

PENGARUH JUMLAH SAMPEL DAN PANJANG TES TERHADAP VALIDITAS KONSTRUK PADA TES PENALARAN MATEMATIKA

Muhammad Ali Gunawan

Dosen tetap Program Studi Pendidikan Matematika Universitas Pekalongan

guns12380@gmail.com

ABSTRAK

Kajian ini bertujuan untuk mengetahui pengaruh jumlah sampel dan panjang tes terhadap validitas dan reliabilitas butir tes menggunakan analisis faktor (*exploratory factor analysis*) dan koefisien alpha. Data yang digunakan merupakan data bangkitan dengan jumlah sampel dan butir berbeda-beda dengan pasangan jumlah sampel dan butir 1) (120: 15); (120:10); (120: 5); 2) (250:15); (250:10); (250:5); 3) (500: 15); (500:10); (500: 5). Data dibangkitkan dengan program WinGen dengan bentuk tes uraian (politomus) dimana penskorannya menggunakan rubrik (skala 0 sampai 4). Data dibangkitkan sebanyak 25 kali ulangan untuk semua pasangan perlakuan. Untuk mengetahui pengaruh dan perbedaan perlakuan terhadap rata-rata koefisien validitas dan rata-rata koefisien reliabilitas digunakan analisis regresi dan *posthoc analysis*. Hasil kajian ini menunjukkan bahwa: (1) Jumlah sampel dan panjang tes berpengaruh terhadap koefisien validitas tes penalaran matematik sebesar 82,30% dengan persamaan garis regresi: $\hat{y} = 0,678 + 0,021x_1 + 0,080x_2$. Perlakuan dengan jumlah sampel 500 dan butir 15 menghasilkan tingkat koefisien validitas yang tinggi dibandingkan dengan dua perlakuan lainnya. (2) Jumlah sampel dan panjang tes berpengaruh terhadap koefisien reliabilitas tes penalaran matematik sebesar 88,30% dengan persamaan garis regresi: $\hat{y} = 0,617 + 0,098x_1 + 0,013x_2$. Perlakuan dengan jumlah sampel 500 dan butir 15 menghasilkan tingkat koefisien reliabilitas yang tinggi dibandingkan dengan dua perlakuan lainnya.

Kata Kunci: Sampel; Tes; Validitas Konstruk; Reliabilitas; Tes Penalaran Matematik

1. PENDAHULUAN

Permasalahan validitas dan reliabilitas menjadi masalah yang utama dan pertama ketika guru atau peneliti menyusun instrumen penilaian. Tidak valid dan tidak reliabelnya alat ukur yang digunakan akan berdampak kepada kesalahan pengambilan keputusan yang sangat fatal. Oleh karena itu, dalam pelaksanaan penilaian pembelajaran matematika guru tidak bisa hanya sekedar mengadopsi soal-soal yang sudah ada dalam buku teks atau dari sumber lainnya tanpa melakukan pengujian terhadap kualitas soal tersebut.

Kegagalan pencapaian hasil belajar matematika secara maksimal sesuai dengan tujuan kurikulum disebabkan karena penilaian dan pengukuran hasil belajar yang belum didasarkan pada pengintegrasian pemahaman konsep dari berbagai materi yang telah diajarkan dan lebih cenderung diukur secara parsial, sebagian besar butir soal hanya mengungkap kemampuan menghafal peserta didik, dimana butir soal sesuai dengan apa yang dicontohkan dalam buku teks, sehingga informasi yang diperoleh guru mengenai

perkembangan siswanya (strategi pemecahan masalah, penalaran matematik, dan lain sebagainya) kurang memadai Romberg [1]. Akibatnya terjadi kesalahan pengukuran yang bersifat acak dan sistemik Djemari Mardapi [2]. Kesalahan acak bisa disebabkan oleh beberapa hal diantaranya: 1) kondisi fisik dan mental yang diukur bervariasi, dan 2) pemilihan materi pengukuran. Sedangkan kesalahan sistemik disebabkan oleh alat ukurnya, yang diukur dan yang mengukur.

Permasalahan yang sering dihadapi guru matematika dalam penyusunan soal adalah (1) bagaimana menyusun soal yang mampu memberikan informasi yang akurat mengenai kesulitan belajar, kecemasan belajar, kemampuan siswa dalam menyelesaikan masalah-masalah yang tidak biasa, seperti berpikir kreatif, penalaran matematik, penyelesaian masalah, dan sebagainya; (2) dalam praktek pengujian soal yang akan digunakan, pertanyaan yang sering muncul adalah berapa banyak sampel dan jumlah butir tes yang harus digunakan agar tes yang dihasilkan memenuhi kriteria soal yang berkualitas.

Dengan demikian, tujuan dari penelitian ini adalah untuk mengetahui : 1) pengaruh jumlah sampel dan panjang tes terhadap validitas konstruk tes penalaran matematik, 2) pengaruh jumlah sampel dan panjang tes terhadap tingkat reliabilitas tes penalaran matematik.

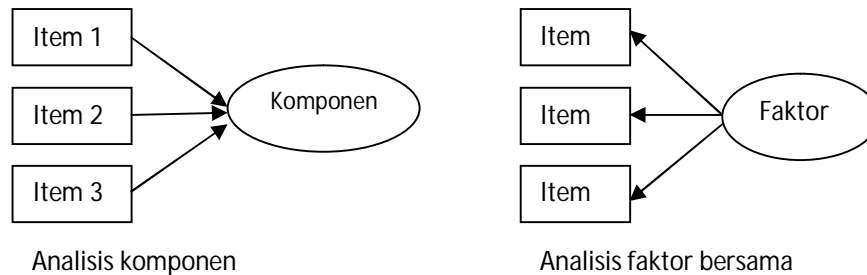
Analisis Faktor

Analisis faktor adalah salah satu metode statistik yang digunakan untuk menguji hubungan antar kelompok pada variabel-variabel yang diamati yang diperoleh melalui pertanyaan atau butir berdasarkan teori tentang konstruk laten yang diukur. Jika teori tentang konstruk yang akan diukur tidak terlalu kuat menjelaskan komponen atau faktor apa saja yang menyusun konstruk yang dimaksud, maka dilakukan *exploratory factor analysis* (EFA), sebaliknya menggunakan CFA untuk membuktikan apakah faktor-faktor hasil dari EFA atau teori memang benar secara struktural dan fungsional membangun konstruk yang hendak diukur.

Proses yang dilakukan dalam analisis faktor adalah proses siklis secara berkelanjutan sampai ditemukannya solusi yang paling bermakna Tabachnick & Fidell [3]. Asumsi yang digunakan dalam analisis faktor sama dengan asumsi yang digunakan dalam teknik statistik multivariat, yaitu: (1) jumlah sampel besar, (2) linier, (3) tidak terjadi outlier, (4) data kontinu, (5) tidak terjadi multikolinieritas, (6) persentase missing data rendah Pett et al. [4].

Pada umumnya analisis faktor dapat dibedakan menjadi dua yaitu analisis komponen (*Principle Component Analysis*) dan analisis faktor bersama (*Common Factor Analysis*). Komponen analisis bertujuan untuk mereduksi sejumlah besar variabel ke dalam komponen yang lebih kecil atau dapat dikatakan bahwa analisis komponen yaitu metode yang digunakan untuk mereduksi data dari sejumlah besar kumpulan variabel, Costello & Osborne [5]. Secara teori, analisis komponen mengasumsikan bahwa komponen adalah komposit variabel-variabel teramati yang disebabkan oleh skor butir secara individual untuk mendefinisikan komponen DeCoster [6]. Sebaliknya, analisis faktor bersama digunakan untuk eksplorasi yang didasarkan pada konstruk, dimana konstruk tidak bisa diukur secara langsung, namun konstruk direpleksikan melalui butir-butir yang mengukur konstruk tersebut Byrne [7]. Analisis faktor bersama mengasumsikan bahwa skor butir secara

individual adalah hasil yang mendasari faktor atau konstruk DeCoster [8]. Skema berikut menggambarkan perbedaan antara analisis komponen dan analisis faktor bersama.



Untuk menentukan apakah data yang dianalisis layak dianalisis menggunakan analisis faktor, maka perlu dilihat nilai Kaiser-Meyer-Olkin (KMO) yang dihasilkan, berikut ini tabel pedoman untuk menginterpretasi hasil KMO.

Tabel 1.1 Pedoman interpretasi untuk uji Kaiser-Meyer-Olkin (KMO)

KMO	Tingkat Varian Bersama
0,90 – 1,00	Sangat tinggi
0,80 – 0,89	Tinggi
0,70 – 0,79	Sedang
0,60 – 0,69	Cukup
0,50 – 0,59	Rendah
0,00 – 0,49	Tidak ada faktor

Beavers. et.al [9]

Berdasarkan tabel di atas, jika $KMO \geq 0,50$ maka analisis faktor dapat dilanjutkan, analisis, namun apakah butir pertanyaan yang diajukan valid atau tidak, belum bisa dijawab hanya dengan menggunakan nilai KMO yang diperoleh. Untuk mengetahui apakah butir-butir pertanyaan valid atau tidak dilihat dari nilai ukuran kecukupan sampel yaitu *Measure of Sampling Adequacy* (MSA), dengan ketentuan jika $MSA \geq 0,50$ maka butir atau item pertanyaan yang diajukan dikatakan valid dan sebaliknya, nilai MSA ditunjukkan oleh nilai diagonal *anti-image correlation* pada tabel *Anti-Image Matrices* Setelah mengetahui butir mana yang valid dan butir yang dibuang/diperbaiki, penyusun tes masih perlu menjawab pertanyaan tentang berapa faktor/komponen yang terbentuk dan butir-butir manakah yang sebaiknya masuk ke dalam komponen yang mana. Hal ini bisa dijawab dengan melihat tabel *Total Variance Explained*. Banyaknya faktor yang terbentuk ditentukan oleh nilai *total initial eigenvalue*, dengan ketentuan bahwa faktor terbentuk jika nilai *eigenvalue* ≥ 1 . Besarnya kontribusi komponen dalam menjelaskan konstruk yang diukur dapat dilihat pada nilai kumulatif %varian yang dihasilkan.

Validitas dan Reliabilitas

Validitas dan reliabilitas alat ukur (tes) adalah dua hal yang menjadi fokus dalam penyusunan tes yang berkualitas, dimana tes dikatakan valid jika tes tersebut mampu mengukur apa yang hendak diukur dan tes dikatakan reliabel (ajeg) jika tes tersebut dapat digunakan untuk mengukur konstruk yang sama pada tempat yang berbeda-beda.

Validitas instrumen terdiri dari 3 jenis, yaitu: (a) validitas isi, mengacu kepada sejauh mana tes dapat mengukur isi domain yang merupakan isi materi yang akan diukur (misalnya tes prestasi belajar matematika pada materi aljabar), (b) validitas kriteria, mengacu pada sejauh mana suatu tes secara empiris terkait variabel kriteria yang relevan, yang dapat dinilai pada waktu yang sama (validitas bersamaan), atau validitas prediktif, dan (c) validitas konstruk, mengacu pada sejauh mana tes mengukur konstruk yang hendak diukur (misalnya kecemasan dalam mengikuti ujian nasional mata pelajaran matematika, penalaran matematik, motivasi belajar dan lain sebagainya) Nunnally & Bernstein [10].

Validitas konstruk merupakan salah satu konsep yang paling penting dalam semuapsikologi. Ini adalah jantung dari setiap studi di mana peneliti menggunakannya untuk mengukur indeks variabel yang tidak bisa secara langsung diamati (Misalnya, kecerdasan, agresi, memori kerja, penalaran). Jika tes hasil belajar dan tes (atau, lebih luas, prosedur psikologis, termasuk tes hasil belajar) tidak memiliki validitas konstruk, hasil yang diperoleh menggunakan tes akan sulit untuk ditafsirkan.

Penalaran Matematika

Penalaran adalah proses menarik kesimpulan yang didasarkan pada bukti atau asumsi lain. Penalaran menurut defnisi ini merupakan hal yang sangat penting pada berbagai disiplin ilmu termasuk matematika. Penalaran dalam matematika memainkan peran khusus dan sangat mendasar, yang di dalamnya mencakup penalaran formal atau pembuktian, di mana kesimpulan secara logis disimpulkan dari asumsi dan definisi NCTM [11].

Ada dua jenis utama dari penalaran matematika, yaitu penalaran matematika yang bersifat kreatif dan penalaran yang bersifat meniru. Penalaran dikatakan kreatif, jika memenuhi persyaratan sebagai berikut: (1) *kebaruan*, membuat urutan solusi penalaran yang baru (bagi penalar) atau mengabaikan urutan solusi yang sebelumnya sudah ada; (2) *masuk akal*. Ada argumen yang mendukung pemilihan strategi dan/atau pelaksanaan strategi, mendukung mengapa kesimpulan itu benar atau masuk akal; dan (3) *berbasis matematika* Boesen, Lithner, dan Palm [12]. Argumentasi ini didasarkan pada sifat intrinsik komponen yang terdapat dalam penalaran matematika.

Penalaran imitatif, berbeda dengan penalaran yang sifatnya kreatif sebagaimana diungkap sebelumnya, dimana jenis penalaran ini sifatnya superfisial, dari sudut pandang penalaran matematika. Penalaran lebih ditekankan kepada kemampuan mengingat strategi jawaban dan menulis jawaban tanpa pertimbangan lain, Lithner [13]. Penalaran ini diklasifikasikan sebagai penalaran algoritmik ketika suatu strategi digunakan kembali pada algoritma tertentu (seperangkat aturan). Strategi ini digunakan pada aturan yang sangat

mudah sekali diingat. Pilihan strategi dalam penalaran algoritmik didasarkan pada tugas-tugas atau latihan soal yang begitu akrab yang diberikan oleh guru dengan penyelesaian yang sesuai dengan apa yang telah dicontohkan.

Kemampuan penalaran matematika bisa dilihat dari: (1) kemampuan menganalisis masalah, (2) menerapkan strategi, (3) mencari dan menggunakan keterkaitan antar domain matematika dari berbagai konteks yang berbeda, (4) merefleksi solusi pada masalah yang dihadapi, NCTM [14]. Kemampuan penalaran matematika ini dipengaruhi oleh faktor: 1) kompeten dalam menggunakan proses-proses penemuan secara matematik, 2) kepercayaan diri dalam mengelola emosi dan masalah-masalah psikologis dan menjadikannya sebagai nilai tambah, 3) pemahaman terhadap isi matematika, terutama pada ranah terapannya, Mason, Burton dan Stacey [15].

Penalaran dalam matematika terdiri dari berbagai bentuk, mulai dari penjelasan informal dan justifikasi untuk deduksi formal, serta pengamatan induktif. Penalaran sering dimulai dengan eksplorasi, perkiraan di berbagai tingkatan, permulaan yang keliru, dan penjelasan parsial sebelum hasilnya tercapai, NCTM [16]. Dengan demikian penalaran diperlukan dalam hal analisis, evaluasi, dan melakukan hal-hal yang kreatif, terutama sekali dalam konteks berpikir tingkat tinggi. Keterampilan penalaran umum meliputi: menilai apakah fakta tunggal atau suatu pendapat benar dan apakah itu relevan dengan argumen pada masalah yang dihadapi, dan menilai apakah dua atau lebih sesuatu berhubungan secara konsisten. Keterampilan ini diperlukan untuk semua jenis penalaran, Brookhart [17].

Sedangkan bila ditinjau dari proses pencapaian kesimpulannya, penalaran secara umum dibagi menjadi dua yaitu penalaran deduktif dan penalaran induktif. Yang dimaksudkan dengan penalaran deduktif adalah proses mencapai kesimpulan dengan menerapkan asumsi, prosedur, atau prinsip-prinsip umum. Sebaliknya, penalaran induktif adalah proses mencapai kesimpulan dengan menguji contoh-contoh khusus, Aufman. et.al, [18] dan 6). Langkah penyimpulan (inferensi) dapat dikatakan valid secara deduktif hanya jika premis-premis yang digunakan adalah benar, maka kesimpulan secara mutlak dijamin benar. Sedangkan, penalaran yang dilakukan dengan menguji kasus sebelumnya kepada kasus baru dan hasilnya dibuat generalisasi atas kasus-kasus serupa dinamakan dengan penalaran induktif, Smith [19].

2. METODE PENELITIAN

Jenis penelitian ini adalah penelitian simulasi dengan pendekatan kuantitatif desain eksperimen 3x3 faktorial (2 perlakuan) yaitu panjang tes (5; 10; 15) dan jumlah sampel (120; 250; 500). Data yang dianalisis dalam penelitian ini merupakan data simulasi yang dibangkitkan dengan program WinGen $N(0;1)$, skor politomus dengan 5 kategori jawaban (0, 1, 2, 3, 4) dengan tingkat kesukaran diatur dari -4 sampai dengan 4, model yang digunakan adalah model parsial kredit (PCM). Data yang dibangkitkan sebanyak 25 kali ulangan pada setiap pasangan perlakuan.

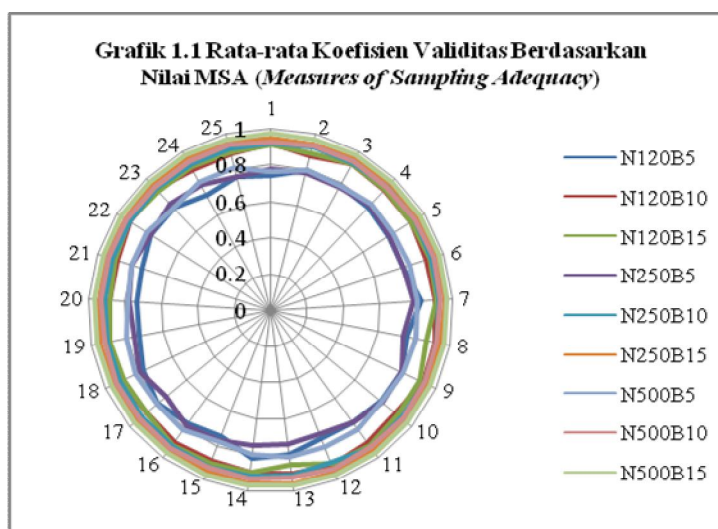
Prosedur yang dilakukan dalam penelitian ini yaitu: 1) mengubah skala data hasil bangkitan (dari skala ordinal ke interval) menggunakan *method of successive interval (MSI)*, 2) menganalisis data dengan *Exploratory Factor Analysis (EFA)* menggunakan program SPSS 12 for Windows dan *Confirmatory Factor Analysis (CFA)* menggunakan program Lisrel 8.80. 3) setelah menemukan koefisien validitas dan reliabilitas menggunakan rumus

alpha Cronbach, analisis dilanjutkan dengan analisis Posthoc (uji t Scheffe) untuk mengetahui perbedaan mean koefisien validitas dan reliabilitas yang diperoleh.

3. HASIL PENELITIAN DAN PEMBAHASAN

1) Validitas Tes Penalaran Matematik

Pembuktian validitas kontruk tes matematika menggunakan *Exploratory Factor Analysis* (EFA) membuktikan bahwa sebagian besar treatmen pada percobaan yang dilakukan sebanyak 25 kali ulangan termasuk ke dalam kategori valid, satu pasangan perlakuan ada yang mengandung butir yang tidak valid (maksimal 2 butir drop) yaitu perlakuan dengan N120B15 butir nomor 10 ulangan ke 13 dan N120B15 butir nomor 10 ulangan ke 17 . Ringkasan hasil penelitian simulasi ini dapat dilihat pada garfik berikut.



Berdasarkan grafik di atas, dapat dilihat bahwa rata-rata nilai MSA semakin tinggi dengan jumlah sampel dan panjang tes (jumlah butir) yang semakin banyak. Namun, hal ini belum bisa menjawab seberapa besar pengaruh jumlah sampel dan panjang tes terhadap koefisien validitas serta pasangan perlakuan (treatment) manakah yang paling baik dalam menghasilkan koefisien validitas yang tinggi. Untuk itu dilakukan analisis lanjut menggunakan analisis regresi dan *posthoc analysis*.

Hasil analisis regresi menunjukkan bahwa panjang tes lebih berpengaruh daripada jumlah sampel, dimana panjang tes mampu memprediksi tingkat koefisien validitas sebesar 77% dengan persamaan regresi : $\hat{y} = 0,720 + 0,080x$, jumlah sampel hanya memberikan kontribusi sebesar 5,3% dengan persamaan regresi $\hat{y} = 0,839 + 0,021x$. Sedangkan pengaruh secara bersama-sama (jumlah sampel dan panjang tes) terhadap koefisien validitas sebesar 82,30% dengan persamaan garis regresi: $\hat{y} = 0,678 + 0,021x_1 + 0,080x_2$.

Selanjutnya uji *posthoc analysis* berikut menunjukkan pasangan perlakuan manakah yang terbaik dalam memprediksi tingkat koefisien validitas, hal ini dapat dilihat pada tabel *multiple comparisons* berikut.

Tabel 1.2 Perbedaan nilai koefisien validitas berdasarkan jumlah sampel yang digunakan.

Dependent Variable: KOEF. VALIDITAS

(I) SAMPEL	(J) SAMPEL	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval
					Lower Bound
1	2	-,0239 [*]	,00231	,000	-,0297
	3	-,0422 [*]	,00231	,000	-,0479
2	1	,0239 [*]	,00231	,000	,0182
	3	-,0182 [*]	,00231	,000	-,0239
3	1	,0422 [*]	,00231	,000	,0365
	2	,0182 [*]	,00231	,000	,0125
		,0182 [*]	,00231	,000	,0127

Dari tabel di atas, diketahui bahwa tingkat koefisien validitas berturut-turut mulai dari yang terbaik adalah perlakuan dengan jumlah sampel 500 dengan perbedaan rata-rata sebesar 0,0182 dengan jumlah sampel 250, dan 0,0422 dengan jumlah sampel 120. Sedangkan perbedaan jumlah sampel 250 dan 120 memberikan perbedaan rata-rata koefisien validitas sebesar 0,0239. Artinya bahwa rata-rata koefisien validitas yang terbaik diperoleh dengan jumlah sampel sebesar 500. Namun, perbedaan yang ditampilkan sangat kecil kurang dari 0,05. Hal ini mengindikasikan bahwa hubungan koefisien validitas dengan jumlah sampel tidak terlalu kuat sebagaimana ditunjukkan pada tabel korelasi berikut.

Tabel 1.3 Hubungan jumlah sampel dan panjang tes terhadap koefisien validitas tes

		KOEF. VALIDITAS	SAMPEL	TES
Pearson Correlation	KOEF. VALIDITAS	1,000	,230	,878
	SAMPEL	,230	1,000	,000
	TES	,878	,000	1,000
Sig. (1-tailed)	KOEF. VALIDITAS	.	,000	,000
	SAMPEL	,000	.	,500
	TES	,000	,500	.
N	KOEF. VALIDITAS	225	225	225
	SAMPEL	225	225	225
	TES	225	225	225

Selanjutnya, perbedaan koefisien validitas berdasarkan panjang tes diperoleh hasil bahwa jumlah butir 15 menghasilkan koefisien validitas yang tinggi bila dibandingkan dengan tes dengan jumlah butir 10 (0,237) ataupun jumlah butir 5 (0,1609), untuk lebih jelasnya dapat dilihat pada tabel berikut.

Tabel 1.4 Perbedaan nilai koefisien validitas berdasarkan panjang tes.

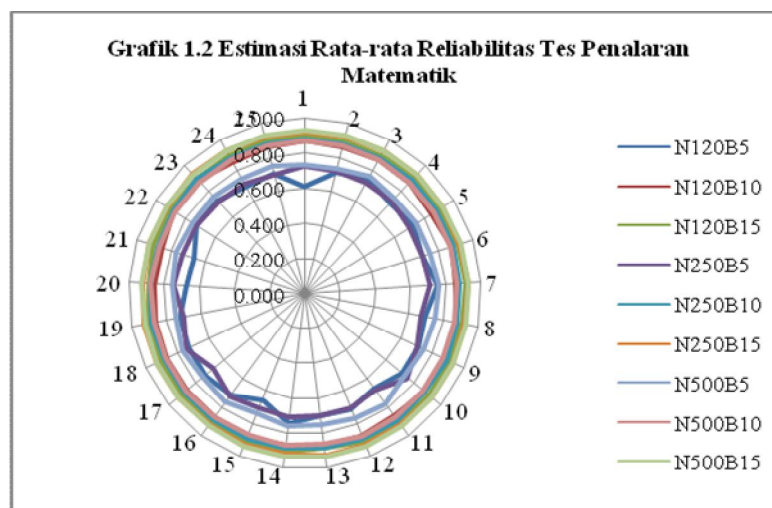
Dependent Variable: KOEF. VALIDITAS

Scheffe

(I) TES	(J) TES	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,1372 [*]	,00231	,000	-,1429	-,1315
	3	-,1609 [*]	,00231	,000	-,1666	-,1552
2	1	,1372 [*]	,00231	,000	,1315	,1429
	3	-,0237 [*]	,00231	,000	-,0294	-,0180
3	1	,1609 [*]	,00231	,000	,1552	,1666
	2	,0237 [*]	,00231	,000	,0180	,0294

2) Reliabilitas tes penalaran matematik

Rata-rata tingkat reliabilitas tes berdasarkan hasil analisis menggunakan rumus alpha diketahui bahwa semakin banyak jumlah sampel dan panjang tes akan semakin tinggi tingkat reliabilitas tes. Hal ini ditunjukkan oleh grafik berikut.



Panjang tes dalam hal ini memiliki hubungan yang lebih kuat dengan koefisien reliabilitas dibanding dengan jumlah sampel, dimana panjang tes dan koefisien reliabilitas memiliki koefisien korelasi sebesar 0,932 dan jumlah sampel dengan koefisien reliabilitas sebesar 0,121. Artinya semakin banyak butir tes yang digunakan akan semakin tinggi tingkat reliabilitas tes tersebut, namun semakin banyak jumlah sampel yang digunakan belum tentu menghasilkan koefisien reliabilitas yang tinggi, sebagaimana tampak pada tabel berikut.

Tabel 1.5 Hubungan tingkat koefisien reliabilitas dengan jumlah sampel dan panjang tes.

		KOEF. VALIDITAS	TEST	SAMPEL
Pearson Correlation	KOEF. VALIDITAS	1,000	,932	,121
	TEST	,932	1,000	,000
	SAMPEL	,121	,000	1,000
Sig. (1-tailed)	KOEF. VALIDITAS	.	,000	,035

	TEST	,000	.	,500
	SAMPEL	,035	,500	.
	KOEF. VALIDITAS	225	225	225
N	TEST	225	225	225
	SAMPEL	225	225	225

Temuan ini diperkuat dengan melihat kontribusi jumlah sampel dan panjang tes melalui analisis regresi, dimana jumlah sampel mampu memprediksi tingkat koefisien reliabilitas sebesar 1,5% dengan persamaan garis regresi : $\hat{y} = 0,814 + 0,013x$ dan panjang tes mampu memprediksi tingkat koefisien reliabilitas sebesar 86,80% dengan persamaan garis regresi: $\hat{y} = 0,643 + 0,098x$. Secara bersama-sama, jumlah sampel dan panjang tes mampu memprediksi koefisien reliabilitas sebesar 88,30% dengan persamaan garis regresi: $\hat{y} = 0,617 + 0,098x_1 + 0,013x_2$.

Selanjutnya, perlakuan jumlah sampel dan panjang tes manakah yang lebih baik dalam menghasilkan koefisien reliabilitas yang tinggi? Hal ini dapat dijawab dengan menggunakan hasil uji beda dengan *posthoc analysis* sebagaimana tampak pada tabel *Multiple Comparisons* berikut.

Tabel 1.6 Perbedaan tingkat koefisien reliabilitas berdasarkan panjang tes.

Dependent Variable: KOEF. RELIABILITAS

(I) TEST	(J) TEST	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound
1	2	-,14956	,002305	,000	-,15524
	3	-,19635*	,002305	,000	-,20203
Scheffe 2	1	,14956	,002305	,000	,14388
	3	-,04679*	,002305	,000	-,05247
3	1	,19635*	,002305	,000	,19066
	2	,04679*	,002305	,000	,04110

Dari tabel di atas, diketahui bahwa koefisien reliabilitas yang paling tinggi diperoleh pada perlakuan 3 (panjang tes 15 butir), perbedaan yang terjadi sangat signifikan dengan perlakuan 2 (panjang tes 10 butir dan 5 butir) yaitu 0,4679 dan 0,19635. Begitu juga bila dilihat dari jumlah sampel yang digunakan, jumlah sampel pada perlakuan 3 (sampel sebanyak 500 orang) memberikan hasil tingkat koefisien reliabilitas yang paling tinggi daripada perlakuan lainnya yaitu berbeda dengan perlakuan 2 sebesar 0,01149 dan perlakuan 1 sebesar 0,02556 sebagaimana diperlihatkan pada tabel berikut.

Tabel 1.7 Perbedaan tingkat koefisien reliabilitas berdasarkan jumlah sampel.

Dependent Variable: KOEF. RELIABILITAS

(I) SAMPEL	(J) SAMPEL	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound
Scheffe 1	2	-,01407	,002305	,000	-,01975
	3	-,02556*	,002305	,000	-,03124
2	1	,01407	,002305	,000	,00838

	3	-.01149 [*]	,002305	,000	-.01718
3	1	,02556 [*]	,002305	,000	,01988
	2	,01149 [*]	,002305	,000	,00581

3) Keterbatasan Penelitian

Penelitian simulasi yang dilakukan ini masih mengandung keterbatasan dalam banyak hal, diantaranya adalah: (1) Analisis yang dilakukan masih sebatas analisis faktor eksploratory dengan *principle component analysis* (PCA), dimana jumlah sampel besar masih bisa diterima, namun perlu dikaji kembali persoalan pembuktian hasil EFA dengan CFA, dimana dengan CFA hasil yang diperoleh melalui EFA seringkali ditolak (tidak fit) terlebih dengan jumlah sampel besar, hal ini dikarenakan metode *chi-square* sangat sensitif dengan jumlah sampel; (2) Asumsi yang digunakan dalam kajian ini adalah asumsi unidimensi, yaitu tes hanya mengukur satu konstruk yaitu penalaran matematik, sehingga kajian ini belum mampu menjawab realitas bahwa karakteristik tes pada dasarnya multidimensi. Pertanyaan spesifiknya yaitu bagaimana pengaruh jumlah sampel dan panjang tes terhadap dimensionalitas tes? (3) Hasil kajian ini perlu dikaitkan dengan metode pembuktian validitas yang lain seperti: validitas isi dan validitas kriteria. Karena bila hanya mengacu pada validitas konstruk sesuai dengan hasil kajian ini, maka sangatlah keliru jika dikatakan bahwa semakin panjang tes akan semakin valid sebuah tes, karena pada kenyataannya se-valid apapun alat ukur (tes) jika jumlah butirnya terlalu banyak bisa menghasilkan bias, dimana peserta tes mengalami kelelahan dan kejenuhan.

D. SIMPULAN

Berdasarkan hasil kajian simulasi ini dapat disimpulkan bahwa: (1) Jumlah sampel dan panjang tes berpengaruh terhadap koefisien validitas tes penalaran matematik sebesar 82,30% dengan persamaan garis regresi: $\hat{y} = 0,678 + 0,021x_1 + 0,080x_2$. Perlakuan dengan jumlah sampel 500 dan butir 15 menghasilkan tingkat koefisien validitas yang tinggi dibandingkan dengan dua perlakuan lainnya. (2) Jumlah sampel dan panjang tes berpengaruh terhadap koefisien reliabilitas te penalaran matematik sebesar 88,30% dengan persamaan garis regresi: $\hat{y} = 0,617 + 0,098x_1 + 0,013x_2$. Perlakuan dengan jumlah sampel 500 dan butir 15 menghasilkan tingkat koefisien validitas yang tinggi dibandingkan dengan dua perlakuan lainnya.

DAFTAR PUSATAKA

- [1] Romberg, Thomas A. (ed.). 2004. *Standards-Based Mathematics Assessment in Middle School: Rethinking Classroom Practice*. New York: Teachers College, Columbia University.
- [2] Djemari Mardapi. 2012. *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta: Nuha Medika.

- [3] Tabachnick, B. & Fidell, L. 2001. *Using multivariate statistics*. Needham Heights: Allyn & Bacon.
- [4] Pett, M., Lackey, N. & Sullivan, J. 2003. *Making sense of factor analysis*. Thousand Oaks: Sage Publications, Inc.
- [5] Costello, A.B. & Osborne, J.W. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10, 1-9. Diambil pada tanggal 12 Desember, 2014 dari <http://pareonline.net/getvn.asp?v=10&n=7>
- [6] DeCoster, J. 1998. *Overview of factor analysis*. diambil tanggal 12 Desember 2014, dari <http://www.stathelp.com/notes.html>
- [7] Byrne, Barbara M. 2010. *Structural Equation Modeling with AMOS Basic Concepts, Applications, and Programming, Second Edition*. New York: Taylor and Francis Group, LLC.
- [8] DeCoster, J. 1998. *Overview of factor analysis*. diambil tanggal 12 Desember 2014, dari <http://www.stathelp.com/notes.html>
- [9] Beavers, Amy S. (et.al). 2013. Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research & Evaluation*, 18. Diambil dari <http://www.google.co.id/search?output=search&client=psy-ab&q=Journal+of+Factor+Analysis+2013.pdf&btnG=>. Tanggal 5 Januari 2015.
- [10] Nunnally, J and Ira. H. Bernstein. 1994. *Psychometric theory*. 3rd ed. New York: MCGraw-Hill.
- [11] National Council of Teacher Mathematics. 2009. *Developing Mathematical Reasoning in Grades K-12: 1999 Yearbook*. Virginia: The National Council Of Teachers Of Mathematics, Inc.
- [12] Boesen, Jesper, Johan Lithner & Torulf Palm. 2010. The Relation Between Types Of Assessment Tasks And The Mathematical Reasoning Students Use. *Educational Studies in Mathematics*, 2010, 75, 89-105.
- [13] Lithner, Johan. 2012. Learning Mathematics by Creative or Imitative Reasoning. 12th *International Congress on Mathematical Education*. Diambil dari http://www.icme12.org/upload/submission/1971_F.pdf. Pada tanggal 01 September 2014
- [14] National Council of Teacher Mathematics. 2009. *Developing Mathematical Reasoning in Grades K-12: 1999 Yearbook*. Virginia: The National Council Of Teachers Of Mathematics, Inc.
- [15] Mason, John., Leone Burton, and Kaye Stacey. 2010. *Thinking Mathematically*. (2th Ed.). England: Pearson Education Limited.
- [16] National Council of Teacher Mathematics. 2009. *Developing Mathematical Reasoning in Grades K-12: 1999 Yearbook*. Virginia: The National Council Of Teachers Of Mathematics, Inc.
- [17] Brookhart, Susan M. 2010. *How to Assess Higher-Order Thinking Skills in Your Classroom*. Virginia: ASCD.
- [18] Aufmann, Richard N. (et.al). 2008. *Mathematical Thinking and Quantitative Reasoning*. Boston: Houghton Mifflin Company.

- [19] Smith, Peter. 2003. *An Introduction to Formal Logic*. New York: Cambridge University Press.