Jurnal Pendidikan Matematika
ISSN-p 2086-8235 | ISSN-e 2597-3592
Vol. 14, No. 2, Juli 2023, Hal: 135-146, Doi: http://dx.doi.org/10.36709/jpm.v14i2.100
Available Online at http://ojs.uho.ac.id/index.php/jpm

Konsistensi Internal Instrumen Tes: Perbandingan Beberapa Metode Estimasi Berdasarkan Variasi Ukuran Sampel

(*Internal Consistency of Test Instruments: Comparison of Several Estimation Methods Based on Variation in Sample Size***)**

Busnawir

[1]Jurusan Pendidikan Matematika, Universitas Halu Oleo, Kampus Hijau Bumi Tridharma Anduonohu, Kendari, Indonesia

Abstrak: Salah satu ukuran reliabilitas instrumen adalah konsistensi internal butir yang menyusunnya. Beberapa metode estimasi sering digunakan tetapi kadang memberikan koefisien reliabilitas yang berbeda sehingga menjadi masalah dalam memberikan kesimpulan. Tujuan penelitian ini untuk menemukan karakteristik metode estimasi konsistensi internal yang paling efektif ditinjau dari ukuran sampel dan model sampling yang berbeda. Metode yang digunakan adalah eksperimen dalam bentuk simulasi. Data primer penelitian merupakan skor hasil tes ujian tengah semester siswa kelas VIII SMPN 9 Kendari, yang merupakan jawaban dari 400 siswa terhadap instrumen tes matematika yang berjumlah 12 item pilihan ganda. Simulasi dilakukan dengan menarik secara acak sampel berukuran 31, 62, 93, 124, dan 155, dengan cara pengembalian dan tanpa pengembalian. Koefisien konsisten internal dihitung menggunakan tiga metode estimasi yang berbeda yaitu metode Sperman-Brown, KR-20, dan KR-21. Hasil penelitian menunjukkan metode *Spearman-Brown* memberikan koefisien konsistensi internal lebih tinggi dibandingkan metode KR-20 dan KR-21 pada semua ukuran sampel baik menggunakan sampling dengan pengembalian maupun sampling tanpa pengembalian. Kesimpulannya adalah metode *Spearman-Brown* paling efektif dalam menentukan koefisien konsistensi internal butir instrumen tes dibandingkan metode KR-20 dan KR-21, baik menggunakan sampel dengan pengembalian maupun tanpa pengembalian.

Kata kunci: konsistensi internal; metode estimasi; ukuran sampel.

Abstract*: One measure of instrument reliability is the internal consistency of the items that compose it. Several estimation methods are often used but sometimes provide different reliability coefficients so it becomes a problem in concluding. The purpose of this study was to find the characteristics of the internal consistency estimation method which is more effective in terms of sample size and different sampling models. The method used is an experiment in the form of a simulation. The primary research data is the scores of the midterm test results for class VIII students of SMPN 9 Kendari, which are the answers of 400 students to the mathematics test instrument, which consists of 12 multiple choice items. The simulation was carried out by randomly drawing samples of sizes 31, 62, 93, 124, and 155, with replacement and without replacement. Then the internal consistent coefficient was calculated using the Sperman-Brown, KR-20, and KR-21 methods. The results showed that the Spearman-Brown method gave a higher reliability coefficient than the other methods, and the KR-20 method gave a higher reliability coefficient than KR-21. The sample size did not affect the high-reliability coefficient linearly. It was concluded that the Spearman-Brown method was more effective in determining the internal consistency of the test instrument items, similarly, KR-20 was more effective than KR-21, both using samples with replacement and without replacement.*

Keywords: *Estimation Method; Internal Consistency; Sample Size.*

PENDAHULUAN

Permasalahan mendasar dalam pengukuran kognitif adalah bagaimana mengetahui bahwa instrumen tes yang digunakan memberikan hasil secara tepat sesuai dengan kemampuan individu yang sebenarnya. Ada dua pertanyaan yang sering dikemukakan berkaitan dengan hal tersebut yaitu, pertama apakah instrumen tes yang digunakan memiliki kualitas baik?, dan kedua apakah instrumen tes yang digunakan dapat menghasilkan keputusan lebih baik dibandingkan pengukuran lainnya?. Kedua pertanyaan ini mempermasalahkan sejauh mana

* Korespondensi Penulis. E-mail: busnawir@uho.ac.id Penerbit: Jurusan Pendidikan Matematika FKIP Universitas Halu Oleo

ketepatan dan ketetapan hasil pengukuran dalam menjelaskan karakteristik individu sesuai dengan obyek ukurnya. Ketepatan dan ketetapan hasil pengukuran merupakan dua unsur pokok yang merepresentasikan kualitas hasil pengukuran, yang pertama berkenaan dengan validitas dan kedua berkenaan dengan reliabilitas (Mustafa & Masgumelar, 2022; Solichin, 2017; Rosita et al., 2021). Validitas suatu instrumen menekankan pada ketepatan mengukur apa yang dimaksudkan untuk mengukur, sedangkan reliabilitas menekankan pada sejauh mana hasil pengukuran dapat diandalkan berdasarkan derajat konsistenannya (Sari et al., 2019; Rapono et al., 2019).

Membahas tentang reliabilitas suatu instrumen tes berarti melihat dua masalah pokok yang berkaitan di dalamnya, yaitu masalah konsistensi internal dan konsistensi eksternal (Listiyandini et al., 2020; Sutriani et al., 2021). Ada tiga terminologi yang menggambarkan reliabilitas pengukuran, yaitu stabilitas, ekuivalensi, dan konsistensi internal. Pendekatan konsistensi internal menunjukkan bahwa antara satu bagian tes dan bagian lainnya menghasilkan pengukuran yang konsisten dan diindikasikan oleh tingginya korelasi antara bagian itu yang biasa dihitung dengan koefisien *Spearman-Brown* (Sarwiningsih, 2017; Sumintono & Widhiarso, 2015), menggunakan KR-20 dan KR-21 (Sarwiningsih, 2017); teknik Hoyt (Khaerudin, 2015); keoefisien alpha (Alwi, 2015). Sementara konsistensi eksternal (reliabilitas eksternal) diperoleh dengan menggunakan skor pengukuran yang berbeda, yang dapat diestimasi dengan dua cara yaitu menggunakan teknik pengukuran ulang (*test-retest-method*) dan teknik paralel (Retnawati, 2016).

Menguji konsistensi internal instrumen merupakan pilihan lain yang dapat digunakan untuk menguji reliabilitas selain menghitung koefisien stabilitas dan kesepadanan (Imania & Bariah, 2019). Konsep reliabilitas menurut pendekatan ini ialah konsistensi pertanyaan pada butir-butir instrumen. Korelasi antara butir pertanyaan atau pernyataan dalam suatu instrumen yang mengukur suatu konstruk tertentu menunjukkan tingkat reliabilitas konsistensi internal instrumen tersebut (Sarwiningsih, 2017; Imania & Bariah, 2019). Konsistensi internal yang tinggi mencerminkan kualitas instruemn yang lebih baik (Taber, 2018).

Konsistensi internal dapat dihitung menggunakan beberapa formula reliabilitas (Alwi, 2015; Khaerudin, 2015; Sarwiningsih, 2017; Sumintono & Widhiarso, 2015), namun setiap formula memiliki karakteristik yang berbeda sehingga ada kemungkinan akan menghasilkan koefisien atau indeks yang berbeda pula (Widayati, 2013; Arifin, 2017). Pemilihan formula yang tepat sangat menentukan ketepatan pendugaan hasil pengukuran sehingga dapat memperkecil terjadinya bias dari nilai yang sebenarnya (Sartika, 2018; Busnawir, 2018). Metode estimasi yang baik akan memberikan indeks relibilitas yang relatif tinggi (Geldhof et al., 2014).

Beberapa penelitian terdahulu telah melakukan kajian terkait dengan konsistensi internal instrumen pengukuran, di antaranya menemukan bahwa terdapat perbedaan besarnya konsistensi internal instrumen tes berdasarkan perbedaan metode atau formula yang digunakan (Sarwiningsih, 2017; Widodo, 2006). Koefisien reliabilitas tidak mengalami perubahan atau perbedaan pada data, sebelum dan sesudah ditransformasi (Setiawati et al., 2013). Model pensekoran dan variasi usia responden mempengaruhi besarnya koefisien relibailitas internal pengukuran pada skala sikap (Busnawir, 2018). Indeks reliabilitas suatu instrumen dipengaruhi oleh jumlah butir tes, variabilitas dalam kelompok, objektivitas pemberian skor, metode estimasi reliabilitas, level dalam kelompok, tingkat kesukaran, serta homogenitas tes (Setiyawan, 2014).

Sayangnya, penelitian terdahulu tidak mengkaji secara spesifik tentang metode yang tepat dalam menentukan indeks konsistensi internal suatu instrumen, serta tidak memperhitungakn ukuran sampel dan teknik sampling yang digunakan. Oleh sebab itu, penelitian ini mencoba mengkaji lebih jauh dengan membandingkan beberapa metode pendugaan dalam menghasilkan

indeks konsistensi internal dengan mempertimbangkan ukuran sampel (*testee*) pada tes formatif matematika bentuk *multiple choice*. Secara statistika, ukuran sampel yang semakin besar diharapkan akan memberikan hasil yang semakin baik. Nitko mengemukakan bahwa dalam mengevaluasi kestabilan koefisien reliabilitas, ukuran sampel harus diperhatikan karena reliabilitas dihitung berdasarkan sampel responden sehingga berbeda-beda tergantung besar kecilnya sampel yang diambil dari populasi (Putri & Nahadi, 2019); dengan sampel yang besar, mean dan standar deviasi yang diperoleh mempunyai probabilitas yang tinggi untuk menyerupai mean dan standar deviasi populasi (Ndiung & Jediut, 2020); tiga metode pendugaan yang sering digunakan dalam menentukan indeks konsistensi internal instrumen tes yaitu *Spearman-Brown* (S-P), *Kurd Ricardson* 20 (KR-20) dan KR-21 (Imania & Bariah, 2019; Sarwiningsih, 2017).

Formula reliabilitas *Spearman-Brown* merupakan metode teknik belah dua, di mana instrumen dikerjakan satu kali oleh sejumlah sampel peserta tes. Butir-butir pada perangkat dibagi menjadi dua bagian yang sama. Pembagian dapat menggunakan nomor ganjil-genap pada instrumen, atau separuh pertama maupun separuh kedua, maupun membelah dengan menggunakan nomor acak atau tanpa pola tertentu. Skor responden (*testee*) setengah perangkat bagian yang pertama dikorelasikan dengan skor setengah perangkat pada bagian yang kedua. Metode *Spearman-Brown* cocok digunakan untuk instrumen dengan jumlah butir genap sehingga dapat dibelah dua secara seimbang, dan juga tidak mempersoalkan tingkat kehomogenan kesukatran butir. Adapun rumus *Spearman-Brown* dinotasikan seperti berikut (Retnawati, 2016; Syamsuryadin & Wahyuniati, 2017):

$$r_i = \frac{2r_b}{1+r_b}$$

$$\textbf{Dengan } r_b = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}} \qquad \text{.....(1)}$$

Keterangan: ri = koefisien reliabilitas skor instrumen; rb = koefisien korelasi antara dua belahan instrumen, N = banyaknya responden, X = belahan pertama, Y = belahan kedua.

Kuder-Richardson 20 yang disingkat KR-20 merupakan rumus reliabilitas komposit yang sering digunakan pada instrumen tes yang terdiri dari banyak butir, di mana butir-butir ini merupakan butir yang berbeda-beda namun membangun suatu konstruk yang sama. Komposit yang dimaksudkan yakni skor akhir merupakan gabungan dari skor butir-butir penyusun instrumen. Prinsip kerjanya, membagi n butir instrumen menjadi n bagian, masing-masing terdiri dari satu butir dan diasumsikan setiap butir memiliki satu faktor persekuruan. Metode KR-20 dapat diterapkan pada skor yang bernilai dikotomi, misalanya 1 dan 0, ya dan tidak, benar dan salah. Rumus reliabilitas KR-20, diformulasikan sebagai berikut (Retnawati, 2016; Syamsuryadin & Wahyuniati, 2017):

$$r_{ii} = \frac{k}{(k-1)}\left\{\frac{s_t^2 - \sum p_i q_i}{s_t^2}\right\} \qquad \text{.....(2)}$$

Keterangan: rii= reliabilitas skor instrumen; k=banyaknya butir pertanyaan atau banyaknya soal; $st2$ = varians skor total; pi= proporsi subjek yang menjawab betul, qi = 1- pi.

Kuder-Richardson 21 yang disingkat KR-20 KR-21 pada prinsipnya sama dengan KR-20, namun dapat digunakan untuk instrumen dengan skor butirnya dikotomi dan juga politomi. Selain itu diasumsikan bahwa setiap butir tes memiliki tingkat kesukaran yang sama atau homogen, dengan formula (Retnawati, 2016; Syamsuryadin & Wahyuniati, 2017):

$$r_{ii} = \frac{k}{k-1}\left(1 - \frac{\bar{X}(k-\bar{X})}{k\sigma_t^2}\right) \qquad \text{.....(3)}$$

Keterangan: $\boldsymbol{rii}$ = koefisien reliabilitas skor instrumen; $\boldsymbol{k}$= banyaknya butir pertanyaan atau banyaknya soal; $\boldsymbol{\sigma t2}$ = varians total; $\boldsymbol{\bar{X}}$= skor rata-rata.

Sifat dari ketiga metode reliabilitas adalah, *Speraman-Brown* lebih cocok digunakan bila jumlah soal genap, sehingga dapat dibelah dua. Jika jumlah soal ganjil, digunakan rumus KR-20 dan KR-21. Rumus KR-20 digunakan untuk menghitung reliabilitas suatu rapid test, yaitu tes kecepatan. Pada saat yang sama, rumus KR-21 lebih cocok untuk tes profisiensi, yang memungkinkan siswa mendemonstrasikan keterampilan mereka sepenuhnya (Alwi, 2015; Gunartha, 2022). Salah satu tujuan dari ketiga metode ini adalah untuk menghindari masalah yang biasanya dihadapi dengan pendekatan paralel dan metode pengujian ulang. Pendekatan konsistensi internal dilaksanakan dengan menggunakan format ujian tunggal yang diberikan kepada sekelompok mata pelajaran dalam sekali (*single administration*), sehingga bernilai lebih praktis (Gunartha, 2022).

Penelitian ini mencoba menerapkan ketiga metode perhitungan reliabilitas yang telah dikemukakan degan ukuran sampel yang berbeda, dengan teknik sampling "*with replacement*" dan "*without replacement*" pada instrumen tes formatif matematika dengan jumlah butir genap. Tujuan penelitian untuk menemukan metode pendugaan koefisien reliabilitas internal atau indeks konsistensi internal yang lebih baik (lebih efektif) pada instumen tes formatif matematika dengan mempertimbangkan ukuran sampel peserta tes (*testee*) baik dengan teknik sampling "*with replacement*" dan "*without replacement*". Metode pendugaan dikatakan lebih baik apabila menghasilkan koefisien reliabilitas yang tinggi. Selain itu, ingin pula diketahui apakah perbedaan ukuran sampel dan teknik sampling menghasilkan indeks koefisien reliabilitas yang berbeda.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dalam bentuk eksperimen dan simulasi (Sari et al., 2017) dengan 5 (lima) tahapan. Pada tahap awal, menentukan data primer (data asli) yang akan digunakan untuk simulasi. Data primer diperoleh dari hasil tes formatif (hasil ujian tengah semester) mata pelajaran matematika pada kelas VIII SMP Negeri 9 Kendari. Tes tersebut memuat materi tentang matriks, deret, dan barisan bilangan. Data berbentuk skor hasil tes yang merupakan jawaban dari 400 siswa sebagai populasi skor data penelitian. Tahap kedua, membangkitkan data yang ditarik secara acak dari populasi data menggunakan bantuan aplikasi *Minitab for Windows*. Sampel data ditarik secara berulang sebanyak 5 kali dalam jumlah sampel yang berbeda-beda dengan kelipatan 31 unit sampel, yaitu 31, 62, 93, 124, dan 155. Kelipatan ini didasarkan pada asumsi distribusi normalitas data (Pujiastuti et al., 2023). Teknik penarikan sampel menggunakan "*Sample with replacement*" (penarikan sampel dengan pengembalian) dan "*Sample without replacement* (penariksan sampel tanpa pengembalian)". Tahap ketiga, berdasarkan data yang telah dibangkitkan, selanjutnya menghitung indeks konsistensi internal menggunakan metode Spearman-Brown, KR-20, dan KR-21. Tahap keempat, membandingkan besarnya indeks konsistensi internal yang dihasilkan oleh ketiga metode yang digunakan pada setiap jumlah sampel. Tahap kelima, membuat justifikasi atau kesimpulan.

Data yang diperoleh berdasarkan ukuran sampel dan metode pendugaan koefisien konsistensi internal, selanjutnya dianalisis secara deskriptif dengan membandingkan besaran kuantitatif nilai statistik yang dihasilkan. Desain analisis dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Desain Penelitian

Ukuran Sampel	Metode Perhitungan Koefisien Konsistensi Internal		
	***Spearman-Brown* (S-B)**	**KR-20**	**KR-21**
31	r_{11}(S-B_{31})	r_{11}(KR-20_{31})	r_{11}(KR-21_{31})
62	r_{11}(S-$B6_2$)	r_{11}(KR-20_{62})	r_{11}(KR-21_{62})
93	r_{11}(S-B_{93})	r_{11}(KR-20_{93})	r_{11}(KR-21_{93})
124	r_{11}(S-B_{124})	r_{11}(KR-20_{124})	r_{11}(KR-21_{124})
155	r_{11}(S-B_{155})	r_{11}(KR-20_{155})	r_{11}(KR-21_{155})
Rata-rata	Rata-rata r_{11} (S-P)	Rata-rata r_{11} (KR-20)	Rata-rata r_{11} (KR-21)
Varians	Var r_{11} (S-P)	Var r_{11} (KR-20)	Var r_{11} (KR-21)
Std. Error	SE r_{11} (S-P)	SE r_{11} (KR-20_	SE r_{11} (KR-21)

Kriteria yang digunakan dalam penelitian ini ialah, metode yang menghasilkan koefisien konsistensi internal atau koefisien reliabilitas internal yang lebih tinggi mengindikasikan sebagai metode yang baik (efektif) dalam menentukan tingkat reliabilitas intrumen tes. Reliabilitas tinggi menunjukkan tingkat error yang kecil (Musliha, 2019); sehingga memberikan tigkat ketetapan yang tinggi pula (Hanifah, 2014). Reliabilitas berkaitan dengan sejauh mana suatu instrumen memberikan hasil dan kesimpulan yang dapat dipercaya (Pramuaji & Loekmono, 2018).

HASIL DAN PEMBAHASAN

Dalam penelitian ini, simulasi dilakukan pada penarikan sampel dengan ukuran yang berbeda terhadap data primer tentang skor hasil tes ujian tengah semester mata pelajaran matematika, yang terdiri atas 12 butir soal multiple choce, bentuk jawaban benar salah yang diberi skor 1 (benar) dan 0 (salah). Terdapat 5 (lima) ukuran sampel yang berbeda dibangkitkan dari 400 peserta tes (testee) sebagai populasi data seperti ditunjukkan pada Tabel 1, menggunakan teknik penarikan sampel secara acak "dengan pengembalian" dan "tanpa pengembalian". Deskriptif statistik nilai-nilai yang dihasilkan menurut ketiga metode perhitungan koefisien reliabilitas internal dan kelima ukuran sampel yang berbeda, menggunakan penarikan sampel dengan pengembalian, dirangkum pada Tabel 2.

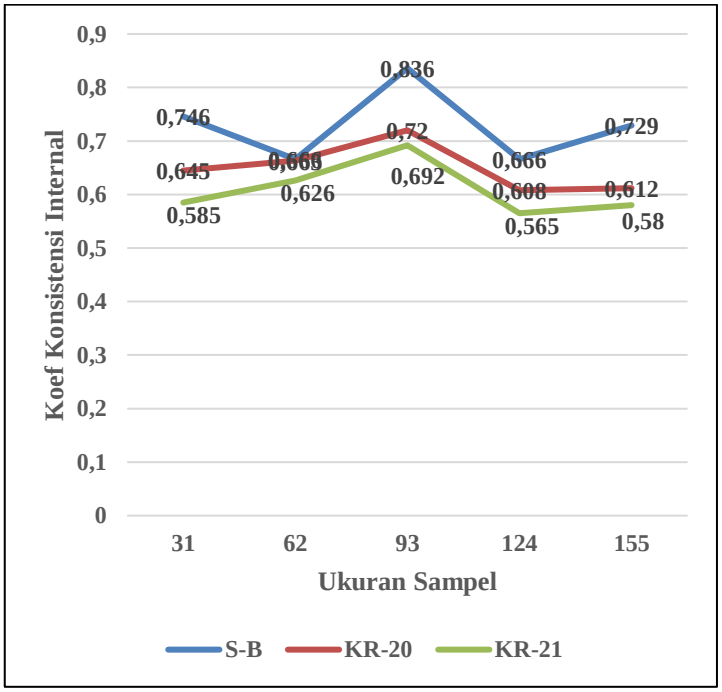
Tabel 2. Koefisien Relibilitas Internal: Kasus Sampel dengan Pengembalian

Ukuran Sampel	Metode Perhitungan		
	Spearman-Brown	**KR-20**	**KR-21**
31	0,746	0,645	0,585
62	0,666	0,663	0,626
93	0,836	0,720	0,692
124	0,666	0,608	0,565
155	0,729	0,612	0,580
Rata-rata	0,7286	0,6496	0,6096
Varians	0,004919	0,002075	0,002632
Std.Error	0,031365	0,020373	0,022945

Berdasarkan Tabel 2, dapat dilihat bahwa koefisien reliabilitas internal yang dihasilkan oleh metode *Spearman-Brown*, KR-20, dan KR-21 memberikan nilai besaran yang berbeda pada semua ukuran sampel. Ketiga metode perhitungan memberikan nilai yang bersifat fluktuatif atau besaran yang tidak konsisten untuk setiap ukuran sampel. Namun, secara

kuantitatif, metode *Spearman-Brown* menghasilkan nilai koefisien reliabilitas internal yang lebih besar dibandingkan mertode KR-20 dan KR-21. Begitu pula, jika dilihat berdasarkan nilai varians dan standar error metode *Spearman-Brown* juga lebih besar dari metode KR-20 dan KR-21. Hal ini memberikan isyarat bahwa koefisien reliabilitas yang dihasilkan metode *Spearman-Brown* berfluktuasi lebih besar.

Secara visual, fluktuasi koefisien reliabilitas yang dihasilkan oleh ketiga metode berdasarkan ukuran sampel yang berbeda ditunjukkan pada Gambar 1.



Gambar 1. Koefisien Reliabilitas Internal Pada Sampel dengan Pengembalian

Pada Gambar 1 tampak jelas, metode *Spearman-Brown* (S-B) memberikan nilai koefisien reilabilitas internal lebih tinggi pada semua ukuran sampel dibandingkan kedua metode yang lainnya. Tampak pula bahwa metode KR-20 memberikan nilai koefisien relibilitas internal lebih tinggi dari pada KR-21 pada semua ukuran sampel. Satu hal yang menarik bahwa pada ukuran sampel 93, semua metode memberikan nilai koefisien reliabilitas internal yang lebih tinggi dibandingkan ukuran sampel lainnya. Dengan demikian maka metode *Spearman-Brown* (S-B) dapat dikatakan paling efektif pada semua ukuran sampel dengan pengembalian dalam menghasilkan koefisien relibalilitas internal intrumen tes karena memberikan koefisien reliabilitas yang lebih tinggi dibandingkan metode lainnya.

Pada penarikan sampel "tanpa pengembalian", ketiga metode yang digunakan memberikan nilai koefisien reliabilitas internal yang juga bersifat fluktuatif atau tidak konsisten ditinjau dari ukuran sampel yang berbeda. Metode *Spearman-Brown* masih memberikan nilai koefisien reliabilitas yang lebih tinggi dibandingkan dengan kedua metode lainnya, seperti halnya pada penarikan sampel dengan pengembalian. Rangkuman nilai koefisien reliabilitas dari ketiga metode menurut ukuran sampel yang berbeda, pada sampel tanpa pengembalian ditunjukkan pada Tabel 3.

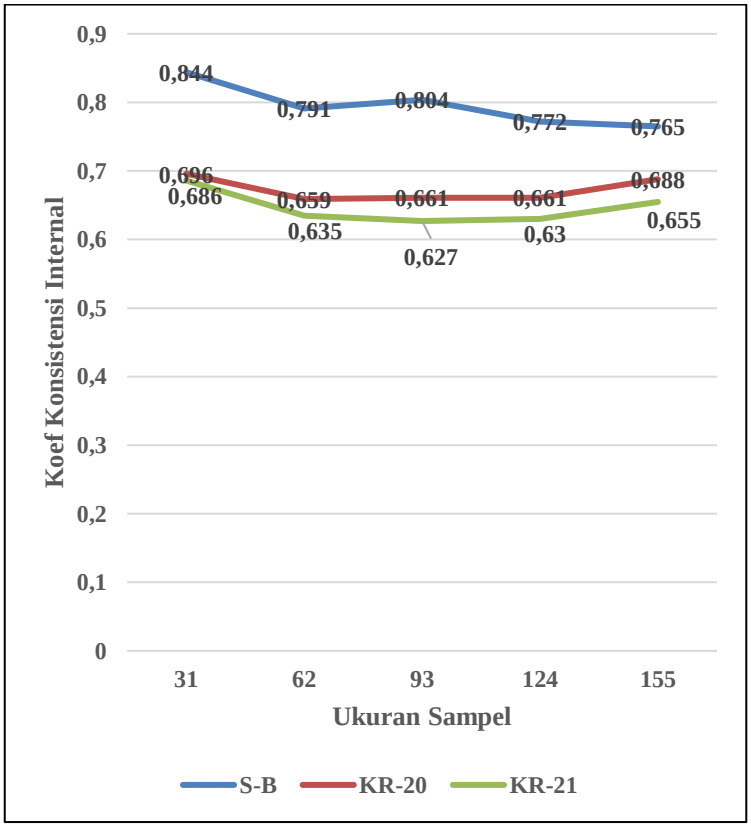
Pada Tabel 3 memperlihatkan nilai koefisien reliabilitas yang dihasilkan oleh metode *Spearman Brown* lebih tinggi dibandingkan metode KR-20 dan KR-21. Metode Spearman Brown memberikan rata-rata nilai koefisien relibilitas di sekitar 0,80, sementara metode KR-20 dan KR-21 masing-masing berkisar 0,67 dan 0,65. Pada sampling tanpa pengembalian, standar error koefisien reliabilitas yang dihasilkan oleh metode Spearman Brown lebih besar

dari pada KR-20 dan KR-21, yang mengindikasikan bahwa fluktuasi nilai-nilai koefisien reliabilitas yang dihasilkan metode *Spearman Brown* lebih besar dibandingkan KR-20 dan KR-21.

Tabel 3. Koefisien Relibilitas Internal: Kasus Sampel Tanpa Pengembalian

Ukuran Sampel	Metode Perhitungan		
	S-B	KR-20	KR-21
31	0,844	0,696	0,686
62	0,791	0,659	0,635
93	0,804	0,661	0,627
124	0,772	0,661	0,630
155	0,765	0,688	0,655
Rata-rata	0,7952	0,673	0,6466
Varians	0,000982	0,000309	0,000604
Std.Error	0,014012	0,007868	0,010994

Secara visual tampak lebih jelas terlihat, bahwa koefisien reliabilitas internal yang dihasilkan oleh metode Spearman Brown jauh lebih tinggi dari pada kedua metode yang lainnya, seperti ditunjukkan pada Gambar 2.



Gambar 2. Koefisien Konsistensi Internal Berdasarkan Sampel Tanpa Pengembalian

Pada Gambar 2 menunjukkan bahwa ukuran sampel 31 cenderung memberikan nilai koefisien reliabilitas lebih tinggi dibandingkan ukuran sampel yang lebih besar dari 31 untuk semua metode yang digunakan dalam menentukan nilai koefisien reliabilitas. Pada Gambar 2 terlihat pula bahwa, nilai koefisien reliabilitas internal yang dihasilkan metode KR-20 cenderung lebih tinggi dari pada KR-21, meskipun dalam selisih yang masih relatif kecil. Dalam hal ini, pada sampling tanpa pengembalian, untuk metode estimasi Spearman Brown

menunjukkan semakin besar ukuran sampel cenderung memberikan nilai koefisien reliabilitas internal yang lebih rendah, sedangkan pada metode KR-20 dan KR-21 cenderung mengalami peningkatan pada ukuran sampel yang lebih besar dari 93. Berdasarkan uraian yang telah dikemukakan menunjukkan bahwa pada sampling tanpa pengembalian metode *Spearman-Brown* (S-B) juga merupakan metode paling efektif pada semua ukuran sampel dalam menghasilkan koefisien reliablilitas internal intrumen tes karena memberikan koefisien reliabilitas yang lebih tinggi dibandingkan kedua metode lainnya yaitu KR-20 dan KR-21.

Hasil penelitian menunjukkan bahwa pada sampling dengan pengembalian (*sample with replacement*) metode estimasi menggunakn formula *Spearman Brown* cenderung memberikan nilai koefisien reliabilitas yang lebih tinggi dibandingkan dengan metode KR-20 dan KLR-21, meskipun nilai-nilai yang dihasilkan bersifat fluktuatif berdasarkan ukuran sampel yang berbeda. Namun demikian, dari 5 (lima) variasi ukuran sampel yang diujicobakan yaitu 31, 62, 93, 124, dan 155, pada ukuran sampel 93 ketiga metode menghasilkan koefisien reliabilitas yang cenderung lebih tinggi dibandingkan ukuran sampel yang lainnya. Sementara pada sampling tanpa pengembalian (*sample without replacement*), ukuran sampel 31 cenderung menghasilkan indeks reliabilitas yang lebih tinggi dibandingkan ukuran sampel lainnya. Menurut (Amalia et al., 2022), menemukan bahwa ukuran sampel 30 memberikan indeks koefisien reliabilitas yang relatif tinggi dibandingkan ukuran sampel 15 dan 39. Besarnya koefisien reliabilitas dipengaruhi oleh, antara lain banyak butir tes, instrumen tes yang digunakan, penyelenggara tes, dan subyek yang diukur (Syamsuryadin & Wahyuniati, 2017), variabilitas kelompok, metode estimasi reliabilitas, level kelompok, homogenitas tes (Setiyawan, 2014). Ukuran sampel tidak signifikan mempengaruhi konsistensi internal (Button et al., 2013).

Kecenderungan nilai koefisien reliabilitas internal yang tinggi pada metode *Spearman Brown* karena pembedahan butir tes menjadi dua bagian yang sama besar. Seperti dikemukakan oleh (Sarwiningsih, 2017), bahwa rumus *Spearman Brown* cocok digunakan jika tes berbentuk dikotomi, pembagian tes antara belahan pertama dan kedua dalam jumlah yang seimbang (paralel) serta korelasi antara kedua belahan itu cukup tinggi. Menurut (Ekolu & Quainoo, 2019) dalam penelitiannya menemukan bahwa terdapat korelasi yang kuat antara KR-21 dan metode split-hlaf. Namun jika dihubungkan dengan ukuran sampel, (Putri & Nahadi, 2019) menemukan bahwa ukuran sampel (30 dan 40) tidak memberikan perbedaan besarnya koefisien reliabilitas pada tes pilihan ganda. Hasil ini juga didukung penelitian yang dilakukan oleh (Amalia et al., 2022), yang menemukan bahwa indeks reliabilitas bersifat fluktuatif pada tiga ukuran sampel yang digunakan, artinya besarnya ukuran sampel tidak berhubungan secara linear terhadap besarnya indeks koefisien reliabilitas. Hal yang sama juga dikemukakan oleh (Button et al., 2013) bahwa reliabilitas bersifat fluktuatif pada ukuran sampel yang berbeda.

Hasil penelitian seperti dirangkun pada Tabel 2 dan Tabel 3, metode estimasi KR-20 dan KR-21 rupanya tidak lebih baik digunakan pada intrumen tes dengan jumlah butir genap. Hal itu ditunjukkan oleh indeks koefisien reliabilitas internal yang dihasilkan cenderung lebih rendah dibandingkan metode *Spearman Brown* baik pada sampling dengan pengembalian maupun sampling tanpa pengembalian. Hal ini diperkuat dengan prinsip penggunaan kedua metode ini, yaitu jika jumlah butir suatu instrumen tes tidak dapat dibagi menjadi dua bagian yang setara, bila dalam setiap bagian tes terdapat sedikit jumal butir maka estimasi reliabilitasnya menjadi tidak cermat (Gunartha, 2022). Dalam hal ini, KR-20 dan KR-21 akan baik digunakan jika jumlah butir intrumen adalah ganjil (Nuswowati et al., 2011) dan butir soal memiliki tingkat kesukaran yang homogen (Mustafa & Masgumelar, 2022).

Jika dibandingkan indeks koefisien reliabilitas berdasarkan konsistensi internal butir instrumen tampak estimasi yang dihasilkan oleh KR-20 relatif lebih tinggi dibandingkan KR-21. Kedua metode estimasi ini dilakukan dengan cara menganalisis skor setiap butir tes secara langsung atau membedah instrumen sebanyak jumah butirnya. Secara spesifik kedua metode

ini memiliki ciri khas tersendiri, kelebihan formula KR-20 adalah lebih teliti dalam perhitungan dibanding formula KR-21. Namun, dalam proses perhitungan KR-20 lebih sulit dibandingkan KR-21. Sementera formula KR-21 dalam penggunaannya relatif lebih sederhana (lebih mudah) tetapi hasil perhitungannya kurang cermat atau kurang teliti (Magdalena et al., 2021). Formula KR-20 bisa menghasilkan estimasi relatif lebih tinggi dibandinghkan KR-21, hal ini bisa terjadi karena tingkat kesukaran butir tes lebih bervariasi (Sarwiningsih, 2017). Pada kasus lain metode KR-21 menghasilkan koefisien reliabilitas relatif lebih tinggi dibandingkan KR-20 (Gambrari et al., 2015).

KESIMPULAN DAN SARAN

Penelitian ini memberikan beberapa kesimpulan, yang berlaku pada kasus instrumen tes matematika dengan jumlah butir genap. Kesimpulan yang dihasilkan adalah bahwa meningkatnya ukuran sampel peserta tes tidak mempengaruhi meningkatnya koefisien reliabilitas internal instrumen tes (tidak berkorelasi positif); jika ukuran sampel meningkat koefisien reliabilitas internal bersifat fluktuatif (naik turun); metode *Spearman-Brown* (S-P) adalah metode paling efektif dibandingkan KR-20 dan KR-21 dalam menghasilkan indeks konsistensi internal instrumen tes baik menggunakan sampel dengan pengembalian maupun tanpa pengembalian; koefisien reliabilitas internal lebih tinggi jika dikenakan pada sampel tanpa pengembalian dibandingkan pada sampel dengan pengembalian; ukuran sampel 31 dianggap paling efektif untuk menghasilkan koefisien konsistensi internal instrumen tes dibandingkan ukuran sampel yang lebih besar pada sampel tanpa pengembalian; ukuran sampel 93 dianggap paling efektif untuk menghasilkan koefisien konsistensi internal instrumen tes dibandingkan ukuran sampel yang lainnya pada sampel dengan pengembalian

Keterbatasan penelitian ini ialah, instrumen yang dianalisis memiliki butir tes relatif sedikit dan berjumlah genap, tidak memperhatikan tingkat kesukaran butir tes, serta tidak melakukan uji validitas butir terlebih dahulu. Untuk itu dikan agar dapat dilakukan penelitian serupa dengan memperhatikan beberapa keterbatasan penelitian ini serta menerapkan metode estimasi reliabilitas lainnya.

DAFTAR PUSTAKA

Alwi, I. (2015). Kriteria Empirik dalam Menentukan Ukuran Sampel Pada Pengujian Hipotesis Statistika dan Analisis Butir. *Formatif: Jurnal Ilmiah Pendidikan MIPA, 2*(2), 140–148. https://doi.org/10.30998/formatif.v2i2.95

Amalia, R. N., Dianingati, R. S., & Annisaa', E. (2022). Pengaruh Jumlah Responden terhadap Hasil Uji Validitas dan Reliabilitas Kuesioner Pengetahuan dan Perilaku Swamedikasi. *Generics: Journal of Research in Pharmacy*, *2*(1), 9–15. https://doi.org/10.14710/genres.v2i1.12271

Arifin, Z. (2017). Kriteria Instrumen Dalam Suatu Penelitian. *Jurnal Theorems (the Original Research of Mathematics), 2*(1), 28–36.

Busnawir. (2018). Pengaruh Model Penskoran Terhadap Kestabilan Reliabilitas Skala Sikap Dengan Mempertimbangkan Variasi Usia Responden. *Jurnal Pembelajaran Matematika*, *2*(1), 1–10.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Ekolu, S. O., & Quainoo, H. (2019). Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods. *Global Journal of Engineering Education*, *21*(1), 24–29.

Gambrari, I. A., Yusuf, M. O., & Thomas, D. A. (2015). Effects of Computer-Assisted STAD, LTM and ICI Cooperative Learning Strategies on Nigerian Secondary School Students' Achievement, Gender and Motivation in Physics. *Journal of Education and Practice*, *6*(19), 16–28. www.iiste.org

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Gunartha, I. W. (2022). Estimasi Kesalahan Pengukuran Dalam Bidang Pendidikan Berdasarkan Teori Tes Klasik. *Jurnal Widyadari*, *23*(1), 34–47. https://doi.org/10.5281/zenodo.6390889

Imania, K. A., & Bariah, S. K. (2019). Rancangan Pengembangan Instrumen Penilaian Pembelajaran Berbasis Daring. *Jurnal Petik*, *5*(1), 31–47. https://doi.org/10.31980/jpetik.v5i1.445

Khaerudin. (2015). Kualitas instrumen hasil belajar. *Jurnal Madaniyah, 2*, 212–235.

Listiyandini, R. A., Nathania, A., Syahniar, D., Sonia, L., & Nadya, R. (2020). Mengukur rasa syukur: Pengembangan model awal Skala Bersyukur versi Indonesia. *Jurnal Psikologi Ulayat*, *2*(2), 473–496. https://doi.org/10.24854/jpu39

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan. *BINTANG : Jurnal Pendidikan Dan Sains*, *3*(2), 198–214. https://ejournal.stitpn.ac.id/index.php/bintang

Mustafa, P. S., & Masgumelar, N. K. (2022). Pengembangan Instrumen Penilaian Sikap, Pengetahuan, dan Keterampilan dalam Pendidikan Jasmani. *Biormatika : Jurnal Ilmiah Fakultas Keguruan Dan Ilmu Pendidikan*, *8*(1), 31–49. https://doi.org/10.35569/biormatika.v8i1.1093

Nani Hanifah. (2014). Perbandingan Tingkat Kesukaran, Daya Pembeda Butir Soal Dan Reliabilitas Tes Bentuk Pilihan Ganda Biasa Dan Pilihan Ganda Asosiasi Mata Pelajaran Ekonomi. *SOSIO E-KONS*, *6*(1), 46.

Ndiung, S., & Jediut, M. (2020). Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(1), 94. https://doi.org/10.25273/pe.v10i1.6274

Nuswowati, M., Binadja, A., Efti, K., & Ifada, N. (2011). Pengaruh validitas dan reliabilitas butir soal ulangan akhir semester bidang studi Kimia terhadap pencapaian kompetensi. *Jurnal Inovasi Pendidikan Kimia*, *4*(1), 566–573.

Pramuaji, K., & Loekmono, A. (2018). Uji Validitas Dan Reliabilitas Alat Ukur Penelitian : Questionnaire Empathy. *Jurnal Ilmiah Bimbingan Konseling Undiksha*, *9*(2), 74–78. https://doi.org/10.23887/jibk.v9i2.18009

Pujiastuti, E., Zahra, A. N., & Utami, N. (2023). Analisis Kualitas Aplikasi Olstorage Menggunakan Metode WebQual 4.0 pada Divisi PPL PT. MNC Play. *Jurnal Ilmiah*

ILKOMINFO - Ilmu Komputer & Informatika, *6*(1), 33–44. https://doi.org/10.47324/ilkominfo.v6i1.157

Putri, D., & Nahadi. (2019). Perbandingan Reliabilitas Tes Hasil Belajar Matematika Sma Berdasarkan Teknik Penskoran Dan Ukuran Sampel. *Journal Education and Chemistry (JEDCHEM)*, *1*(1), 10–24.

Rapono, M., Safrial, S., & Wijaya, C. (2019). Urgensi Penyusunan Tes Hasil Belajar: Upaya Menemukan Formulasi Tes Yang Baik dan Benar. *Jupiis: Jurnal Pendidikan Ilmu-Ilmu Sosial*, *11*(1), 95. https://doi.org/10.24114/jupiis.v11i1.12227

Retnawati, H. (2016). *Heri Retnawati 9 786021 547984*.

Rosita, E., Hidayat, W., & Yuliani, W. (2021). Uji Validitas Dan Reliabilitas Kuesioner Perilaku Prososial. *FOKUS (Kajian Bimbingan & Konseling Dalam Pendidikan)*, *4*(4), 279. https://doi.org/10.22460/fokus.v4i4.7413

Sari, A. Q., Sukestiyarno, Y., & Agoestanto, A. (2017). Batasan Prasyarat Uji Normalitas dan Uji Homogenitas pada Model Regresi Linear. *Unnes Journal of Mathematics*, *6*(2), 168–177. http://journal.unnes.ac.id/sju/index.php/ujm

Sari, I. K., Fajri, N., & Mulyani, S. (2019). Profil Validitas Dan Reliabilitas Butir Soal Matematika Ujian Akhir Semester Kelas VIII SMP di Banda Aceh. *Jurnal Numeracy*, *6*(1), 132–142.

Sartika, E. (2018). Analisis Metode K Nearest Neighbor Imputation (KNNI) untuk Mengatasi Data Hilang Pada Estimasi Data Survey. *Jurnal TEDC*, *12*(3), 219–227.

Sarwiningsih, R. (2017). The Comparison Accuracy Estimation of Test Reliability Coefficients for National Chemistry Examination in Jambi Province on Academic Year 2014/2015. *JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *2*(1), 34. https://doi.org/10.20961/jkpk.v2i1.8740

Setiawati, F. A., Mardapi, D., & Azwar, S. (2013). Penskalaan Teori Klasik Instrumen Multiple Intelligences Tipe Thurstone Dan Likert. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 259–274. https://doi.org/10.21831/pep.v17i2.1699

Setiyawan, A. (2014). Reliabilitas Tes. *Jurnal An Nûr*, *VI*(2), 341–354.

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirasat: Jurnal Manajemen & Pendidikan Islam*, *2*(2), 192–213. www.depdiknas.go.id/evaluasi-proses-

Sumintono, B., & Widhiarso, W. (2015). Penilaian Pendidikan dan Ujian. *AplikAsi RascH PemodelAn Pada Assessment Pendidikan*, 1–4.

Sutriani, S., Sukmawati, S., & Rukli, R. (2021). Pengembangan instrumen tes hasil belajar matematika berbasis pendekatan kontekstual siswa kelas IV sekolah dasar wilayah II marioriwawo kabupaten soppeng. *Delta-Pi: Jurnal Matematika Dan Pendidikan Matematika*, *10*(1), 1–20. https://doi.org/10.33387/dpi.v10i1.2559

Syamsuryadin, S., & Wahyuniati, C. F. S. (2017). Tingkat Pengetahuan Pelatih Bola Voli Tentang Program Latihan Mental Di Kabupaten Sleman Yogyakarta. *Jorpres (Jurnal Olahraga Prestasi)*, *13*(1), 53–59. https://doi.org/10.21831/jorpres.v13i1.12884

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Wahyu Widayati, C. S. (2013). Komparasi Beberapa Metode Estimasi Kesalahan Pengukuran. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *13*(2), 182–197. https://doi.org/10.21831/pep.v13i2.1409

Widodo, P. B. (2006). Reliabilitas Dan Validitas KonstrukSkala Konsep Diri Untuk Mahasiswa Indonesia. *Jurnal Psikologi Universitas Diponegoro*, *3*(1), 1–9.