan atau konteksnya. Validasi untuk UKARA secara sistem sebagai pencapaian tujuannya tentunya telah dilakukan pada sistem kecerdasan buaatannya. Algoritma yang digunakan dianggap telah valid sebagai suatu penskor otomatis. Sehingga dalam penelitian ini tidak dilakukan validasi terhadap UKARA sebagai penskor otomatis.

Kemajuan dalam penilaian otomatis memiliki implikasi signifikan bagi proses pengukuran pengetahuan. Saat ini penilaian pilihan ganda banyak digunakan di ruang kelas sains dan pengukuran kompetensi guru, yang sebagian alasannya adalah tantangan bagi guru dan instruktur untuk menghabiskan waktu dalam membuat skor pertanyaan terbuka. Meskipun butir soal pilihan ganda memiliki kelebihan sendiri, dalam proses asesmen untuk guru diperlukan pertanyaan terbuka yang dapat menangkap kemampuan analisis, penalaran dan argumentasi yang mendalam (Nehm & Haertig, 2012). Alat penilaian otomatis dengan keakuratan yang terbukti juga akan dapat membantu meningkatkan penggunaan penilaian terbuka (Liu, dkk., 2016). Implikasi penting lainnya dari penilaian otomatis adalah umpan balik yang diberikan kepada pihak berkepentingan dalam asesmen dapat mengarahkan program tindak lanjut yang lebih akurat. Hal ini serupa dengan penjelasan Linn, dkk. (2014), bahwa penggunaan penskor otomatis dapat mencapai tingkat untuk memberikan bimbingan yang sama efektifnya dengan bimbingan guru dalam mendorong siswa untuk meningkatkan hasil belajar. Pada tingkat keseluruhan, hasil penskoran dan masukan otomatis dapat menginformasikan tentang hambatan umum untuk memahami dan memfasilitasi peningkatan kompetensi guru secara efisien.

Kesimpulan

Nilai reliabilitas berupa ICC antara penskor esai otomatis UKARA dan penskor manusia pada penelitian ini termasuk ke dalam kategori tinggi untuk semua butir soal dalam instrumen pengukuran pengetahuan guru. Analisis berdasarkan korelasi Pearson juga menunjukkan bahwa skor yang diberikan oleh UKARA memiliki korelasi yang kuat dengan skor yang diberikan oleh penskor manusia. Dalam hasil analisis juga dijelaskan bahwa hasil tersebut tidak memiliki korelasi yang nyata dengan jenis kalimat jawaban dan komposisi jawaban terkait skor. Oleh karena itu faktor yang mempengaruhi reliabilitas antar penskor dalam penelitian ini terutama adalah kapabilitas sistem penskoran otomatis yang digunakan itu sendiri. Keterbatasan penelitian yang mungkin menjadi penyebab adalah kekurangan dalam jumlah subjek asesmen sehingga uji coba terkait dua faktor tersebut belum optimal.

Daftar Pustaka

- Aalaei, S., Ahmadi, M.A.T., & Aalaei, A. 2016. A Comparison of multiple-choice and essay questions in the evaluation of dental students. *International Journal of Advanced Biotechnology and Research*, 7(5):1674–1680.
- Ahmad, R. 2019. E-learning automated essay scoring system menggunakan metode searching text similarity matching text. *Jurnal Penelitian Enjiniring*, 22(1):38–43.
- Aji, R.B., Baizal, A., & Firdaus, Y. 2011. Automatic essay grading system menggunakan metode latent semantic analysis E-78 E-79. *Prosiding Seminar Nasional Aplikasi Teknologi Informasi*, 2011(Snati): 1–9.
- Akturk, A.O. & Ozturk, H.S. 2019. Teachers' TPACK levels and students' self-efficacy as predictors of students' academic achievement. *International Journal of Research in Education and Science*, 5(1):283–294.
- Attali, Y. & Burstein, J. 2006. Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from <http://www.jtla.org>

- Sevcikova, B.L. 2018. Human versus automated essay scoring: a critical review. *Arab World English Journal*, 9(2):157–174.
- Brew, C, & Leacock, C. 2013. *Automated Shor Answer Scoring: Principles and Prospects. In Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge. pp. 136–152.
- Byrne, R., Tang, M., Tranduc, J., & Tang, M. 2010. E Grader, a software application that automatically scores student essays: With a postscript on ethical complexities. *Journal of Systemics, Cybernetics and Informatics*, 8(6):30–35.
- Chung, G.K.W.K. & Baker, E.L. 2003. *Issues in the reliability and validity of automated scoring of constructed responses*, dalam Shermis, M.D. & Burnstein, J. (Eds.), *Automated essay scoring: A cross disciplinary approach* (pp. 23–40). Lawrence Erlbaum Associates.
- Clauser, B.E., Kane, M.T., & Swanson, D. B. 2002. Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4):413–432.
- Cohen, Y. 2017. Estimating the Intra-Rater Reliability of Essay Raters. *Frontiers in Education*, 2(September): 1–11.
- Crossley, S.A., Allen, L.K., Snow, E.L., & Mcnamara, D.S. 2016. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining*, 8(2):1–19.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. 2012. Evaluating teacher evaluation. *Phi Delta Kappa*. 9(3):8–15.
- Dawis, R.V. 1987. Scale construction. *Journal of Counseling Psychology*, 34(4):481–489.
- Ebel, R.L. & Frisbie, D.A. 1991. *Essentials of educational measurement* (5th ed.).
- Etikan, I. 2016. Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1):1-7.
- Finch, W.H. & French, B.F. 2018. Educational and Psychological Measurement. Routledge, New York and London.
- Fleenor, J.W., Fleenor, J.B., & Grossnickle, W.F. 1996. Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology*, 10(3):367–380.
- Graham, M., Milanowski, A., & Miller, J. 2012. *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Inter-Rater Agreement of Teacher and Principal*. United States: Center for Educator Compensation Reform.
- Ha, M., Nehm, R.H., Urban-Lurain, M., & Merrill, J.E. 2011. Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, 10(4):379–393.
- Haley, D.T., Thomas, P., De Roeck, A., & Petre, M. 2007. Measuring improvement in latent semantic analysis-based marking systems: Using a computer to mark questions

about HTML. *Proceedings of the Ninth Australasian Conference on Computing Education*, 66(January):35–42.

Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., & Pribadi, F.S. 2019. A scoring rubric for automatic short answer grading system. *Telkomnika (Telecommunication Computing Electronics and Control)*, 17(2):763–770.

Horbach, A. & Zesch, T. 2019. The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, 4(April).

Ke, Z. & Ng, V. 2019. Automated essay scoring: A survey of the state of the art. *IJCAI International Joint Conference on Artificial Intelligence*, 2019-Augus, 6300–6308. <https://doi.org/10.24963/ijcai.2019/879>

Keith, T.Z. 2003. *Validity and Automated Essay Scoring Systems*, dalam Shermis, M.D & Burstein, J. (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates. pp.147–168.

Khan, M.A. & Ishfaq, U. 2016. Test-score reliability of essay questions in BA examination. *Global Language Review*, 1(1):58–65.

Kuntarto, E., Nurhayat, W.I, Handayani, H., Trianto, A., & Maryono, M. 2019. Teacher'S competency assessment (TCA) in Indonesia: a New Frame Work. *International Conference on Educational Assessment and Policy (ICEAP 2019)*, 8–20.

Kuo, C.Y. & Wu, H.K. 2013. Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers and Education*, 68:388–403.

Landauer, T.K., Laham, D., & Folt, P. 2003. Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice*, 10(3):295–308.

Latifi, S. & Gierl, M. 2021. Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1):62–85.

Linn, M.C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O.L., & Rafferty, A.N. 2014. Computer-guided inquiry to improve science learning. *Science*, 344(6180):155–156.

Liu, O L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. 2014. Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2):19–28.

Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., & Linn, M.C. 2016. Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, 53(2):215–233.

Luckin, R. 2017. Towards artificial intelligence-based assessment systems. *Nature Human Behaviour*, 1(3):725–731.

Nehm, R.H., & Haertig, H. 2012. Human vs. Computer Diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, 21(1):56–73.

- Nkhoma, C., Nkhoma, M., Thomas, S., & Quoc Le, N. 2020. The Role of Rubrics in Learning and Implementation of Authentic Assessment: A Literature Review. *Proceedings of the 2020 InSITE Conference*, (October):237–276.
- Powers, D.E., Escoffery, D.S., & Duchnowski, M.P. 2015. Validating automated essay scoring: a (modest) refinement of the "gold standard." *Applied Measurement in Education*, 28(2):130–142.
- Ramalingam, V.V., Pandian, A., Chetry, P., & Nigam, H. 2018. Automated essay grading using machine learning algorithm. *Journal of Physics: Conference Series*, 1000(1).
- Rios, J.A. & Wang, T. 2018. *Essay Items*, dalam Frey, B.B. (Ed.), *The SAGE Encyclopedia of Educational, Measurement, and Evaluation*. Thousand Oak CA: SAGE Publications, Inc. pp.602–605.
- Shermis, M.D. & Burstein, J. 2003. *Introduction*, dalam Shermis, M.D. & Burnstein, J. (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah NJ: Lawrence Erlbaum Associates. pp. xiii-xvi.
- Shermis, M.D., Burstein, J., & Bursky, S.A. 2013. *Introduction to Automated Essay Evaluation*, dalam Shermis, M.D. & Burnstein, J. (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge/Taylor & Francis. pp.1–15.
- Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. 2010. *Automated essay scoring: Writing assessment and instruction*. International Encyclopedia of Education, Vol. 4, pp. 20–26.
- Shin, J. & Gierl, M.J. 2020. More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, 1–26.
- Sukardi, R.R., Widodo, A., & Sopandi, W. 2017. Describing teachers' pedagogic content knowledge about reasoning development and students' reasoning test. *Advances in Social Science, Education and Humanities Research (ASSEHR)*. ICMSEd 2016:14–20.
- Taghipour, K. & Ng, H.T. 2016. A Neural Approach to Automated Essay Scoring. *Proceedings of the 2016 conference on empirical methods in natural language processing*. pp.1882–1891.
- Toranj, S. & Ansari, D.N. 2012. Automated versus human essay scoring: A comparative study. *Theory and Practice in Language Studies*, 2(4):719–725.
- Vacha-Haase, T., Kogan, L.R., & Thompson, B. 2000. Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, 60(4):509–522.
- Valenti, S., Neri, F., & Cucchiarelli, A. 2003. An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education: Research*, 2:319–330.
- van Helvoort, J., Brand-Gruwel, S., Huysmans, F., & Sjoer, E. 2017. Reliability and validity test of a Scoring Rubric for Information Literacy. *Journal of Documentation*, 73(2): 305–316.

- Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., & Horbach, A. 2020. Automated scoring of teachers' pedagogical content knowledge – a comparison between human and machine scoring. *Frontiers in Education*, 5(August):1–10.
- Williamson, D., Bennett, R., Lazer, S., Bernstein, J., Foltz, P., Landauer, T., ... Sweeney, K. 2010. *Automated Scoring for the Assessment of Common Core Standards*. Working Paper. Educational Testing Service.
- Zhang, M. 2013. Contrasting automated and human scoring of essays. *ETS R & D Connections*, (21):1–11.
- Zimmerman, W.A., Kang, H. Bin, Kim, K., Gao, M., Johnson, G., Clariana, R., & Zhang, F. 2018. Computer-automated approach for scoring short essays in an introductory statistics course. *Journal of Statistics Education*, 26(1):40–47.
- Zlatkin-Troitschanskaia, O., Kuhn, C., Brückner, S., & Leighton, J.P. 2019. Evaluating a technology-based assessment (tba) to measure teachers' action-related and reflective skills. *International Journal of Testing*, 19(2):148–171.