ANALISIS KORELASI ANTAR INSTRUMEN DAN PENGGUNAAN TEKNOLOGI PENSKORAN DALAM PENGUKURAN PENGETAHUAN GURU ILMU PENGETAHUAN ALAM

Raden Ahmad Hadian Adhy Permana[1], Ari Widodo[2]

[1]Widyaiswara LPMP Banten/Mahasiswa Sekolah Pasca Sarjana UPI
[2]Program Studi Pendidikan IPA, Sekolah Pasca Sarjana Universitas Pendidikan Indonesia
Email: radenahmadhadian@upi.edu

ABSTRAK

Kelemahan dari pertanyaan pilihan ganda bukan validitas atau reliabilitas instrumennya, tetapi tingkat keaslian jawaban yang diberikan. Diperlukan instrumen lain yang otentisitas hasilnya lebih baik untuk mengukur pengetahuan guru. Tujuan penelitian untuk mengetahui hubungan antara hasil UKG 2015 dan hasil tes menggunakan instrumen esai. Tujuan kedua mengetahui reliabilitas skor oleh penilai dan menggunakan program komputer. Penelitian melibatkan 30 guru IPA SMP di Provinsi Banten sebagai sampel, guru-guru tersebut adalah guru yang telah mengikuti UKG pada tahun 2015. Metode penelitian adalah penelitian kuantitatif dengan menerapkan convenience sampling. Pengumpulan data dilakukan langsung untuk instrumen tes esai dan data hasil UKG diambil dari dokumen yang ada di LPMP Banten. Peskoran dilakukan oleh peneliti dan perangkat lunak UKARA yang dimiliki oleh Pusat Asesmen dan Pembelajaran Kemdikbud. Analisis korelasi berdasarkan koefisien korelasi produk-momen Pearson dan reliabilitas penskor berdasarkan intraclass coefisien correlation (ICC). Hasil perhitungan menunjukkan korelasi positif yang kuat antara kedua hasil tes dengan nilai r = 0,61 dan tingkat signifikansi 0,01. Hasil tersebut menunjukkan bahwa guru yang mendapat skor tinggi di UKG cenderung menjawab pertanyaan esai lebih baik daripada guru yang mendapat skor UKG lebih rendah. Hasil perhitungan ICC antara penilai orang dan perangkat lunak UKARA adalah 0,62. Ini menunjukkan bahwa reliabilitas antar penskor masih di bawah level yang dianggap andal. Penelitian ini menunjukkan bahwa pertanyaan esai memiliki potensi sebagai instrumen untuk menguji pengetahuan guru dan berpotensi untuk diterapkan pada sejumlah besar peserta sama seperti instrumen pilihan ganda.

Kata kunci: instrumen esai; penskoran esai berbantuan komputer; korelasi antar instrumen; reliabilitas antar penskor

LATAR BELAKANG

Pengembangan instrumen tes pengetahuan saat ini umumnya masih terfokus hanya kepada satu bentuk instrumen, yaitu pilihan ganda. Intrumen tersebut bukan berarti tidak valid atau tidak reliabel, tetapi pada beberapa penelitian masih dikemukakan bahwa instrumen pilihan ganda memiliki nilai otentisitas yang kurang baik (Kastner dan Stangla, 2011). Peserta tes masih selalu dapat menjawab benar pilihan ganda dengan hanya menebak dan bukan karena pengetahuan yang mereka miliki.

Tes skala besar, seperti uji kompetensi guru, masih dibutuhkan dan tidak semua instrumen penilaian dapat menjangkau hal tersebut secara efisien. Efisiensi dalam pelaksanaan Uji Kompetensi Guru (UKG) 2015 ditunjukkan dengan penggunaan media berbantuan komputer. Tes tersebut bertujuan untuk mengetahui peta penguasaan guru dalam kompetensi pedagogik dan kompetensi profesional. Peta penguasaan kompetensi tersebut akan dijadikan acuan untuk pertimbangan dalam program pembinaan dan pengembangan profesi Guru (Sofiah, dkk., 2016). Sesuai dengan tujuan tersebut maka tes harus

dilakukan dalam skala nasional untuk seluruh guru di Indonesia dalam kriteria tertentu. Jika otentisitas menjadi perhatian, seperti efisiensinya, maka tes tersebut perlu ditingkatkan kualitas hasilnya menggunakan instrumen yang relevan.

Pengembangan instrumen esai untuk tes skala besar telah dimulai sejalan dengan berkembangnya sistem penskoran berbantuan komputer (Clauser, Kane, & Swanson, 2002; Aji, dkk., 2011). Potensi tes esai sebagai alternatif telah dikaji dan diterapkan di beberapa negara. Pengembangan tes esai ini dilatarbelakangi oleh karakternya yang dianggap lebih otentik, karena peserta tes tidak akan mudah menjawab dengan cara menebak (Rios & Wang, 2018). Jika berusaha menebak pun, lebih mudah diketahui, karena jawabannya akan menyimpang dari rubrik yang diharapkan. Soal uraian yang dimaksud disini bukan soal isian singkat yang bersifat objektif, tetapi soal yang memerlukan jawaban dalam bentuk kalimat bahkan beberapa kalimat sehingga termasuk ke dalam instrumen subjektif walaupun terbatas.

Kendala dalam penerapan soal uraian untuk peserta tes dalam jumlah yang banyak terkait dengan efisiensi dan konsistensi penskor atau penilai. Umumnya jawaban uraian peserta tes diberi skor oleh seorang rater atau beberapa rater. Jika jumlah peserta ratusan bahkan ribuan, tentunya diperlukan banyak rater. Di sisi lain kemampuan rater juga berbeda-beda dan akan menghasilkan skor yang tidak konsisten (Powers, Escoffery & Duchnowski, 2015). Penggunaan rubrik tentunya menjadi keharusan, tetapi beberapa penelitian menunjukkan masih selalu ada inkonsistensi dari orang yang menjadi penilai tersebut. Maka kedua hal tersebut menjadi pertimbangan utama dalam penerapan intrumen uraian dalam tes massal.

Maka solusinya saat ini adalah penggunaan penskor berupa perangkat lunak komputer. Seperti juga pada intrumen soal berbentuk pilihan ganda, perangkat komputer menjadi mesin untuk memberikan skor terhadap jawaban peserta secara efisien. Program komputer untuk esai tersebut tidak hanya yang sederhana tetapi bahkan *artificial intelegent* (AI) atau kecerdasan artifisial.

Permasalahan yang dikaji pada penelitian ini meliputi 2 hal, yaitu: (1) Bagaimana korelasi antara nilai UKG (2015) dan hasil tes uraian guru IPA SMP?; (2) Bagaimana reliabilitas antara hasil penskoran oleh penilai dan program komputer?

Permasalahan tersebut kami anggap dapat menghasilkan kebaruan informasi yaitu terkait dengan pengembangan instrumen alternatif dan hasil uji coba penggunaan perangkat lunak sebagai penskor esainya.

Tujuan yang ingin dicapai adalah mengetahui hubungan antara hasil UKG 2015 dan hasil tes menggunakan instrumen esai dan mengetahui reliabilitas skor oleh penilai orang dan menggunakan program komputer. Kedua tujuan khusus tersebut akan membawa ke arah yang lebih luas, yaitu kriteria asesmen yang valid, reliabel, lebih otentik, dan efisien untuk mengukur pengetahuan guru.

METODE PENELITIAN

Metode penelitian yang kami lakukan untuk mencapai tujuan tersebut adalah penelitian kuantitatif. Populasi adalah jumlah guru IPA di Provinsi Banten, dan sampel diambil secara *convenience sampling* (Gall, Gall & Borg, 2003). Diperoleh 30 sampel dengan status guru IPA SMP di 5 kab/kota di Banten yang sebelumnya telah mengikuti UKG tahun 2005.

Instrumen yang digunakan pada penelitian ini adalah instumen soal esai dan

rubrik penskoran yang disusun berdasarkan kisi-kisi UKG 2015. Instrumen tambahan berupa kuisioner uji keterbacaan pertanyaan pada instrumen soal esai. Instrumen soal esai divalidasi isinya oleh ahli, dan diuji secara standar statisitik.

Analisis data secara statistik untuk menghitung koefisien korelasi Pearson Product-Moment sebagai analisis data pertama (Gilchrist & Samuels, 2014) dan menghitung Intraclass Coefisien Corelation (ICC) untuk data kedua. ICC dipilih karena data skor berupa data rentang sehingga dianggap lebih relevan dibandingkan analisis tipe yang lain untuk interrater (Bartko, 1966; Akhtar, 2018). Analisis data secara statisitik menggunakan program SPSS.

Data pada penelitian terbagi menjadi data yang diambil secara langsung dan data dari dokumen. Data untuk analisis pertama terbagi 2, yaitu data UKG 2015 yang kami ambil dari dokumen di LPMP Banten (2016) dan data tes uraian yang kami ambil langsung dari para guru di sekolah menggunakan instrumen tertulis. Data untuk analisis kedua adalah hasil penskoran terhadap tes esai dari 30 orang guru menggunakan 2 cara, penskoran oleh penilai orang dan menggunakan perangkat lunak komputer. Penilai (orang) adalah peneliti sendiri menggunakan rubrik penskoran yang telah disusun dan divalidasi untuk isi dan pemeringkatannya. Penskoran berbantuan komputer menggunakan aplikasi UKARA miliki Pusat Asesmen dan Pembelajaran Kementerian Pendidikan dan Kebudayaan. Aplikasi atau perangkat komputer tersebut merupakan penskor yang dapat memberikan skor terhadap jawaban dikotomi maupun politomi. Untuk penelitian ini instrumen penskoran yang dikembangkan bersifat politomi. Model pengklasifikasian jawaban yang digunakan pada perangkat lunak tersebut dipilih Adaboost Classifier (Schapire, 2013). UKARA sendiri memiliki beberapa pilihan model algoritma untuk klasifikasi jawaban atau prediksi. Program komputer tersebut bekerja dengan sistem kecerdasan artifisial. Perangkat lunak komputer ini bukan perangkat yang dapat diakses secara umum atau tersedia terbuka dalam jaringan internet secara luas, tetapi perlu ijin dari pengelolanya untuk dapat menggunakannya secara terbatas.

Tahapan kegiatan penelitian ini adalah: (1) kajian teori; (2) penyusunan instrumen; (3) pengambilan data; (4) pengolahan dan analisis data (tahap 1 dan 2); dan (5) penyusunan laporan.

HASIL DAN PEMBAHASAN

Instrumen Soal Uraian

Instrumen dalam penelitian ini dibuat berdasarkan kisi-kisi soal UKG IPA 2015. Penggunaan kisi-kisi UKG tentunya dengan alasan bahwa soal esai akan setara dengan soal UKG yang berbentuk pilihan ganda. Jika dibuat sendiri kisi-kisinya maka instrumen esai tersebut perlu diuji coba lebih menyeluruh. Soal UKG, termasuk kisi-kisinya, dianggap telah valid dan reliabel sebagai suatu instrumen untuk menguji pengetahuan guru.

Instrumen yang dihasilkan terdiri dari 15 nomor soal yang mencakup 37 butir soal. Konstruksi soal terdiri dari stimulus berupa studi kasus atau data dan beberapa rumusan pertanyaan (pokok soal) pada setiap nomor soal. Berdasarkan uji coba awal, instrumen tersebut dapat diselesaikan dalam waktu sekitar 60 menit. Jangka waktu tersebut dianggap relevan dengan jangka waktu UKG dan jangka waktu tes pada umumnya. Keterbacaan soal juga ditanyakan kepada para responden dalam kuisioner yang diberikan setelah soal dikerjakan. Seluruh responden menyatakan dapat memahami setiap pertanyaan dalam soal yang diberikan.

Instrumen yang disusun mencakup rubrik penskoran yang dibuat secara bertingkat menggunakan rentang skor 0 – 2

atau 0 – 3 tergantung kompleksitas jawaban yang diharapkan. Model penskoran yang digunakan dengan pendekatan berdasarkan Teori Respon Butir untuk politomi (Ostini & Neuring, 2006; Retnawati, 2017). Pilihan jenis rubrik penskoran ini karena berdasarkan kisi-kisi soalnya akan dapat diperoleh jawaban-jawaban yang relevan dengan penskoran bertingkat, dibandingkan penskoran parsial ataupun benar-salah (dikotomi). Instrumen penskoran ini tetap memiliki batasan jawaban sesuai konsep atau prosedur dalam pertanyaannya. Artinya rubrik yang telah disusun tidak bersifat subjektif atau terbuka sepenuhnya tetapi memiliki panduan yang jelas sebagai batasan.

Pengetahuan yang diujikan mencakup kemampuan berpikir pemahaman, aplikasi, analisis, dan evaluasi terhadap substansi pelajaran IPA SMP secara komprehensif. Cakupan pengerahuan ini relevan dengan bentuk soal uraian yang dijadikan intrumen utama penelitian ini. Tingkatan pemahaman sampai dengan evaluasi dapat diukur dengan instrumen penskoran yang bertingkat dan jawaban masih dapat dibatasi. Berbeda jika pengetahuan tingkat kreasi yang diujikan, maka instrumen penskoran yang dibuat akan terbuka atau subjektif sepenuhnya.

Perbandingan Hasil Tes

Hasil tes esai dan UKG untuk 30 responden secara ringkas dapat dilihat pada tabel 1 di bawah ini.

Tabel 1. Perbandingan Nilai Antara 2 Tes

	N	*Min.*	*Max.*	*Mean*	*Std. Deviation*
UKG	30	39.68	82.20	62.85	12.75
Esai	30	11.00	65.00	41.30	14.72

Perbandingan antara hasil UKG IPA 2015 dan hasil tes esai seperti ditunjukkan dalam tabel 1, bahwa secara rata-rata hasil UKG (62,85) lebih tinggi daripada hasil tes esai (41,30). Data rinci juga menunjukkan bahwa tidak ada responden yang nilai tes esainya lebih tinggi daripada nilai UKG. Perbedaan (penurunan) nilai berkisar dari 0,85 – 47,03. Umumnya hasil tes pilihan ganda lebih tinggi daripada hasil tes uraian untuk soal yang serupa (Aalaei, Ahmadi & Aalaei, 2016). Hasil ini tidak sepenuhnya menyatakan bahwa soal pilihan ganda lebih mudah daripada esai, tetapi ada beberapa faktor yang mempengaruhi. Jika berdasarkan tingkat kesulitan maka relatif sama, karena soal dikembangkan dari kisi-kisi yang sama. Faktor yang dapat mempengaruhi perbedaan tersebut antara lain kesiapan dan keseriusan peserta dalam mengerjakan tes. UKG adalah tes formal yang hasilnya akan dipakai untuk berbagai kepentingan dan dapat mempengaruhi status seorang guru. Sehingga kebanyakan guru akan mempersiapkan diri dengan serius dan mengerjakan soal juga secara serius. Sementara tes pada penelitian ini hasilnya tidak akan mempengaruhi status pekerjaan para guru tersebut, sehingga kecenderungan kurang serius dalam mengerjakan sangat besar, jadi hanya dikerjakan sebisanya, tanpa persiapan. Hal ini telah menjadi pertimbangan, karena walaupun soal-soal tersebut adalah materi yang umumnya diajarkan oleh para guru secara substansi, tetapi di sisi lain ada pengetahuan pedagogi yang secara konsep masih belum dikuasai dengan baik oleh para guru. Jadi dapat dikatakan soal pada penelitian ini tidak lebih sulit secara konstruksi, tetapi lebih sulit ketika harus dikerjakan karena faktor persiapan dan konteks tes.

Peserta yang memperoleh nilai tertinggi pada UKG dalam kelompok respoden ini (dengan nilai 82,20) ternyata menjadi salah satu yang mendapatkan nilai tertinggi pula pada nilai tes esai (65,00). Tetapi responden yang memiliki nilai terendah pada tes esai bukanlah responden yang mendapat nilai UKG terendah pada

kelompok ini. Bahkan responden yang mendapat nilai terendah (11,00) pada tes esai, mengalami perbedaan nilai cukup tinggi (nilai UKG 39,68) dan bukan nilai UKG terendah pada kelompok ini. Indikasi ini menujukkan bahwa penjelasan di atas sangat terkait, kemungkinan sebagian peserta tidak serius mengerjakan soal esai dibanding saat kegiatan UKG 2015.

Jika asumsi bahwa otentisitas hasil tes dengan instrumen esai lebih baik daripada hasil pilihan ganda, maka hasil penelitian ini menggambarkan bahwa nilai asli guru-guru responden lebih rendah daripada nilai yang diperoleh dari UKG. Masalahnya adalah nilai UKG guru IPA dan secara keseluruhan masih berada di bawah harapan (Sofiah, dkk., 2016). Artinya jika digunakan soal berupa uraian maka akan menunjukkan nilai yang cenderung lebih rendah lagi. Secara konsisten data yang diperoleh menunjukkan bahwa tidak ada satupun responden yang nilai tes esainya lebih baik daripada nilai UKG.

Keterbasan penelitian ini, dalam jumlah responden, membuat hasil ini tidak terlalu kuat untuk digeneralisir, tetapi hasil ini menunjukkan bahwa soal dalam bentuk esai akan selalu dianggap lebih sulit untuk dikerjakan dibandingkan tes pilihan ganda. Kelebihan tes esai adalah peserta tes tidak akan mendapat petunjuk sedikitpun mengenai jawaban yang diharapkan, sementara soal pilihan ganda akan selalu memberikan pilihan yang salah satunya adalah jawaban benar. Walaupun peserta tidak mencoba menebak, tetapi kalimat pilihan dapat memberikan petunjuk pada memori peserta mengenai jawaban yang tepat. Kelebihan lain menurut Wahyuni, dkk. (2015), ketika menjawab soal uraian peserta tes dapat dinilai mengenai kreativitas, kemampuan menganalisis dan mensintesis suatu persoalan.

Korelasi Antara Hasil Tes Esai dan UKG

Hasil perhitungan SPSS seperti pada tabel 2 di bawah ini.

Tabel 2. Korelasi Hasil UKG dan Tes Esai

		UKG	Esai
UKG	Pearson Correlation	1	.610**
	Sig. (2-tailed)		.000
	N	30	30
Esai	Pearson Correlation	.610**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. *Correlation is significant at the 0.01 level*

Koefisien korelasi Product-Moment (Pearson) yang diperoleh adalah korelasi positif dengan r = 0,61 pada derajat signifikansi 0,01. Berdasarkan penjelasan menurut Gall, Gall & Borg (2003) nilai koefisien tersebut termasuk dalam tingkat korelasi yang kuat. Maka nilai tes esai responden penelitian ini memiliki korelasi positif yang kuat terhadap nilai UKG.

Menurut Schober & Schwarte (2018), korelasi digunakan pada konteks dari suatu hubungan linier antara dua variabel yang kontinu, dalam hal ini hubungan antara hasil UKG dan tes esai. Hubungan tersebut mengindikasikan bahwa terdapat kecenderungan responden yang memiliki nilai UKG lebih tinggi akan memiliki nilai lebih tinggi pada tes esai. Sebaliknya yang memiliki nilai UKG lebih rendah akan mendapat nilai lebih rendah pula pada tes esai. Hubungan tersebut terlihat dari pola yang terbentuk ketika uji korelasi dilakukan.

Tentunya masih ada syarat-syarat lain pula yang harus dipenuhi sehingga instrumen tes esai dapat sepenuhnya menggantikan tes pilihan ganda. Syarat-syarat dasar seperti validitas, reliabilitas, dan daya pembeda harus tetap diuji sehingga untuk memenuhi kriteria suatu tes yang berkualitas tinggi. Satu hal penting lagi adalah efisiensi dalam pelaksanaan untuk tes yang dilakukan dalam skala besar seperti

UKG. Menurut Carlson (1990), dalam pengembangan tes pilihan ganda untuk menilai pengetahuan konten pedagogis, kita harus memperhatikan banyak isu, antara lain tujuan tes, keterpaduan antara pengetahuan pedagogi dan konten, serta validitas jawaban. Prinsip tersebut juga akan menjadi prinsip yang dapat diterapkan dalam penyusunan soal esai sebagai alternatif instrumen yang setara. Pemangku kebijakan perlu memperhatikan hasil penelitian-penelitian yang relevan, seperti pada penelitian ini, untuk menjadikan dasar penerapan instrumen tes alternatif.

Reliabilitas Antar Penskor

Kendala utama bagi soal uraian adalah efisiensi. Penskoran esai umumnya melibatkan seorang atau beberapa penskor dan hal ini perlu waktu. Menurut Aalaei, dkk. (2016), ujian dengan bentuk soal uraian memerlukan waktu yang banyak untuk proses koreksi karena peserta ujian menunjukkan pengetahuannya dalam format uraian. Sementara untuk tes pilihan ganda, penskor ini sepenuhnya telah dapat digantikan oleh perangkat komputer yang dapat memberikan skor 1 atau 0 sesuai kunci jawaban. Untuk jawaban uraian perlu diperhatikan bahwa skor yang diberikan tidak hanya 0 dan 1 atau benar – salah, tetapi ada penjenjangan skor karena kompleksitas jawaban yang diharapkan. Untuk penelitian ini, penskoran tidak mengalami kesulitan berarti karena jumlah responden hanya 30 orang atau skala kecil.

Untuk mengetahui kemampuan sistem komputer yang dapat membantu penskoran jawaban esai maka perangkat yang digunakan perlu diuji lebih dahulu. Pengujian utama adalah menghitung *Intraclass Correlation Coeficient* (ICC). Berdasarkan Powers, Escoffery & Duchnowski (2015), membandingkan hasil penskor orang dengan hasil komputer adalah cara yang paling sering digunakan untuk memvalidasi. Sementara Wuensch, (2014) menyatakan bahwa menghitung ICC adalah salah satu solusi untuk mengetahui kehandalan sistem penskoran oleh komputer.

Hasil perhitungan ICC pada penelitian ini dapat dilihat pada tabel 3 di bawah ini. Penskor 1 adalah peneliti sendiri dan penskor 2 adalah program komputer (UKARA).

Tabel 3. ICC Antara Penskor 1 dan 2

	Intraclass Correlation	*95% Confidence Interval*	
		Lower	*Upper*
Single Measures	.62	.34	.80
Average Measures	77	.51	.89

Berdasarkan hasil perhitungan menggunakan SPPS, diperoleh nilai 0,62 (tunggal) antara penskor orang (peskor 1) dan UKARA (penskor 2). Hasil tersebut menunjukkan belum mencapai nilai yang dianggap reliabel atau handal. Menurut Polgar dan Thomas (2000, dalam Harry, 2015) dan Akhtar (2018), alat ukur memiliki stabilitas memadai jika ICC antar pengukuran > 0.50, stabilitas tinggi jika ICC antar pengukuran ≥ 0.80. Untuk tes atau pemetaan diperlukan stabilitas atau kehandalan yang tinggi sehingga instrumen dapat dipercaya hasilnya.

Hasil ini mengindikasikan bahwa masih ada kelemahan pada sistem yang digunakan. Jika dianggap bahwa skor yang diberikan oleh penskor 1 telah valid dan konsisten, maka skor yang diberikan oleh komputer juga seharusnya relatif sama. ICC yang rendah menunjukkan bahwa skor yang diberikan komputer masih memiliki pola yang tidak mirip (similar) dengan pola skor yang diberikan penilai. Perbedaan tersebut dipengaruhi oleh beberapa faktor, antara lain kapasitas algoritma yang digunakan, konsistensi skor, dan jumlah responden yang

menyebabkan perangkat lunak tersebut tidak bekerja secara optimal. Akurasi pemberian skor otomatis oleh program komputer juga tergantung dari beberapa faktor, termasuk di dalamnya adalah domain konten, kompleksitas tugas, tingkatan dalam rubrik pensokran, dan jumlah dari respon yang tersedia untuk membangun sistem skor otomatis (Liu, dkk., 2016).

Algoritma AdaBoost yang digunakan pada UKARA tentunya memiliki kapasitas tertentu. Boosting adalah suatu pendekatan terhadap cara belajar mesin berdasarkan ide dari penciptaan aturan prediksi yang berakurasi tinggi dengan cara mengkombinasikan banyak aturan yang relatif lemah dan kurang akurat (Schapire, 2013). Sistem ini bekerja dengan komputer yang menerapkan AI (kecerdasan artifisial), maka perlu "pelatihan" yang memadai sehingga sistem dapat bekerja optimal. Pelatihan yang menjadi pola kerja sistem komputer yang digunakan pada penelitian ini adalah membuat data dasar penskoran dari jawaban responden dan pemberian skor oleh penilai. Berdasarkan data dasar tersebut, sistem kecerdasan artifisial akan bekerja membuat pola skor sendiri sesuai algoritma yang digunakan (Adaboost Classifier). Pada dasarnya program komputer ini memerlukan input data sebanyak mungkin untuk membuat data dasar yang berkualitas tinggi. Sementara pada penelitian ini hanya diberikan 30 jawaban peserta, tentunya sistem akan memiliki data yang kurang optimal.

Walaupun hasil penelitian ini belum menunjukkan bahwa perangkat komputer dapat dijadikan penskor yang handal, tetapi potensi penggunaannya dapat dianggap tetap tinggi. Kehandalam memang masih harus ditingkatkan, tetapi efisiensi sudah terlihat antara lain efisiensi waktu. Perbaikan pada beberapa aspek akan dapat membuat perangkat lunak ini menunjukkan kualitasnya dan dapat digunakan pada penskoran skala luas.

KESIMPULAN

Pertama, dari hasil analisis terhadap 2 instrumen yang berbeda, Guru yang mendapat skor tinggi di UKG cenderung menjawab pertanyaan esai lebih baik daripada guru yang mendapat skor UKG lebih rendah dan berlaku juga sebaliknya. Hal ini mengindikasikan hasil soal esai yang dibuat dianggap setara dengan soal pilihan ganda UKG secara statistik dan dianggap dapat mengukur pengetahuan guru

Kedua, reliabilitas penskoran oleh perangkat komputer yang dikorelasikan dengan peskoran oleh penilai orang masih di bawah level yang dianggap handal. Hal ini menunjukkan alat penskor memiliki kelemahan, tetapi bukan berarti lemah secara sistem AI, kelemahan karena penerapan sistemnya belum ideal, seperti jumlah responden dan konsistensi penksoran oleh orang sebagai basis data sistem AI.

Penelitian ini menunjukkan bahwa pertanyaan esai memiliki potensi sebagai instrumen untuk menguji pengetahuan guru dan berpotensi untuk diterapkan pada sejumlah besar peserta sama seperti instrumen pilihan ganda. Penelitian lebih lanjut secara spesifik dan mendalam mengenai penggunaan teknologi untuk penskoran atau asesmen dapat menambah informasi yang dibutuhkan pada masa yang akan datang.

DAFTAR PUSTAKA

Aalaei, S., Ahmadi, M. A. T. & Aalaei, A. (2016). Comparison of Multiple-Choice and Essay Questions In The Evaluation of Dental Students. *International Journal of Advanced Biotechnology and Research (IJBR)*. Vol. 7, Special Issue-Number 5, pp.1674-1680. http://www.bipublication.com.

Aji P,. R.B., Baizal, Z. K. A. & Firdaus, Y. (2011). Automatic Essay Grading System Menggunakan Metode Latent Semantic Analysis. *Prosiding Seminar Nasional Aplikasi Teknologi Informasi*.

Akhtar, H. (2018). Estimasi Reliabilitas Antar Rater (Interrater Reliability) dengan SPSS. https://www.semestapsikometrika.com/2018/10/estimasi-reliabilitas-antar-rater.html.

Bartko, J.J. (1966). The Intraclass Correlation Coefficient As A Measure of Reliability. *Psychological Reports*, 19, 3-11. Southern Universities Press.

Carlson, R. E. (1990). Assessing Teachers' Pedagogical Content Knowledge: Item Development Issues. *Journal of Personnel Evaluation in Education*, 4:157-173.

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. Lawrence Erlbaum Associates, Inc.

Gall, M.D., Gall, J.P. & Borg, W.R. (2003). *Educational Research*. An Introduction. 7th edition. Pearson Education. pp 175 – 176.

Gilchrist, M. & Samuels, P. (2014). *Pearson Correlation*. Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre.

Harry. (2015). Intraclass Coeficient Corellation. http://research-indonesia.blogspot.com/2015/01/intraclass-correlation-coefficient-icc.html.

Kastner, M. & Stangla, B. (2011). Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter? *Procedia Social and Behavioral Sciences*, (12) 263-273.

Liu, O. L., Rios, L.A., Heilman, M., Gerard, L. & Linn, M.C. (2016). Validation of Automated Scoring of Science Assessments. *Journal of research in science teaching*, Vol. 53, No. 2, pp. 215-23. Wiley Periodicals, Inc.

Ostini, R & Neuring, M. (2006). *Polytomous Item Response Theory Models*. Abstract. Sage Publisher. https://www.researchgate.net/publication/37621870_Polytomous_Item_Response_Theory_Models.

Powers, D.A., Escoffery, D.S. & Duchnowski, M.P. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the "Gold Standard. *Applied Measurement in Education*, 28: 130–142. London: Routledge Informa Ltd.

Retnawati, Heri. (2017). *Mengestimasi Kemampuan Peserta Tes Uraian Matematika dengan Pendekatan Teori Respons Butir dengan Penskoran Politomus dengan Generalized Partial Credit Model*. http://staff.uny.ac.id/sites/default/files/132255129/GPCM1_1.pdf.

Rios, J. A. dan Wang, T. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks: SAGE Publications, Inc.

Schapire, R. E. (2013). *Explaining AdaBoost*. http://rob.schapire.net/papers/explaining-adaboost.pdf.

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation.

Anesthesia & Analgesia. www.anesthesia-analgesia.org.

Sofiah, dkk. (2016). *Analisis Gambaran Kompetensi Guru Terhadap Prestasi Belajar Siswa SMP Pada Ujian Nasional Tahun 2015 Provinsi Daerah Istimewa Yogyakarta*. Pusat Data dan Statistik Pendidikan dan Kebudayaan, Kementerian Pendidikan dan Kebudayaan.

Wahyuni, I. T., Yamtinah, S & Budi, T. (2015). Pengembangan Instrumen Pendeteksi Kesulitan Belajar Kimia Kelas X Menggunakan Model Testlet. *Jurnal Pendidikan Kimia*, Vol. 4, No. 4.

Wuensch, K.L. (2014). Inter-Rater Agreement. *InterRater*. East Carolina University, Department of Psychology.