KOMPARASI MODEL PENSKORAN KLASIK PADA ESTIMASI *OBSERVED SCORE* SISWA

Rizki Nor Amelia[1*], Anisa Ratna Nugraini[1], Kriswantoro2, Dian Normalitasari Purnama[3]

[1]Universitas Negeri Semarang, Semarang
[2]Sekolah Tinggi Agama Islam Ma'arif Jambi
[3]Universitas Negeri Yogyakarta, Yogyakarta
*Email korespondensi: rizkinoramelia@mail.unnes.ac.id

ABSTRAK

Masalah skoring masih menjadi isu yang relevan bagi pengukuran dan pengujian pendidikan karena perilaku *guessing* merupakan salah satu faktor yang berkontribusi pada skor yang dihasilkan, padahal idealnya skor tes haruslah sedekat mungkin dengan tingkat penguasaan *testee* yang sebenarnya. Untuk mengatasi hal tesebut, penelitian ini bertujuan untuk mengkomparasikan model penskoran klasik yang berbasis *guessing* (variasi 4 dan 5 pilihan jawaban) dengan model penskoran konvensional (*number right scoring*) terhadap estimasi *observed score* yang dihasilkan. Data yang digunakan adalah data simulasi yang dibangkitkan melalui software WinGen, dengan ukuran sampel 300 responden berdasarkan *Item Response Theory* 1-PL. Hasil analisis menunjukkan bahwa penggunaan model penskoran klasik yang berbeda ternyata tidak memberikan perbedaan yang signifikan dalam hal kesesuaian peringkat *testee*. Selain itu, karakteristik deskriptif dan karakteristik distribusi skornya juga tidak berbeda. Namun demikian, model penskoran berbasis *guessing* tetap dapat dipertimbangkan penggunaannya sebagai upaya pencegahan *guessing* mengingat tidak mungkinnya pembuat soal dalam membedakan jawaban benar berdasarkan penguasaan atau *lucky guessing*.
Kata kunci: klasik; model penskoran; skor.

PENDAHULUAN

Salah satu tujuan utama pengukuran dan pengujian dalam konteks pendidikan adalah memperoleh informasi tentang kemampuan individu yang nantinya dapat digunakan pada berbagai tujuan. Metode yang paling sederhana dan mudah dilakukan untuk mengestimasikan kemampuan siswa adalah dengan memberikan seperangkat butir dan mengukur responnya (Jacob, 2016) sehingga didapatkan suatu ukuran kuantitatif yang umum disebut dengan skor. Skor tersebut menjadi menarik ketika digunakan untuk mendukung klaim yang berupa kemampuan atau kinerja peserta tes pada tugas-tugas tertentu dalam kondisi tertentu (Kane, 2013) karena keakuratannya, keterwakilannya terhadap estimasi kemampuan, hingga penggunaannya dapat mencerminkan dan menjamin kecermatan pengujian yang dilakukan oleh pembuat dan penyelenggara tes (Romanoski & Douglas, 2022).

Pilihan ganda adalah bentuk tes yang umum dimanfaatkan oleh guru dalam melakukan kegiatan pengukuran kemampuan siswa di kelas. Tes ini menjadi favorit karena penilaiannya bersifat obyektif sebagai akibat dari responden yang hanya perlu memilih jawaban paling benar dari daftar pilihan jawaban yang telah disediakan (Yazdi, et.al. 2021) juga karena memberikan banyak kemudahan dalam penilaian, konsistensi penilaian, dan kemampuan untuk mencakup sejumlah besar topik dalam satu kali pengujian (Stankous, 2016). Dengan bentuk tes apapun, skor yang diestimasi dalam pendekatan klasik berupa skor tampak (*observed score*, X) yang merupakan penjumlahan dua komponen independen, yakni skor sesungguhnya (*true score*, T) dan error pengukuran (*random error*, E) yang dinotasikan sebagai : X = T + E (Moses, 2017). Berdasarkan definisi tersebut, maka terdapat beberapa model penskoran klasik yang dapat diaplikasikan untuk melakukan estimasi skor siswa pada tes bentuk pilihan ganda, yakni *number right scoring*, *correction for guessing (rights minus wrongs correction*, dan *correcting raw score* (Crocker & Algina, 2008).

Model penskoran *number right* adalah model penskoran konvensional, dimana skor siswa merupakan penjumlahan dari butir-butir yang direspon secara benar (Lesage, Valcke, & Sabbe, 2013; Lord, 1975). Model ini merupakan model penskoran yang paling tua dan memiliki keunggulan dalam hal ketidakberpihakan, dalam arti tidak ada *testee* yang diuntungkan atau dirugikan oleh faktor-faktor kepribadian yang tidak relevan (Rowley & Traub, 1977). Namun faktanya, ada kekhawatiran bahwa *testee* dapat menjawab benar melalui tebakan (Choppin, 1988), padahal mereka tidak memiliki kemampuan untuk menyelesaikan butir tersebut (Kubinger, et.al, 2010), sehingga justru menyebabkan terjadinya kesalahan acak dan bertambahnya sumber varians pada tes yang berimplikasi pada turunnya validitas dan reliabilitas (Bereby-Meyer, et.al., 2002; Burton, 2001; Lesage, Valcke, & Sabbe, 2013; Prihoda, et.al., 2006).

Model penskoran *correction for guessing (rights minus wrongs correction*) muncul untuk mengatasi masalah yang ditimbulkan akibat *guessing* tersebut. Dengan model penskoran ini, jawaban yang salah dan diduga hasil *guessing* diberikan hukuman berupa pengurangan skor, sehingga model penskoran ini juga dikenal dengan nama *punishment scoring* (Ostrosky, et.al., 2022) atau *negative marking* (Lesage, Valcke, & Sabbe, 2013). Satu lagi model penskoran alternatif yang diusulkan Traub, Hambleton, & Singh (1969) dengan berprinsip pada pendekatan psikologis berupa menghargai perilaku yang diinginkan daripada menghukum perilaku yang tidak diinginkan (Crocker & Algina, 2008) dikenal dengan nama *reward scoring* atau *correcting raw score*. Pada model ini, *testee* diberikan penghargaan karena tidak melakukan *guessing* pada butir yang tidak bisa dijawab, sehingga mereka merasa tidak terancam, dibandingkan dengan menerima hukuman untuk jawaban yang salah (Lesage,

Valcke, & Sabbe, 2013). Prietro & Delgado (1999) juga mendukung model ini sebagai model penskoran terbaik dengan merujuk pada indikator kinerja dan nilai reliabilitas yang dihasilkan.

Penerapan model penskoran yang berbeda tentu dapat menghasilkan estimasi skor yang berbeda. Hal ini dikarenakan *testee* akan mempertimbangkan kemungkinan apakah akan melakukan *guessing* atau lebih memilih untuk tidak merespon butir soal yang tidak sesuai dengan kemampuannya. Beberapa penelitian memang telah mempelajari masalah ini dan berusaha membuat rekomendasi yang sesuai, tetapi hasil penelitian masih sangat jarang dan kontradiktif (Yazdi, et.al., 2021). *Guessing* jelas diketahui berimplikasi pada menurunnya reliabilitas dan validitas (Bereby-Meyer et al., 2002; Kubinger, et. al., 2010; Prihoda, et.al., 2006).). Menambahkan jumlah butir tes maupun pilihan jawaban bukan hal yang bijak jika tujuannya hanya untuk memenuhi reliabilitas dan validitas saja (Burton, 2004; Karandikar, 2010). Penambahan butir seringkali menyulitkan pengembang tes dan merugikan karena membutuhkan waktu tes yang lebih panjang (Kurz, 1999) dan penambahan pilihan jawaban hampir dipastikan kurang berfungsi sebagai distraktor yang efektif (Lesage, Valcke, & Sabbe, 2013). Dari awal memang tidak ada aturan terkait berapa jumlah pilihan jawaban yang pas untuk meminimalisir *guessing* (Santoso, 2011). Oleh sebab itu, penelitian ini dilakukan untuk mengetahui karakteristik skor yang diestimasikan menggunakan model penskoran klasik yang berbeda, yakni *number right* (NR), *correction for guessing* variasi 4-pilihan jawaban (CFG-4), dan *correction for guessing* variasi 5-pilihan jawaban (CFG-5), ditinjau dari karakteristik statistik, karakteristik distribusi, maupun kesesuaian peringkat pada skor yang dihasilkan.

METODE PENELITIAN

Penelitian ini menggunakan metode penelitian deskriptif komparatif dengan pendekatan kuantitatif. Data dengan ukuran sampel sejumlah 300 responden (θ, rerata = -0,014 logit; SD = 1,024 logit) dibangkitkan dengan software WinGen berdasarkan *Item Response Theory* 1-Parameter Logistik (IRT 1-PL) dimana karakteristik parameter tingkat kesukaran butir *(b*) disajikan dalam Tabel 1 (rerata = -0,029 logit; SD = 1,105 logit) serta Fungsi Informasi Tes (*Test Information Function*, TIF) dan Kesalahan Pengukuran (*Standard Error of Measurement*, SEM) yang diestimasi dengan persamaan (1) dan (2) disajikan dalam Gambar 1.

$$\mathrm{I}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{I_i}(\theta) \text{ dengan } \mathrm{I_i}(\theta, \mathrm{X}) = \frac{\mathrm{D}^2 \mathrm{e}^{\mathrm{D}(\theta - \mathrm{b_j})}}{[1+\mathrm{e}^{\mathrm{D}(\theta-\mathrm{b_j})}]^2} \quad \ldots (1)$$

$$SEM\ (\theta) = \frac{1}{\sqrt{I_i\ (\theta)}} \quad \ldots (2)$$

Keterangan:

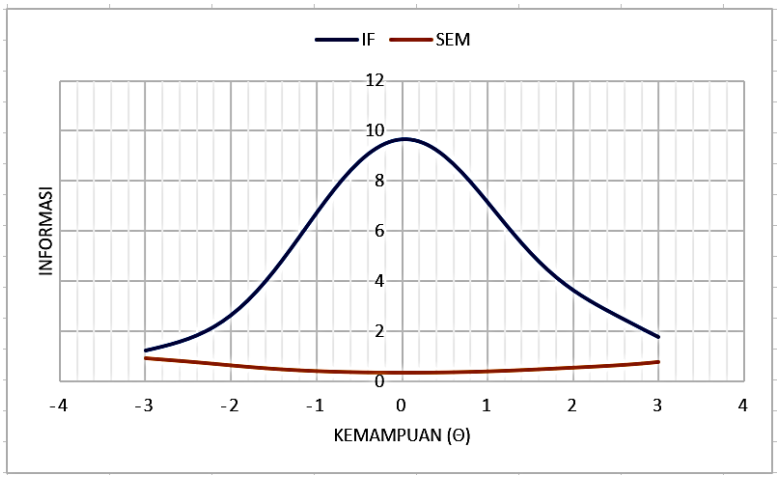
θ : Tingkat kemampuan (ability) peserta tes
b_j : Indeks kesukaran butir ke-j
e : bilangan natural yang nilainya mendekati 2,718
D : faktor penskalaan yang nilainya 1,7
$I_i(\theta)$: fungsi informasi butir
$I\ (\theta)$: fungsi informasi tes

Tabel 1. Karakteristik parameter tingkat kesukaran butir (*b*)

Butir	b	Butir	B	Butir	b	Butir	b
1	0,115	6	0,825	11	-1,050	16	-0,059
2	-3,025	7	2,500	12	0,429	17	-0,353

3	0,055	8	1,266	13	0,464	18	-1,067
4	-0,143	9	-0,729	14	0,372	19	-0,590
5	-0,453	10	1,062	15	-0,523	20	0,320

Data parameter butir yang dibangkitkan (Tabel 1) selanjutnya diolah pada rentang kemampuan -3,0 logit ≤ θ ≤ +3 logit untuk mendapatkan fungsi informasi butir (yang dijumlahkan menjadi TIF) dan SEM sebagaimana Gambar 1. Berdasarkan Gambar tersebut, tes akan memberikan informasi yang baik, dengan kesalahan pengukuran terkecil yaitu 0,32, apabila dikerjakan oleh peserta tes yang memiliki kemampuan sekitar 0,0 logit (kategori sedang). Selain memberikan informasi berupa TIF yang setara dengan reliabilitas, data bangkitan ini juga diestimasi reliabilitasnya dan didapatkan koefisien reliabilitas sebesar 0,984.



Gambar 1. Hubungan fungsi informasi tes dengan kesalahan baku pengukuran

Selain digunakan untuk mengolah TIF dan SEM, data bangkitan juga digunakan sebagai bahan untuk *scoring* menggunakan tiga model penskoran klasik, yakni *number right scoring* (X_a) dan *correction for guessing scoring* (X_c) untuk 4 pilihan jawaban dan 5 pilihan jawaban yang berturut-turut dipaparkan dalam persamaan (3) dan (4).

$$X_a = \sum_{i=1}^{n} x_{ai} \ldots (3)$$

$$X_c = R - \frac{W}{(k-1)} \ldots (4)$$

Keterangan:
X_a : skor hasil estimasi dengan model penskoran *number right*
X_c : skor hasil estimasi dengan model penskoran *correction for guessing.*
R : jumlah jawaban benar
W : jumlah jawaban salah
k : jumlah option jawaban

Adapun dua analisis yang dilakukan berbantuan software SPSS v.25, yakni analisis karakteristik deskriptif dan distribusi skor yang dihasilkan serta analisis komparatif ketiga model penskoran terhadap estimasi skor yang dihasilkan jika ditinjau dari kesesuaian skornya.

HASIL DAN PEMBAHASAN

Karakteristik deskriptif skor hasil estimasi dengan model penskoran NR, CFG-4, dan CFG-5 dapat dilihat pada Tabel 2. Berdasarkan tabel tersebut, nampak bahwa rerata skor hasil estimasi dengan model penskoran NR lebih tinggi dan lebih menyebar daripada rerata skor hasil estimasi dengan model penskoran CFG-4 dan CFG-5. Hasil serupa juga ditunjukkan pada data median dan modus, dimana model penskoran NR memiliki skor yang lebih tinggi. Sementara itu, terkait skor yang dihasilkan model penskoran CFG, nampak bahwa CFG-5 memberikan hasil estimasi skor yang lebih menguntungkan dibandingkan skor yang dihasilkan oleh model penskoran CFG-4. Hal ini dikarenakan besarnya hukuman yang diberikan tergantung pada jumlah pilihan jawaban yang disediakan oleh bentuk tes pilihan ganda (Kamaruddin, et.al., 2023). Pilihan jawaban yang disediakan oleh CFG-5 menyebabkan bilangan penyebutnya lebih banyak, sehingga hukuman yang dihasilkan relatif lebih kecil daripada CFG-4; yang pada akhirnya berimplikasi pada lebih tingginya skor hasil estimasi CFG-5 dibandingkan CFG-4.

Tabel 2. Karakteristik deskriptif skor

Statistik	Model Penskoran		
	NR	CFG-4	CFG-5
Minimum	0,00	0,00	0,00
Maksimum	20,00	13,00	15,00
Rata-rata	10,07	6,71	7,69
Median	10,00	7,00	8,00
Modus	12,00	9,00	11,00
Standar Deviasi	4,93	3,35	3,69
Skewness	0,017	0,035	0,009
Std. Error Skewness	0,141	0,141	0,141
Kurtosis	-0,929	-0,917	-0,969
Std. Error Kurtosis	0,281	0,281	0,281

Keterangan:
NR = *Number Right*
CFG-4 = *Correction for Guessing* variasi 4 pilihan jawaban
CFG-5 = *Correction for Guessing* variasi 5 pilihan jawaban

Untuk mengetahui karakteristik distribusi skor hasil estimasi model penskoran number right dapat diinterpretasi dari skewness dan kurtosis (Orcan, 2020). Skewness terkait dengan status modus, median, dan mean data yang relatif terhadap satu sama lain (Demir, 2022). Disebut distribusi simetris ketika rata-rata berada di tengah-tengah distribusi; dengan demikian, tidak ada kemencengan (skewness); tetapi ketika rata-rata tidak berada di tengah-tengah distribusi, maka terdapat distribusi yang tidak simetris (distribusi miring) (Tabachnick & Fidell, 2013). Sementara itu, kurtosis terkait dengan seberapa jauh data dari mean atau seberapa dekat data dengan mean, atau dengan kata lain, kurtosis terkait dengan standar deviasi data (Demir, 2022). Ketika standar deviasi kecil, maka distribusinya runcing (leptokurtik, berekor pendek); sedangkan, ketika standar deviasi besar, distribusinya rata (platykurtik, berekor panjang) (Field, 2013; Tabachnick & Fidell, 2013).

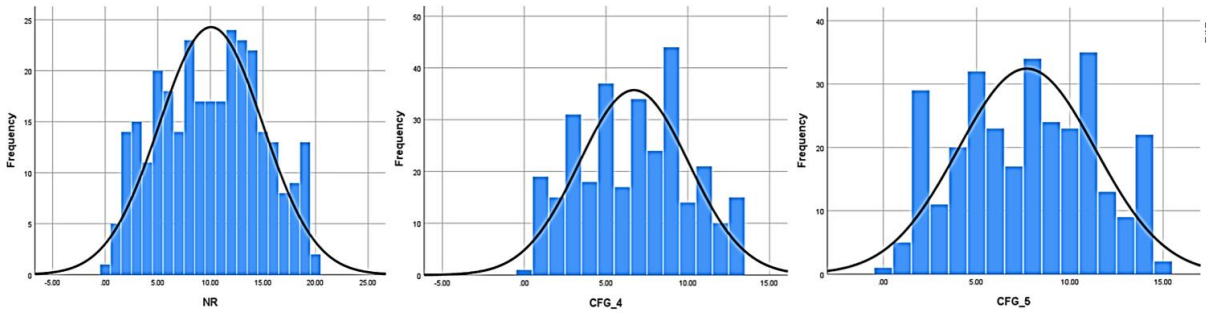
Interpretasi distribusi dengan skewness dan kurtosis dilakukan dengan uji-Z yang nilainya diolah dengan cara membagi nilai skewness atau kurtosis dengan kesalahan standar pengukurannya (Kim, 2013). Untuk sampel berukuran sedang ($50 < n \leq 300$), hipotesis nol akan ditolak pada nilai Z absolut di atas 3,29 ($\alpha = 0,05$) sehingga dapat disimpulkan jika distribusi

sampel tidak normal (Kim, 2013). Pada penelitian ini, uji-Z dilakukan dengan 2 sisi, sehingga dikatakan skor berdistribusi nomal jika Z skewness dan Z kurtosis berada pada $-3{,}29 \leq Z \leq 3{,}29$. Berdasarkan kriteria ini, maka dapat disimpulkan bahwa semua skor tidak berdistribusi normal sebagaimana disajikan ringkasan hasil uji-Z pada Tabel 3. Pada tabel tersebut, sekaligus dipaparkan hasil uji normalitas menggunakan statistik Kolmogorov-Smirnov maupun statistik Shapiro-Wilk yang juga sepakat bahwa semua skor tidak berdistribusi normal. Kedua statistik ini digunakan karena cukup *powerfull* bagi sampel berukuran kecil hingga sedang ($n \leq 300$), namun kurang dapat diandalkan jika digunakan sampel yang besar (Kim, 2013).

Tabel 3. Hasil uji karakteristik distribusi skor

Model Penskoran	Z skewness	Z kurtosis	Kolmogorov-Smirnov			Shapiro-Wilk		
			Statistik	df	Sig	Statistik	df	Sig
NR	0,12	-3,31	0,079	300	0,000	0,973	300	0,000
CFG-4	0,25	-3,26	0,099	300	0,000	0,965	300	0,000
CFG-5	0,06	-3,45	0,094	300	0,000	0,964	300	0,000

Meskipun sama-sama tidak berdistribusi normal, namun Gambar 2 mengindikasikan bahwa bentuk histogram dari ketiga skor hasil estimasi memiliki perbedaan. Histogram skor NR menunjukkan data yang lebih beragam dibandingkan histogram skor CFG-4 dan CFG-5, yakni berturut-turut sebesar 20, 14, dan 15 data; sehingga dapat disimpulkan bahwa penskoran yang diestimasi dengan model NR lebih menguntungkan *testee*, sebagaimana yang didukung data skor maksimum pada Tabel 2. Akan tetapi, jika ditinjau dari nilai skewness dan nilai kurtosisnya, ketiga skor hasil estimasi kembali memiliki persamaan lagi. Ditinjau dari nilai skewnessnya yang positif, ketiga skor yang dihasilkan dari model penskoran yang berbeda ini menunjukkan bahwa ekor di sisi kanan distribusi lebih panjang daripada ekor di sisi kiri, dengan sebagian besar nilai terletak di sebelah kiri rata-rata atau dapat dikatakan bahwa kurva juling ke kanan. Lalu jika ditinjau dari nilai kurtosisnya yang negatif, maka dapat disimpulkan bahwa distribusi puncak kurvanya adalah datar atau agak tumpul (platikurtik). Ini berarti, meskipun estimasi skor dilakukan dengan model penskoran yang berbeda, namun karakteristik distribusinya tetaplah serupa.



Gambar 2. Karakteristik distribusi skor berdasarkan model penskoran *number right*, *correction for guessing* variasi 4-pilihan jawaban dan 5-pilihan jawaban

Penggunaan model penskoran untuk melakukan estimasi skor pada akhirnya memang memberikan skor numerik yang berbeda. Untuk mengetahui apakah perbedaan numerik tersebut mempengaruhi peringkat *testee* secara signifikan, maka dilakukan Analisis Korelasi Intraklas yang hasilnya ditampilkan pada Tabel 4. Pada tabel tersebut, koefisien intraklas yang dihasilkan dari pengukuran tunggal (0,807) maupun rerata pengukuran (0,926) menunjukkan

hasil yang memuaskan, artinya ada kesesuaian yang tinggi pada peringkat *testee* meskipun skornya diestimasi menggunakan model penskoran yang berbeda. Hasil ini tentu kontradiktif dengan hasil analisis sebelumnya (Tabel 2 dan Gambar 2) yang menunjukkan bahwa model penskoran NR memberikan estimasi skor yang lebih menguntungkan dibandingkan CFR-4 dan CFR-5. Namun, ketika model penskoran yang digunakan sudah berbasis *Item Response Theory* sebagaimana penelitian Amelia & Setiawati (2016), maka skor yang dihasilkan justru memberikan peringkat yang berbeda sehingga dapat lebih dipertimbangkan lagi model penskoran mana yang akan digunakan pada bentuk tes pilihan ganda.

Tabel 4. Korelasi intraklas

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0,807	0,225	0,928	62,293	299	598	0,000
Average Measures	0,926	0,466	0,975	62,293	299	598	0,000

KESIMPULAN

Penggunaan model penskoran klasik yang berbeda sebagai alat bantu estimasi skor *testee* ternyata memberikan karakteristik deskriptif, karakteristik distribusi, dan kesesuaian peringkat yang tidak berbeda. Meskipun begitu, model penskoran CFG dapat dipertimbangkan penggunaannya sebagai upaya pencegahan *guessing* yang dilakukan *testee*, mengingat tidak mungkinnya pembuat soal dalam membedakan jawaban benar berdasarkan penguasaan atau *lucky guessing*. Disisi lain, estimasi skor yang dilakukan masih sebatas pada skor tampak (*observed score*) bukan pada skor sesungguhnya (*true score*) sebagai akibat penggunaan pendekatan klasik yang diacu, sehingga dalam skor yang dihasilkan ini tentu masih mengandung kesalahan pengukuran (*random error*). Oleh sebab itu, perlu digunakan model penskoran yang berbasis *Item Response Theory* yang memberikan hasil estimasi yang lebih *powerfull*, meskipun pada pelaksanaannya sedikit rumit karena perlunya pemenuhan asumsi yang tidak sedikit; atau alternatif lain berupa pengukuran langsung *ability* (θ) yang juga hanya bisa dilakukan melalui *Item Response Theory* yang kemudian dikonversi dalam ukuran yang mudah diinterpretasikan oleh *testee*, pembuat tes, hingga penyelenggara tes agar hasil tes dapat dimanfaatkan sesuai tujuannya.

DAFTAR PUSTAKA

Amelia, R.N., & Setiawati, F.A. (2016). Aplikasi model penskoran equal weighting dan differential weighting untuk mengestimasi skor kimia siswa. *Jurnal Evaluasi Pendidikan, 4*(1), 80-89

Bereby-Meyer, Y., Meyer, Y., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, *15*(4), 313–327. https://doi.org/10.1002/bdm.417

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41–50. https://doi.org/10.1080/02602930020022273

Burton, R. F. (2004). Multiple choice and true/false tests: Reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*, *29*(5), 585–595. https://doi.org/10.1080/02602930410001689153

Choppin, B. H. (1988). Correction for guessing. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 384–386). Oxford: Pergamon Press

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York: Holt, Reinhart, and Winston, Inc

Demir, S. (2022). Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients. *International Journal of Assessment Tools in Education*, *9*(2), 397-409. https://doi.org/10.21449/ijate.1101295

Field, A. (2013). *Discovering statistics using SPSS*. London: Sage Publications

Jacob, B. (2016). Student test scores: How the sausage is made and why you should care. *Brookings Institute*. Retrieved Juny, 20, 2023

Kamaruddin, E., Hanafi, I., Salman, I., & Ningtyas, L. D. (2023). The punishment score model to the mathematics learning outcomes of high school students in Jakarta. *Journal of Data Acquisition and Processing*, *38*(1), 225-231

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1), 1–73.* http://doi.org/10.1111/jedm.12000

Karandikar, R. L. (2010). On multiple choice tests and negative marking. *Current Science*, *99*(8), 1042–1045

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52-54. http://doi.org/10.5395/rde.2013.38.1.52

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*(1), 111–115. https://doi.org/10.1111/j.1468-2389.2010.00493.x

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests*. Paper Presented at the Annual Meeting of the Southwest Educational Research Association.

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking?. *Studies in Educational Evaluation, 39*(3), 188–193. http://doi.org/10.1016/j.stueduc.2013.07.001

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*(1), 7-11

Moses, T (2017). Psychometric contributions: Focus on test scores. In R.E. Bennett, M. von Davier (eds.), *Advancing human assessment, methodology of educational measurement and assessment* (pp. 47-78). New Jersey: Springer

Orcan, F. (2020). Parametric or non-parametric: Skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education*, *7*(2), 255-265. https://doi.org/10.21449/ijate.656077

Ostrosky, B. D., Reeve, K. F., Day-Watkins, J., Vladescu, J. C., Reeve, S. A., & Kerth, D. M. (2022). Comparing group-contingency and individualized equivalence-based instruction to a power point lecture to establish equivalence classes of reinforcement and punishment procedures with college students. *The Psychological Record*, *72*, 1–32. https://doi.org/10.1007/s40732-021-00495-6

Prieto, G., & Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment*, 15(2), 143–150. https://doi.org/10.1027//1015-5759.15.2.143

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, *70*(4), 378–386.

Romanoski, J., & Douglas, G. (2002). Test scores, measurement, and the use of analysis of variance: an historical overview. *Journal of applied measurement*, *3*(3), 232-242.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*(1), 15–22. http://doi.org/10.1111/j.1745-3984.1977.tb00024.x

Santoso, B. (2011). Inappropriatness score based on item response theory. *Jurnal Evaluasi Pendidikan*, *2*(2), 132-146. http://doi.org/10.21009/JEP

Stankous, N. V. (2016). Constructive response vs. multiple-choice tests in math: American experience and discussion. In *2nd PAN-American interdisciplinary conference, PIC 2016 24-26 February, Buenos Aires Argentina* (p. 321).

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics*. New Jersey: Pearson

Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement*, *29*, 847–861. https://doi.org/10.1177/001316446902900

Yazdi, M. S. G., Haghighat Shoar, S. M., Sobhani, G., Vafi Sani, F., Khoshkholgh, R., Mousavi Bazaz, N., & Mansourzadeh, A. (2021). Factors affecting students' guesswork in multiple choice questions and corrective strategies. *Medical Education Bulletin*, *2*(4), 341-349. https://doi.org/10.22034/MEB.2021.312176.1032