Jurnal AL-Muta`aliyah: Jurnal Pendidikan Guru Madrasah Ibtidaiyah
ISSN 2614-1612 (e) | ISSN 2502-2474 (p) | Vol.05, No.02, (2025), 122 -128 | DOI: 10.51700/mutaaliyah.v5i2.

Analisis Generalizabilitas Dua Faset dalam Penilaian Ujian Skripsi di IAI Nurul Hakim Kediri Lombok Barat

Ahmad[1*], Mariano Dos Santos[2]
[1]*Ilmu Komputer, Fakultas Teknik, Universitas Bumigora Mataram, Indonesia;*
[2]*Instituto Católico para a Formação de Professores (ICFP) Baucau, Timor - Leste*

Abstrak

Penelitian ini mengevaluasi konsistensi dan keadilan penilaian ujian skripsi di IAI Nurul Hakim Kediri Lombok Barat dengan pendekatan Teori Generalisasi *(G-Theory).* Data diperoleh dari penilaian tiga dosen terhadap 13 mahasiswa menggunakan empat kriteria standar kampus. Analisis dengan perangkat lunak *EduG* menunjukkan bahwa sumber error terbesar berasal dari interaksi antara mahasiswa, rater, dan kriteria (PRQ). Koefisien generalisasi sebesar 0,93 menunjukkan reliabilitas penilaian yang sangat baik. Untuk meningkatkan keadilan, disarankan mengurangi variabilitas interaksi antar faset. Simulasi menunjukkan bahwa penambahan jumlah rater dan kriteria dapat meningkatkan keandalan. Opsi 5 menjadi yang paling optimal, namun Opsi 3 dan 4 juga layak dipertimbangkan karena menawarkan keseimbangan antara reliabilitas dan efisiensi sumber daya. Penelitian ini menekankan pentingnya desain penilaian yang tepat untuk menjamin penilaian skripsi yang adil dan konsisten.

Kata kunci: generalizability theory; konsistensi penilaian; variabilitas penilaian

Abstrack

This study aims to evaluate the consistency and fairness of undergraduate thesis assessments at IAI Nurul Hakim Kediri, West Lombok, using Generalizability Theory (G-Theory). Data were collected from the assessments of 13 students by three examiners using four standardized evaluation criteria. Analysis using the EduG software revealed that the largest source of error stemmed from the interaction among persons, raters, and criteria (PRQ). The generalizability coefficient of 0.93 indicates excellent reliability of the assessment instrument. To enhance fairness, efforts should be made to reduce variability in facet interactions. Simulations showed that increasing the number of raters and criteria could improve reliability. While Option 5 yielded the most optimal results, Options 3 and 4 are also worth considering as they offer a balance between reliability and resource efficiency. This study underscores the importance of well-designed assessment systems to ensure fair and consistent undergraduate thesis evaluations.

Keywords: *generalizability theory; assessment consistency; assessment variability*

Pendahuluan

Penilaian akademik memegang peranan sentral dalam menentukan keberhasilan proses pendidikan, terutama di tingkat perguruan tinggi. Evaluasi yang tepat dan objektif tidak hanya memberikan umpan balik penting bagi pengembangan program pendidikan, tetapi juga membantu dalam pengambilan keputusan untuk peningkatan kualitas Pendidikan (Rashid & McGuinness, 2018; Simion, 2023). Di dunia akademis, proses evaluasi menjadi landasan untuk mengenali kontribusi akademik, membangun karir, serta memperkuat batas-batas disiplin ilmu (Hegde & Shushruth, 2022; Prabhavathy, 2023; Simion, 2023). Di Indonesia, ujian skripsi merupakan salah satu bentuk evaluasi akhir yang kritis dalam menilai kompetensi akademik mahasiswa sebelum mereka dinyatakan

[*] Corresponding to the author: Ahmad, Ilmu Komputer, Fakultas Teknik Universitas Bumigora Mataram, Indonesia; e-mail: ahmad@universitasbumigora.ac.id

lulus(Handoyono, 2020; Hsiao et al., 2023). Namun demikian, kualitas penilaian skripsi sering kali dipertanyakan karena berbagai masalah yang timbul dalam praktik penilaiannya, seperti yang terlihat di IAI Nurul Hakim Kediri Lombok Barat, di mana terdapat variabilitas signifikan dalam penilaian yang diberikan oleh dosen penguji.

Salah satu masalah utama yang dihadapi adalah variabilitas antar dosen penguji. Meskipun terdapat pedoman penilaian, interpretasi dan penerapan kriteria tersebut sering kali bervariasi antara satu dosen dengan dosen lainnya. Variasi ini dapat disebabkan oleh perbedaan latar belakang akademik, pengalaman mengajar, dan persepsi subjektif terhadap kualitas karya mahasiswa (Bauer et al., 2002; Duerksen et al., 2021; Tierney, 2022). Akibatnya, penilaian menjadi tidak konsisten dan cenderung tidak adil, yang pada akhirnya dapat merugikan mahasiswa. Penelitian sebelumnya menunjukkan bahwa subjektivitas dalam penilaian dapat mempengaruhi persepsi siswa dan mengurangi reliabilitas hasil penilaian (Bacon et al., 2017). Dosen penguji, meskipun berusaha objektif, sering kali memiliki preferensi dan bias yang mempengaruhi hasil penilaian, seperti hubungan personal antara dosen dan mahasiswa, penampilan saat presentasi, dan persepsi terhadap relevansi serta keunikan topik skripsi (Azizah et al., 2021).

Selain itu, instrumen penilaian yang digunakan sering kali tidak mampu menangkap seluruh aspek penting dari kompetensi mahasiswa. Banyak instrumen penilaian yang belum dapat mengevaluasi aspek-aspek seperti kemampuan analisis kritis, kreativitas, dan penerapan teori ke dalam praktik dengan memadai (Helmanda et al., 2022; Thomas et al., 2022). Akibatnya, penilaian yang diberikan tidak memberikan gambaran yang utuh mengenai kemampuan akademik mahasiswa. Masalah lain yang dihadapi adalah tekanan eksternal dan institusional yang mempengaruhi kualitas penilaian. Institusi pendidikan sering kali berada di bawah tekanan untuk mempertahankan reputasi akademik dan tingkat kelulusan yang tinggi, yang dapat mempengaruhi dosen penguji dalam memberikan penilaian yang lebih permisif atau lebih ketat, tergantung pada kebijakan dan budaya institusi penilaian (Ilma et al., 2021; Sebhatu & Wennberg, 2023).

Di berbagai institusi pendidikan, termasuk IAI Nurul Hakim Kediri Lombok Barat, muncul permasalahan terkait inkonsistensi dan subjektivitas dalam penilaian ujian skripsi. Variabilitas penilaian antar dosen penguji sering kali menimbulkan keraguan terhadap keadilan dan reliabilitas hasil penilaian, yang dapat merugikan mahasiswa. Selain itu, interpretasi kriteria penilaian yang berbeda-beda oleh dosen penguji juga berpotensi menciptakan ketidakadilan dalam evaluasi akhir mahasiswa. Dalam konteks ini, *Generalizability Theory (G-Theory)* menawarkan pendekatan yang komprehensif untuk mengevaluasi dan meningkatkan konsistensi serta keadilan dalam penilaian akademik (Kim et al., 2022; Ten Hove et al., 2021).

G-Theory merupakan pendekatan statistik yang digunakan untuk menilai keandalan dan validitas alat pengukuran dalam berbagai pengaturan penelitian, termasuk penelitian sosial dan perilaku (Andersen et al., 2021; Kim et al., 2022). Teori ini memungkinkan dekomposisi sumber kesalahan pengukuran seperti komponen terkait subjek, kluster, dan penilai, memberikan pemahaman komprehensif tentang faktor-faktor yang mempengaruhi presisi pengukuran. *G-Theory* telah diterapkan untuk mengevaluasi prosedur penilaian dalam konteks pendidikan, penilaian keterampilan medis, dan alat pengukuran depresi, menunjukkan keserbagunaan dan efektivitasnya dalam berbagai domain (ten Hove et al., 2021). Dengan memanfaatkan *G-Theory*, peneliti dapat mengidentifikasi sumber kesalahan, memperkirakan koefisien reliabilitas, dan meningkatkan kualitas pengukuran, yang pada akhirnya meningkatkan validitas dan kepercayaan temuan penelitian dalam ilmu sosial dan perilaku (Sturgis et al., 2022).

Di samping itu Pendekatan ini juga memungkinkan analisis multifaset dari berbagai sumber varians, memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang mempengaruhi keandalan pengukuran (Andersen et al., 2021; Gorges et al., 2017). Melalui penelitian ini, yang melibatkan dua faset utama yaitu mahasiswa dan dosen penguji, kami bertujuan untuk mengungkap sejauh mana variabilitas penilaian dipengaruhi oleh kedua faset tersebut dan bagaimana perbaikan dapat dilakukan.

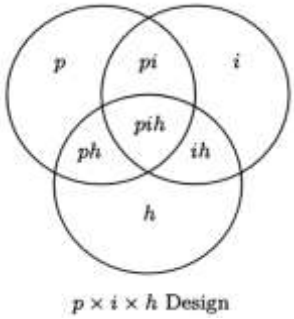
Penelitian ini bertujuan untuk mengevaluasi generalizabilitas penilaian ujian skripsi di IAI Nurul Hakim Kediri Lombok Barat dengan menggunakan pendekatan dua faset, yaitu dosen penguji dan mahasiswa. Melalui penerapan *Generalizability Theory (G-Theory)*, penelitian ini akan mengidentifikasi dan mengukur sumber-sumber variabilitas dalam penilaian serta mengevaluasi konsistensi dan keadilan dari penilaian tersebut. Diharapkan, penelitian ini dapat memberikan kontribusi signifikan dalam meningkatkan kualitas dan keadilan penilaian akademik di institusi pendidikan tinggi.

Metode

Penelitian ini menggunakan pendekatan kuantitatif dengan desain studi eksperimental. Tujuan utama penelitian adalah untuk mengevaluasi generalizabilitas penilaian ujian skripsi di IAI Nurul Hakim Kediri Lombok Barat dengan menggunakan *Generalizability Theory (G-Theory)* (Huebner & Lucht, 2019). Pendekatan ini dipilih karena kemampuannya untuk mengidentifikasi dan mengukur berbagai sumber variabilitas dalam penilaian, serta mengevaluasi reliabilitas dan validitas hasil evaluasi.

Data penelitian dikumpulkan dari penilaian ujian skripsi yang dilakukan oleh dosen penguji di IAI Nurul Hakim. Sampel penelitian terdiri dari tiga belas mahasiswa yang telah menyelesaikan ujian skripsi mereka, serta tiga dosen penguji yang terlibat dalam proses penilaian dengan menggunakan empat kriteria penilaian. Penilaian dilakukan berdasarkan rubrik yang telah distandardisasi, yang mencakup berbagai aspek penting seperti kualitas konten, kemampuan analisis, presentasi, dan penyusunan laporan skripsi.

Analisis data dalam penelitian ini dilakukan dengan menggunakan *(G-Theory)* p x r x q untuk mengevaluasi konsistensi dan keadilan penilaian. *G-Theory* memungkinkan peneliti untuk menguraikan total variabilitas penilaian ke dalam komponen-komponen yang disebabkan oleh faktor dosen penguji, mahasiswa, dan interaksi antara keduanya (Van Hooijdonk et al., 2022). Dalam lingkup Teori G *Two-Facet*, desain p x r x q mengacu pada salah satu rancangan eksperimental yang digunakan untuk mengevaluasi reliabilitas dan validitas pengukuran. Desain ini melibatkan tiga komponen utama yaitu person (p), rater (r), dan kriteria (q). Desain ini dapat ditulis diilustrasikan p x i x h seperti pada gambar 1 berikut.



Gambar 1. p x i x h Design

Analisis dilakukan dengan menggunakan perangkat lunak statistik EduG, untuk menghitung analisis varians (ANOVA), koefisien generalizabilitas (teori G), dan indeks reliabilitas. Selain itu, dilakukan juga analisis dependabilitas *(phi coefficient)* (teori D) untuk menilai sejauh mana hasil penilaian dapat digeneralisasikan ke situasi penilaian lain yang serupa. Studi simulasi dilakukan untuk mengevaluasi efek dari jumlah dosen penguji dan kriteria penilaian terhadap reliabilitas hasil penilaian.

Hasil dan Diskusi

Hasil

Penelitian ini menggunakan *Generalizability Theory (G-Theory)* untuk mengevaluasi reliabilitas penilaian ujian skripsi di IAI Nurul Hakim Kediri Lombok Barat, dengan memperhatikan tiga faset utama: mahasiswa *(P)*, dosen penguji *(R)*, dan kriteria penilaian *(Q)*. Data yang dihasilkan dari analisis varians menunjukkan beberapa temuan penting yang dapat digunakan untuk meningkatkan kualitas penilaian akademik.

a. Varians dan Jumlah Kuadrat

Untuk memahami struktur data yang digunakan dalam penelitian ini, langkah awal yang dilakukan adalah menyusun desain observasi dan estimasi berdasarkan faset-faset yang terlibat dalam proses penilaian. Desain ini mencerminkan konfigurasi kombinasi antara mahasiswa sebagai objek penilaian (person), dosen penguji sebagai pemberi nilai (rater), dan kriteria penilaian yang digunakan (criteria). Informasi ini penting untuk menentukan tingkat kompleksitas pengukuran serta dasar dalam pelaksanaan analisis varians menggunakan *Generalizability Theory (G-Theory)*.

Tabel 1 berikut menyajikan jumlah level pada masing-masing faset serta karakteristik pengamatannya yang akan digunakan sebagai dasar dalam analisis lebih lanjut.

Tabel 1. Desain Observasi dan Estimasi

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSON	P	13	INF	
RATER	R	3	INF	
KRITERIA	Q	4	INF	

Tabel 1 menunjukan bahwa jumlah *person* 13 orang, *rater* 3 orang dan komponen kriteria yang digunakan sama dengan 4 item. Selanjutnya unutk nilai Analisis varian dapar silihat pada tabel 2 di bawah ini.

Tabel 2. Nilai Analisis Varians

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	6431.41026	12	535.95085	41.54647	41.54647	41.54647	62.8	16.90710
R	35.89744	2	17.94872	-0.13800	-0.13800	-0.13800	0.0	0.32528
Q	128.20513	3	42.73504	0.78793	0.78793	0.78793	1.2	0.71939
PR	834.93590	24	34.78900	3.93073	3.93073	3.93073	5.9	2.53627
PQ	780.12821	36	21.67023	0.86806	0.86806	0.86806	1.3	1.95904
RQ	56.41026	6	9.40171	-0.74341	-0.74341	-0.74341	0.0	0.43462
PRQ	1372.75641	72	19.06606	19.06606	19.06606	19.06606	28.8	3.13444
Total	9639.74359	155					100%	

Dari Tabel 2 diketahui bahwa variabilitas terbesar dalam penilaian berasal dari faset person (mahasiswa) (P) dengan sum of squares (SS) sebesar 6431.41026 dan mean square (MS) sebesar 535.95085. Kontribusi variabilitas ini mencapai 62.8%, yang menunjukkan bahwa penilaian sangat dipengaruhi oleh perbedaan individual di antara mahasiswa. Variabilitas ini merupakan hal yang wajar, mengingat bahwa setiap mahasiswa memiliki kemampuan dan kinerja yang berbeda. Faset rater (dosen penguji) (R) menunjukkan SS sebesar 35.89744 dan MS sebesar 17.94872. Variabilitas dari dosen penguji ini relatif kecil, dengan kontribusi hanya sebesar 0.3%. Sementara itu, kriteria penilaian (Q) memiliki SS sebesar 128.20513 dan MS sebesar 42.73504, dengan kontribusi sebesar 1.2%.

Interaksi antara mahasiswa dan dosen penguji (PR) menyumbang variabilitas sebesar 5.9% (SS = 834.93590, MS = 34.78900), sedangkan interaksi mahasiswa dan kriteria penilaian (PQ) serta interaksi dosen penguji dan kriteria (RQ) masing-masing menyumbang 0.3% dan 0.4% (SS = 780.12821 dan 56.41026). Komponen gabungan PRQ memiliki SS terbesar kedua yaitu 1372.75641 dan MS sebesar 19.06606, menyumbang 28.8% dari total variabilitas. Ini menunjukkan bahwa interaksi kompleks antara mahasiswa, dosen penguji, dan kriteria memiliki pengaruh yang signifikan terhadap hasil penilaian.

b. Uji Teori G *Study*

Uji G *study* dilakukan sebagai langkah penting dalam menganalisis sejauh mana masing-masing faset baik *person* (mahasiswa), *rater* (dosen penguji), maupun *kriteria penilaian* berkontribusi terhadap total variabilitas atau kesalahan pengukuran dalam proses penilaian ujian skripsi. Melalui pendekatan ini, peneliti dapat mengidentifikasi sumber error terbesar dari setiap komponen secara sistematis, sehingga dapat diketahui faset mana yang paling memengaruhi ketidakstabilan skor penilaian. Informasi ini sangat krusial untuk menentukan strategi peningkatan reliabilitas instrumen penilaian, baik melalui penguatan instrumen, penambahan jumlah penguji, atau penyempurnaan kriteria. Adapun hasil lengkap dari uji G study disajikan pada tabel 3 berikut.

Tabel 3. Nilai G Study (Measurement design P/RQ)

Source of variance	Differ-Entiation variance	Source Of variance	Relative error variance	% relative	Absolute Error variance	% absolute
P	41.54647	.....	.....		.....	
.....	.....	R	.....		(0.00000)	0.0
.....	.....	Q	.....		0.19698	5.9
.....	.....	PR	1.31024	42.0	1.31024	39.5
.....	.....	PQ	0.21701	7.0	0.21701	6.6
.....	.....	RQ	.....		(0.00000)	0.0
.....	.....	PRQ	1.58884	51.0	1.58884	48.0
Sum of variances	41.54647		3.11610	100%	3.31308	100%
Standard deviation	6.44566		Relative SE:	1.76525	Absolute SE:	1.82019
Coef_G relative	0.93					
Coef_G absolute	0.93					

Pada tabel 3 di atas, dilakukan analisis varians menggunakan desain pengukuran P/RQ untuk mengevaluasi sumber-sumber kesalahan dan mengukur keandalan instrumen penelitian. Tabel G-Study yang diperoleh menunjukkan berbagai komponen varians yang memberikan gambaran tentang kontribusi setiap sumber varians terhadap total varians.

Dari tabel tersebut diketahui bahwa varians utama berasal dari subjek (P) dengan nilai 41.54647. Komponen ini menyumbang sebagian besar varians total. Selanjutnya, komponen

varians untuk interaksi antara pengamat dan subjek (PR) tercatat sebesar 1.31024, yang menyumbang 42% dari total varians relatif dan 39.5% dari total varians absolut.

Varians untuk interaksi antara pengamat dan kriteria penilaian (PQ) sebesar 0.21701, dengan kontribusi 7% relatif dan 6.6% absolut. Sementara itu, komponen varians gabungan dari interaksi pengamat, kriteria, dan subjek (PRQ) tercatat sebesar 1.58884, memberikan kontribusi terbesar kedua, yaitu 51% dalam hal varians relatif dan 48% dalam varians absolut.

Total varians relatif yang dihitung adalah 3.11610, dengan standar error relatif sebesar 1.76525. Sedangkan total varians absolut sebesar 3.31308, dengan standar error absolut 1.82019. Koefisien G relatif dan absolut masing-masing tercatat sebesar 0.93, yang menunjukkan bahwa instrumen penilaian memiliki tingkat keandalan yang tinggi.

c. Uji Teori D

Untuk mengetahui apakah penambahan jumlah *rater* (dosen penguji) atau *kriteria penilaian* dapat meningkatkan koefisien generalisasi (G relatif), maka dilakukan analisis lanjutan berupa D study (Decision Study). Uji ini bertujuan untuk melakukan simulasi terhadap berbagai skenario desain pengukuran dengan variasi jumlah level pada faset-faset tertentu, guna mengevaluasi dampaknya terhadap keandalan instrumen penilaian. Dengan menggunakan data hasil G study sebagai dasar, D study menyajikan proyeksi perubahan nilai koefisien G serta besarnya error pengukuran dalam skenario-skenario yang berbeda. Hasil dari analisis ini akan menjadi acuan dalam menentukan rancangan penilaian yang optimal, yaitu desain yang mampu memberikan tingkat reliabilitas tinggi namun tetap efisien dari segi penggunaan sumber daya institusi. Adapun hasil simulasi dari uji D study disajikan pada tabel 4 berikut.

Tabel 4. Nilai D Study

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	13	INF	13	INF	13	INF	13	INF	13	INF	13	INF
R	3	INF	4	INF	5	INF	6	INF	7	INF	8	INF
Q	4	INF	5	INF	6	INF	7	INF	8	INF	9	INF
Observ.		156		260		390		546		728		936
Coef_G rel.		0.93023		0.95168		0.96367		0.97118		0.97626		0.97989
rounded		0.93		0.95		0.96		0.97		0.98		0.98
Coef_G abs.		0.92615		0.94825		0.96074		0.96863		0.97400		0.97787
rounded		0.93		0.95		0.96		0.97		0.97		0.98
Rel. Err. Var.		3.11610		2.10960		1.56636		1.23308		1.01051		0.85260
Rel. Std. Err. of M.		1.76525		1.45245		1.25154		1.11044		1.00524		0.92336
Abs. Err. Var.		3.31308		2.26718		1.69768		1.34565		1.10900		0.94015
Abs. Std. Err. of M.		1.82019		1.50572		1.30295		1.16002		1.05309		0.96961

Tabel 4 menunjukkan berbagai opsi optimasi desain pengukuran dengan memvariasikan jumlah level (*level*) dan *universum* pada faktor mahasiswa (P), dosen penguji (R), dan kriteria penilaian (Q). Pada desain awal, diperoleh koefisien generalisasi relatif (G_rel) sebesar 0.93023 dan absolut (G_abs) sebesar 0.92615, yang menunjukkan tingkat keandalan instrumen yang cukup tinggi. Namun demikian, varians kesalahan relatif dan absolut masing-masing tercatat cukup besar, yaitu 3.11610 dan 3.31308, dengan standar error relatif sebesar 1.76525 dan absolut sebesar 1.82019. Dalam upaya mengoptimalkan reliabilitas, dilakukan simulasi terhadap lima skenario

peningkatan level Q. Hasilnya menunjukkan tren peningkatan signifikan pada nilai koefisien G dan penurunan varians kesalahan. Pada Opsi 1, penambahan level Q menjadi 5 menghasilkan peningkatan G_rel menjadi 0.95168 dan G_abs menjadi 0.94825, dengan penurunan varians kesalahan menjadi 2.10960 dan 2.26718. Opsi 2, dengan level Q sebanyak 6, mencatat peningkatan G_rel dan G_abs masing-masing menjadi 0.96367 dan 0.96074. Peningkatan tersebut terus berlanjut hingga Opsi 5, yang memberikan hasil terbaik dengan koefisien G_rel sebesar 0.97989 dan G_abs sebesar 0.97787. Pada opsi ini, varians kesalahan relatif dan absolut pun turun drastis menjadi 0.85260 dan 0.94015, serta standar error juga mengalami penurunan signifikan.

Diskusi

a. Varians dan Jumlah Kuadrat

Hasil analisis varians dalam Tabel 2 menunjukkan bahwa faset mahasiswa (P) merupakan sumber utama variabilitas penilaian dengan kontribusi sebesar 62.8% dari total varians. Hal ini mencerminkan bahwa terdapat perbedaan yang cukup besar dalam kemampuan dan performa mahasiswa dalam ujian skripsi. Secara pedagogis, temuan ini sejalan dengan prinsip bahwa mahasiswa merupakan aktor utama dalam proses asesmen, sehingga variasi skor yang tinggi menunjukkan adanya perbedaan nyata dalam kompetensi individual. Dalam konteks pendidikan tinggi, perbedaan ini bisa disebabkan oleh sejumlah faktor seperti latar belakang akademik, kemampuan berpikir kritis, keterampilan presentasi, serta penguasaan materi skripsi.

Sementara itu, variabilitas dari faset dosen penguji (R) hanya sebesar 0.3%, yang menunjukkan bahwa para dosen memiliki konsistensi yang cukup baik dalam memberikan penilaian. Ini merupakan hasil positif yang menunjukkan bahwa pelatihan atau pedoman penilaian yang digunakan dalam proses evaluasi sudah cukup efektif untuk mengurangi subjektivitas antar penilai. Namun, meskipun nilainya kecil, keberadaan variabilitas ini tetap perlu diperhatikan karena dalam konteks evaluasi skripsi, perbedaan sekecil apapun dapat berdampak besar terhadap nilai akhir mahasiswa. Oleh karena itu, institusi tetap perlu melakukan kalibrasi berkala dan forum penilaian bersama (scoring consensus) untuk memastikan kesamaan persepsi antar dosen penguji.

Faset kriteria penilaian (Q) berkontribusi sebesar 1.2% terhadap variabilitas total. Meskipun kontribusinya tergolong rendah, hal ini mengindikasikan bahwa kriteria yang digunakan masih memiliki ruang untuk penyempurnaan. Kriteria yang kurang jelas, terlalu umum, atau interpretatif bisa menimbulkan perbedaan dalam penerapannya. Oleh karena itu, penting bagi lembaga untuk menyusun rubrik penilaian yang rinci dan terukur, disertai contoh konkret agar penggunaannya dapat konsisten di antara semua dosen penguji.

Lebih lanjut, interaksi antara mahasiswa dan dosen penguji (PR) menyumbang 5.9% dari total variabilitas. Ini menunjukkan adanya potensi subjektivitas selektif, di mana dosen penguji mungkin memberikan penilaian yang berbeda tergantung pada siapa mahasiswa yang dinilai. Kondisi ini dapat disebabkan oleh bias kognitif, efek halo, atau bahkan hubungan interpersonal antara dosen dan mahasiswa. Untuk meminimalkan hal ini, strategi seperti double-blind assessment atau rotasi penguji bisa dipertimbangkan.

Interaksi PQ dan RQ, masing-masing hanya menyumbang 0.3% dan 0.4% dari total variabilitas, namun tetap relevan sebagai indikator bahwa terdapat sedikit perbedaan dalam bagaimana kriteria digunakan terhadap mahasiswa yang berbeda (PQ), serta bagaimana tiap dosen

menafsirkan kriteria (RQ). Ini memperkuat perlunya penguatan standarisasi instrumen dan pelatihan rater dalam memahami serta menerapkan kriteria penilaian secara seragam.

Komponen PRQ, yaitu interaksi kompleks antara mahasiswa, dosen penguji, dan kriteria, menyumbang variabilitas sebesar 28.8%, menjadikannya sumber error terbesar kedua setelah faset mahasiswa itu sendiri. Temuan ini sangat penting karena menunjukkan bahwa meskipun dosen penguji dan kriteria tampak cukup konsisten secara terpisah, ketika ketiganya berinteraksi secara simultan, muncul variabilitas yang cukup besar. Hal ini dapat diartikan sebagai ketidaksesuaian antara karakteristik mahasiswa tertentu, gaya penilaian dosen tertentu, dan kriteria tertentu yang diterapkan yang pada akhirnya menciptakan inkonsistensi. Oleh karena itu, PRQ adalah target utama dalam intervensi perbaikan penilaian.

Salah satu solusi yang dapat diambil adalah dengan meningkatkan jumlah dosen penguji (R) dan kriteria (Q) seperti yang dilakukan dalam uji simulasi D-study. Strategi ini dapat mengurangi variabilitas PRQ karena memberikan ruang triangulasi pendapat dan memperkaya dimensi penilaian, sehingga satu faktor tidak terlalu mendominasi hasil akhir (Moore et al., 2019). Namun, solusi ini tentu harus mempertimbangkan ketersediaan sumber daya dan efisiensi waktu, sehingga pemilihan opsi optimal harus disesuaikan dengan kapasitas institusi.

b. Hasil Uji Teori G Study

Hasil G-study menunjukkan bahwa subjek (P) merupakan sumber utama variabilitas penilaian, menandakan adanya perbedaan individu yang signifikan di antara mahasiswa. Hal ini menguatkan bahwa setiap mahasiswa memiliki karakteristik kemampuan yang unik, sehingga variabilitas skor adalah konsekuensi alami dari heterogenitas peserta. Interaksi antara pengamat dan subjek (PR) yang menyumbang 42% dari varians relatif mengindikasikan adanya inkonsistensi antar dosen penguji dalam menilai mahasiswa yang sama. Ini menunjukkan perlunya peningkatan kesepahaman antar dosen dalam penerapan kriteria, misalnya melalui pelatihan atau forum diskusi kalibrasi.

Sementara itu, nilai varians pada PQ yang rendah memperlihatkan bahwa aspek kuantifikasi kriteria tidak memberikan banyak kontribusi terhadap variasi penilaian, yang dapat diartikan bahwa dosen cenderung konsisten dalam menerapkan kriteria penilaian secara umum. Namun, perlu diperhatikan bahwa interaksi PRQ menjadi komponen varians terbesar kedua, yang menandakan adanya kompleksitas dalam penilaian ketika ketiga faset (P, R, dan Q) berinteraksi secara bersamaan. Koefisien G sebesar 0.93 mengonfirmasi bahwa instrumen penilaian ini sangat andal. Meski begitu, untuk meningkatkan akurasi dan konsistensi lebih lanjut, perlu dilakukan revisi terhadap aspek-aspek yang menyumbang varians tinggi, terutama pada komponen interaksi PRQ. Oleh karena itu, disarankan untuk melakukan simulasi penambahan jumlah dosen penguji (R) dan jumlah kriteria (Q) dalam desain pengukuran. Simulasi ini dapat dievaluasi melalui pendekatan *Decision Study (D-study)* untuk mengetahui desain optimal yang memberikan reliabilitas tinggi namun tetap efisien dari sisi sumber daya (Li et al., 2019). Hasil dari D-study ini nantinya dapat menjadi acuan praktis dalam merancang sistem penilaian skripsi yang lebih adil dan konsisten ke depannya.

c. Hasil Uji Teori D

Hasil dari simulasi D-Study ini menunjukkan bahwa peningkatan jumlah level pada faset kriteria penilaian (Q) secara nyata berkontribusi terhadap peningkatan keandalan dan presisi instrumen pengukuran. Koefisien G yang mendekati angka 1 pada setiap opsi optimasi menandakan bahwa desain yang diubah menjadi lebih rinci mampu menghasilkan data penilaian yang lebih stabil

dan konsisten. Penurunan varians kesalahan pada setiap simulasi mengindikasikan bahwa ketidakpastian dalam hasil penilaian dapat diminimalkan secara signifikan melalui strategi desain yang tepat. Opsi 5, sebagai skenario terbaik, memberikan bukti kuat bahwa peningkatan jumlah kriteria hingga sembilan level secara substansial meningkatkan reliabilitas sistem penilaian. Namun, dalam konteks implementasi praktis, pemilihan opsi optimasi tidak hanya bergantung pada keandalan semata, melainkan juga pada efisiensi dan ketersediaan sumber daya institusi. Oleh karena itu, meskipun Opsi 5 paling optimal secara statistik, Opsi 3 atau 4 dapat dipertimbangkan sebagai alternatif rasional yang tetap menjamin kualitas penilaian namun lebih hemat dari segi tenaga dan waktu (Robertson et al., 2024). Temuan ini menegaskan pentingnya pendekatan berbasis simulasi dalam merancang sistem penilaian yang adil, objektif, dan efisien, serta memberikan kontribusi signifikan dalam pengembangan metodologi evaluasi akademik di lingkungan perguruan tinggi.

Keterbatasan

Adapun penelitian ini dibatasi hanya pada: Pertama, jumlah sampel yang digunakan relatif kecil, yaitu hanya melibatkan 13 mahasiswa dan 3 dosen penguji, sehingga hasil generalisasi temuan masih terbatas pada konteks institusi IAI Nurul Hakim Kediri Lombok Barat. Kedua, studi ini hanya menggunakan dua faset utama, yaitu mahasiswa (P) dan dosen penguji (R), serta satu tambahan faset berupa kriteria (Q), tanpa mempertimbangkan faktor-faktor lain yang mungkin memengaruhi penilaian, seperti latar belakang akademik penguji atau aspek non-kognitif mahasiswa. Ketiga, meskipun simulasi D-study memberikan proyeksi optimal terhadap penambahan rater dan kriteria, penerapannya di dunia nyata dapat terkendala oleh keterbatasan sumber daya, seperti waktu, tenaga pengajar, dan infrastruktur. Oleh karena itu, hasil penelitian ini sebaiknya digunakan sebagai dasar awal untuk pengembangan sistem penilaian yang lebih komprehensif dan kontekstual.

Kesimpulan

Penelitian ini berhasil mengevaluasi konsistensi dan keadilan penilaian ujian skripsi di IAI Nurul Hakim Kediri Lombok Barat melalui penerapan Teori Generalisasi (Generalizability Theory, G-Theory). Hasil penelitian menunjukkan bahwa penyumbang error terbesar berasal dari interaksi PRQ, sehingga perlu dilakukan simulasi dengan menambah jumlah penilai (R) dan kriteria (Q). Koefisien generalisasi yang tinggi (0,93) menunjukkan bahwa instrumen penilaian yang digunakan memiliki reliabilitas yang sangat baik. Penelitian ini menyimpulkan bahwa untuk meningkatkan keadilan dan konsistensi penilaian ujian skripsi, perlu dilakukan upaya untuk mengurangi variabilitas dari interaksi antar faset, khususnya interaksi PRQ. Untuk implementasi praktis, disarankan untuk memilih opsi optimasi yang paling sesuai dengan sumber daya yang ada. Meskipun Opsi 5 memberikan hasil yang optimal, Opsi 4 atau Opsi 3 juga dapat dipertimbangkan karena menawarkan keseimbangan antara keandalan yang tinggi dan efisiensi sumber daya yang ada di kampus.

Referensi

Andersen, S. A. W., Nayahangan, L. J., Park, Y. S., & Konge, L. (2021). Use of Generalizability Theory for Exploring Reliability of and Sources of Variance in Assessment of Technical Skills: A Systematic Review and Meta-Analysis. In *Academic Medicine* (Vol. 96, Issue 11). https://doi.org/10.1097/ACM.0000000000004150

Azizah, A., Wahyuningsih, S., Kusumasari, V., Asmianto, A., & Setiawan, D. (2021). Validity and reliability of mathematical instruments in online learning using the Rasch measurement model at

UM lab school. *AIP Conference Proceedings*, *2330*. https://doi.org/10.1063/5.0043356

Bacon, R., Holmes, K., & Palermo, C. (2017). Exploring subjectivity in competency-based assessment judgements of assessors. *Nutrition and Dietetics*, *74*(4). https://doi.org/10.1111/1747-0080.12326

Bauer, J., Capra, S., & Ferguson, M. (2002). Use of the scored Patient-Generated Subjective Global Assessment (PG-SGA) as a nutrition assessment tool in patients with cancer. *European Journal of Clinical Nutrition*, *56*(8). https://doi.org/10.1038/sj.ejcn.1601412

Duerksen, D. R., Laporte, M., & Jeejeebhoy, K. (2021). Evaluation of Nutrition Status Using the Subjective Global Assessment: Malnutrition, Cachexia, and Sarcopenia. In *Nutrition in Clinical Practice* (Vol. 36, Issue 5). https://doi.org/10.1002/ncp.10613

Gorges, J., Koch, T., Maehler, D. B., & Offerhaus, J. (2017). Same but different? Measurement invariance of the PIAAC motivation-to-learn scale across key socio-demographic groups. *Large-Scale Assessments in Education*, *5*(1). https://doi.org/10.1186/s40536-017-0047-5

Handoyono, N. A. S. (2020). Online Thesis Exam Evaluation Using Zoom Cloud Meeting During the Covid-19 Pandemic. *VANOS Journal of Mechanical Engineering Education*, *5*(2).

Hegde, V., & Shushruth, S. (2022). Evaluation of Student's Performance in Programming Using Item Response Theory. *IEEE International Conference on Data Science and Information System, ICDSIS 2022*. https://doi.org/10.1109/ICDSIS55133.2022.9915978

Helmanda, C. M., Novrizal, A. M. N., & Safura, S. (2022). Students problems in the introduction section of thesis writing. *ACCENTIA: Journal of English Language and Education*, *2*(1). https://doi.org/10.37598/accentia.v2i1.1264

Hsiao, Y.-P. (Amy), van de Watering, G., Heitbrink, M., Vlas, H., & Chiu, M.-S. (2023). Ensuring bachelor's thesis assessment quality: a case study at one Dutch research university. *Higher Education Evaluation and Development*. https://doi.org/10.1108/heed-08-2022-0033

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research and Evaluation*, *24*(5).

Ilma, A. Z., Adhelacahya, K., & Ekawati, E. Y. (2021). Assessment for Learning Model in Competency Assessment of 21stCentury Student Assisted by Google Classroom. In A. null, A. M., & D. U.A. (Eds.), *Journal of Physics: Conference Series* (Vol. 1805, Issue 1). IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1805/1/012005

Kim, S. Y., Malatesta, J. L., & Lee, W. C. (2022). Generalizability theory and applications. In *International Encyclopedia of Education: Fourth Edition*. https://doi.org/10.1016/B978-0-12-818630-5.10009-0

Li, G., Xie, J., An, L., Hou, G., Jian, H., & Wang, W. (2019). A generalizability analysis of the mobile phone addiction tendency scale for Chinese college students. *Frontiers in Psychiatry*, *10*(APR). https://doi.org/10.3389/fpsyt.2019.00241

Moore, L. J., Freeman, P., Hase, A., Solomon-Moore, E., & Arnold, R. (2019). How consistent are challenge and threat evaluations? A generalizability analysis. *Frontiers in Psychology*, *10*(JULY). https://doi.org/10.3389/fpsyg.2019.01778

Prabhavathy, P. (2023). Kirkpatrick's Model Evaluation in Business English Training With The Humanistic Approach- An Overview. *Global Research Journal*, *2*(2). https://doi.org/10.57259/grj1068

Rashid, S. M., & McGuinness, D. L. (2018). Creating and using an education standards ontology to improve education. *CEUR Workshop Proceedings*, *2182*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053242084&partnerID=40&md5=b17f6ed4efa58c708912feb72ba76fcd

Robertson, S. E., Steingrimsson, J. A., Joyce, N. R., Stuart, E. A., & Dahabreh, I. J. (2024). Estimating Subgroup Effects in Generalizability and Transportability Analyses. *American Journal of Epidemiology*, *193*(1). https://doi.org/10.1093/aje/kwac036

Sebhatu, A., & Wennberg, K. (2023). Institutional Pressure and Failure Dynamics in the Swedish Voucher School Sector. *Scandinavian Journal of Public Administration*, *21*(3).

https://doi.org/10.58235/sjpa.v21i3.11566

Simion, A. (2023). The impact of socio-emotional learning (SEL) on academic evaluation in higher education. *Educatia 21*, *24*. https://doi.org/10.24193/ed21.2023.24.11

Sturgis, P. W., Marchand, L., Miller, M. D., Xu, W., & Castiglioni, A. (2022). Generalizability Theory and Its Application to Institutional Research. *AIR Professional File*, *Spring 2022*. https://doi.org/10.34315/apf1562022

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2021). Interrater Reliability for Multilevel Data: A Generalizability Theory Approach. *Psychological Methods*, *27*(4). https://doi.org/10.1037/met0000391

Thomas, L. J. G., Lee, M. G., Todd, C. S., Lynch, K., Loeb, S., McConnell, S., & Carlis, L. (2022). Navigating Virtual Delivery of Assessments for Head Start Children During the COVID-19 Pandemic. *Journal of Early Intervention*, *44*(2), 151–167. https://doi.org/10.1177/10538151221085942

Tierney, R. D. (2022). Fairness in Educational Testing and Assessment. In *Fairness in Educational Testing and Assessment*. https://doi.org/10.4324/9781138609877-ree35-1

Van Hooijdonk, M., Mainhard, T., Kroesbergen, E. H., & Van Tartwijk, J. (2022). Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks☆. *Thinking Skills and Creativity*, *43*. https://doi.org/10.1016/j.tsc.2021.100994