JURNAL JENDELA PENDIDIKAN

Jurnal Jendela Pendidikan
Volume 5 No 04 November 2025
ISSN: 2776-267X (Print) / ISSN: 2775-6181 (Online)
The article is published with Open Access at: https://www.ejournal.jendelaedukasi.id/index.php/JJP

Analisis Generalizabilitas Dua Facet Dalam Penilaian Internship Program Primary School Teaching College Timor-Leste

Mariano Dos Santos, Instituto Católico para a Formação de Professores (ICFP) Baucau, Timor - Leste
Ahmad ✉, Ilmu Komputer, Fakulatas Teknik, Universitas Bumigora, Mataram, Indonesia

✉ *ahmad@universitasbumigora.ac.id*

Abstract: This study aims to evaluate the reliability of assessments in the internship program at the Teacher College of Timor-Leste using the Two-Facet Generalizability Theory. The internship program plays an important role in preparing prospective teachers to apply theoretical knowledge in real teaching practice. The research design employed the i × r × p model, where *i* represents the assessment items, *r* the raters, and *p* the participants. Data were collected from 15 fourth-semester students across five schools, assessed by lecturers and teachers based on five indicators: teaching preparation, material presentation, explanation skills, pedagogical ability, and participation related to gender and special needs. Analysis using EduG software produced a relative generalizability coefficient of 0.59 and an absolute coefficient of 0.57, indicating moderate reliability. The main source of variance was the interaction among participants, raters, and criteria (PRQ), contributing 72.5% of the total variance, while participants accounted for only 13.9%. Optimization results showed that increasing the number of raters from two to three and criteria from five to twenty-five could raise the relative coefficient to 0.88 and the absolute coefficient to 0.86. The study concludes that assessment reliability can be improved by increasing the number of raters and criteria, as well as providing training and regular evaluation for raters to ensure consistency and accuracy in the assessment instrument.
Keywords: Intership Program, Two-Facet Generalizability Theory, Assessment

Abstrak: Penelitian ini bertujuan mengevaluasi reliabilitas penilaian dalam program magang di Teacher College Timor-Leste menggunakan Teori Generalisasi Dua Facet. Program magang berperan penting dalam melatih calon guru agar mampu menerapkan teori dalam praktik nyata. Desain penelitian menggunakan model i *x* r *x* p, di mana *i* adalah item penilaian, *r* penilai, dan *p* peserta. Data diperoleh dari 15 mahasiswa semester empat di lima sekolah, dinilai oleh dosen dan guru berdasarkan lima indikator: persiapan pengajaran, presentasi materi, kemampuan menjelaskan, kemampuan pedagogi, serta partisipasi gender dan kebutuhan khusus. Analisis dengan software EduG menghasilkan koefisien generalizabilitas relatif 0,59 dan absolut 0,57, menunjukkan reliabilitas moderat. Sumber utama variasi berasal dari interaksi peserta, penilai, dan kriteria (PRQ) sebesar 72,5%, sedangkan variasi dari peserta hanya 13,9%. Hasil optimasi menunjukkan bahwa menambah jumlah penilai dari dua menjadi tiga dan kriteria dari lima menjadi 25 dapat meningkatkan koefisien generalizabilitas relatif menjadi 0,88 dan absolut 0,86. Penelitian ini menyimpulkan bahwa reliabilitas penilaian dapat ditingkatkan melalui penambahan penilai dan kriteria, serta pelatihan dan evaluasi berkala bagi penilai untuk menjaga konsistensi dan akurasi instrumen penilaian.
Kata kunci: Intership Program, Teori Generalisasi Dua Facet, Penilaian

Received 23 Oktober 2025; **Accepted** 30 Oktober 2025; **Published** 10 November 2025
Citation: Santos, M.D., & Ahmad. (2025). Analisis Generalizabilitas Dua Facet Dalam Penilaian Internship Program Primary School Teaching College Timor-Leste. *Jurnal Jendela Pendidikan*, 5 (04), 819-827.

PENDAHULUAN

Intership Program adalah komponen integral dalam pelatihan guru yang memungkinkan calon guru untuk mengaplikasikan teori yang telah mereka pelajari dalam lingkungan pengajaran yang nyata (Anwar and Winsor 2024; CHED 2017). Intership Program memberikan kesempatan bagi calon guru untuk memperoleh pengalaman praktis, mengasah keterampilan pedagogis, dan mengembangkan kompetensi yang diperlukan untuk menjadi guru yang efektif (Edy et al. 2019; Newmark and Hutchins 1981). Dalam konteks pendidikan guru, Intership Program bukan hanya tentang menerapkan teori, tetapi juga tentang mengembangkan keterampilan interpersonal, manajemen kelas, dan kemampuan beradaptasi dengan berbagai situasi pengajaran.

Internship program di Teacher College Timor-Leste dirancang dengan tujuan memberikan pengalaman praktis yang kaya kepada calon guru. Program ini menggabungkan berbagai komponen yang penting untuk pengembangan profesional calon guru, termasuk perencanaan pelajaran, presentasi materi, interaksi dengan siswa, dan manajemen kelas. Feiman-Nemser (2001) menekankan pentingnya pengalaman pengajaran langsung dalam mempersiapkan guru yang efektif dengan menyatakan bahwa praktik pengajaran langsung memainkan peran kunci dalam mengembangkan kesiapan mengajar calon guru. Melalui Intership Program, calon guru dapat menguji dan menyesuaikan pendekatan pengajaran mereka, menerima umpan balik yang konstruktif, dan belajar dari pengalaman nyata di kelas (Koşar 2021; Li, Xie, and Zeng 2023).

Penilaian yang efektif dari program magang ini sangat penting untuk memastikan bahwa calon guru benar-benar siap untuk mengajar secara mandiri. Penilaian yang akurat dan dapat diandalkan membantu mengidentifikasi kekuatan dan kelemahan calon guru, memberikan umpan balik yang konstruktif, dan menentukan area yang perlu diperbaiki. Grossman (2010) menyoroti bahwa penilaian yang tepat dalam praktik pengajaran membantu calon guru untuk memahami dan menginternalisasi strategi pengajaran yang efektif serta meningkatkan kemampuan manajemen kelas mereka. Penilaian dalam konteks magang harus mencakup berbagai aspek keterampilan mengajar dan kemampuan interpersonal untuk memberikan gambaran yang komprehensif tentang kinerja calon guru.

Meskipun pelaksanaan Internship Program di *Teacher College* Timor-Leste telah berjalan rutin setiap tahun, pelaksanaan penilaiannya masih menghadapi sejumlah kendala. Salah satu masalah utama yang muncul adalah ketidakkonsistenan hasil penilaian antarpenilai. Nilai yang diberikan oleh dosen pembimbing dan guru lapangan sering menunjukkan perbedaan yang cukup besar untuk kinerja mahasiswa yang relatif serupa. Kondisi ini menimbulkan keraguan terhadap reliabilitas hasil penilaian, karena belum ada instrumen baku dan pelatihan penilai yang sistematis untuk menyamakan persepsi dalam menilai kompetensi calon guru. Akibatnya, hasil akhir penilaian magang belum sepenuhnya dapat dijadikan dasar yang kuat untuk menggambarkan kemampuan mengajar mahasiswa secara objektif. Kondisi inilah yang menjadi dasar perlunya penelitian ini menggunakan pendekatan Teori Generalisasi Dua Facet untuk mengevaluasi dan meningkatkan reliabilitas penilaian dalam *Internship Program*.

Teori Generalisasi Dua Facet menawarkan kerangka kerja yang berguna untuk mengevaluasi reliabilitas penilaian dalam konteks Intership Program. Teori ini mengusulkan bahwa ada dua aspek utama yang perlu dipertimbangkan dalam penilaian: generalisasi dan spesifikasi (Tasdelen Teker and Guler 2019)(Lee and Ye 2023). Generalisasi berkaitan dengan kemampuan calon guru untuk menerapkan keterampilan pengajaran yang telah dipelajari ke berbagai situasi pengajaran yang berbeda, sementara spesifikasi mengacu pada keterampilan pengajaran yang spesifik dalam konteks tertentu. Menurut Jiang and Skorupski (2018), Teori Generalisasi memberikan pendekatan yang lebih komprehensif untuk menilai reliabilitas instrumen penilaian dengan mempertimbangkan berbagai sumber variasi dalam hasil penilaian. Pendekatan ini

memungkinkan penilai untuk memahami sejauh mana hasil penilaian dapat digeneralisasi ke berbagai situasi dan konteks pengajaran.

Di Teacher College Timor-Leste, Intership Program tidak hanya menilai kemampuan pedagogis calon guru tetapi juga bagaimana mereka berinteraksi dengan siswa dan mengelola kelas. Choi and Park (2022) menyatakan bahwa pengembangan guru yang efektif harus mencakup penilaian yang holistik yang melibatkan berbagai aspek keterampilan mengajar dan kemampuan interpersonal. Dengan menggunakan Teori Generalisasi Dua Facet, penilaian dalam Intership Program dapat dioptimalkan untuk memastikan bahwa hasil penilaian mencerminkan kinerja yang sebenarnya dari calon guru dalam berbagai kondisi pengajaran (Medvedev et al. 2021; Uzun et al. 2018).

Penelitian ini bertujuan untuk menganalisis penerapan Teori Generalisasi Dua Facet dalam penilaian Intership Program di Teacher College Timor-Leste. Dengan menganalisis data penilaian dari berbagai penilai dan situasi penilaian, penelitian ini bertujuan untuk mengidentifikasi sumber utama variasi dalam hasil penilaian dan memberikan rekomendasi untuk meningkatkan reliabilitas penilaian. Gugiu, Gugiu, and Baldus (2012) menunjukkan bahwa penggunaan Teori Generalisasi dapat memberikan wawasan yang lebih mendalam mengenai variabilitas penilaian dan membantu dalam mengembangkan instrumen penilaian yang lebih andal dan valid.

Penelitian ini juga berfokus pada pengembangan strategi untuk meningkatkan kualitas penilaian dalam Intership Program. Dalam konteks pendidikan guru, penting untuk memiliki instrumen penilaian yang dapat diandalkan dan valid untuk memastikan bahwa calon guru mendapatkan umpan balik yang akurat dan konstruktif. Hasil penelitian ini diharapkan dapat memberikan kontribusi yang signifikan terhadap pengembangan kurikulum dan strategi penilaian di Teacher College Timor-Leste serta memberikan rekomendasi praktis untuk meningkatkan efektivitas program magang dalam melatih calon guru yang kompeten dan siap mengajar.

Desain penelitian ini menggunakan pendekatan i x r x p, di mana 'i' adalah item penilaian, 'r' adalah penilai, dan 'p' adalah peserta yang dinilai. Sampel terdiri dari 15 mahasiswa semester 4 dari lima sekolah berbeda yang dipilih secara acak. Data penilaian dikumpulkan dari dosen dan guru penilai di sekolah dasar menggunakan lima indikator penilaian: persiapan pengajaran, presentasi materi, kemampuan penjelasan materi, kemampuan pedagogi, dan partisipasi gender serta kebutuhan khusus. Analisis data dilakukan menggunakan software EduG untuk menghitung koefisien generalizabilitas serta melakukan studi-G dan studi-D (Cardinet, Johnson, and Pini 2011; Clauser 2008).

Analisis hasil penelitian menunjukkan bahwa interaksi antara peserta, penilai, dan kriteria (PRQ) adalah sumber utama variasi dalam penilaian, menunjukkan pentingnya mempertimbangkan kombinasi unik dari ketiga facet ini untuk meningkatkan konsistensi penilaian. Studi G menegaskan bahwa meskipun ada upaya konsistensi, variabilitas interaksi facet masih menantang reliabilitas penilaian. Optimasi menunjukkan bahwa menambah jumlah penilai dan kriteria adalah langkah efektif untuk meningkatkan reliabilitas dan presisi penilaian, menjadikannya lebih dapat diandalkan dan kurang terpengaruh oleh variabilitas antar penilai atau kondisi penilaian.

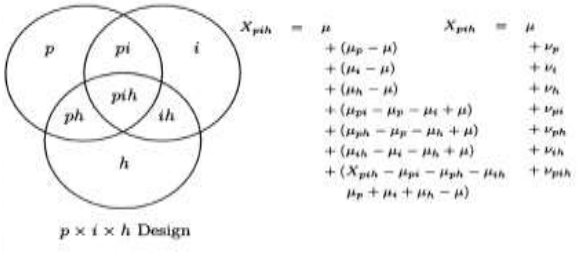
Dalam upaya untuk meningkatkan reliabilitas penilaian, penting untuk mempertimbangkan pelatihan tambahan bagi penilai dan evaluasi berkala terhadap instrumen penilaian yang digunakan. Pelatihan tambahan dapat membantu penilai memahami standar penilaian yang seragam dan teknik penilaian yang objektif, sementara evaluasi berkala dapat memastikan bahwa instrumen penilaian tetap relevan dan efektif dalam mengukur kompetensi peserta.

Dalam konteks ini, penelitian ini tidak hanya berfokus pada evaluasi reliabilitas penilaian, tetapi juga pada pengembangan strategi untuk meningkatkan kualitas penilaian dalam program magang. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi yang signifikan terhadap pengembangan kurikulum dan strategi penilaian di Teacher College Timor-Leste serta memberikan rekomendasi praktis untuk meningkatkan efektivitas program magang dalam melatih calon guru yang kompeten dan siap mengajar. Penelitian ini juga diharapkan dapat memberikan wawasan yang lebih

mendalam mengenai variabilitas penilaian dan membantu dalam mengembangkan instrumen penilaian yang lebih andal dan valid.

METODE

Penelitian ini menggunakan desain i x r x p, di mana 'i' adalah item penilaian, 'r' adalah penilai, dan 'p' adalah peserta yang dinilai. Sampel terdiri dari 15 mahasiswa semester 4 dari lima sekolah berbeda yang dipilih secara acak. Data penilaian dikumpulkan dari dosen dan guru penilai di sekolah dasar menggunakan lima indikator penilaian: persiapan pengajaran, presentasi materi, kemampuan penjelasan materi, kemampuan pedagogi, dan partisipasi gender serta kebutuhan khusus.



$$X_{pih} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (\mu_h - \mu) + (\mu_{pi} - \mu_p - \mu_i + \mu) + (\mu_{ph} - \mu_p - \mu_h + \mu) + (\mu_{ih} - \mu_i - \mu_h + \mu) + (X_{pih} - \mu_{pi} - \mu_{ph} - \mu_{ih} + \mu_p + \mu_i + \mu_h - \mu)$$

$$= \mu + \nu_p + \nu_i + \nu_h + \nu_{pi} + \nu_{ph} + \nu_{ih} + \nu_{pih}$$

Design	Main Effects	Interaction Effects
p × i × h	p, i, h	pi, ph, ih, pih
p × (i:h)	p, h, i:h	ph, pi:h
(i:p) × h	p, h, i:p	ph, ih:p
i: (p × h)	p, h, i:ph	ph
(i × h):p	p, i:p, h:p	ih:p
i:h:p	p, h:p, i:h:p	

Gambar 1. Desain Penelitian *i x r x p*

Teknik pertama dalam analisis data adalah menghitung estimasi koefisien generalizabilitas dengan menggunakan software EduG. Koefisien ini merupakan ukuran reliabilitas yang menunjukkan sejauh mana hasil penilaian dapat digeneralisasi ke berbagai situasi dan kondisi penilaian yang berbeda. Software EduG digunakan karena kemampuannya untuk menghitung koefisien ini dengan mempertimbangkan berbagai sumber variasi dalam data penilaian seperti interaksi antara peserta, penilai, dan kriteria penilaian.

Langkah kedua adalah melakukan analisis data Studi-G. Studi-G adalah metode analisis yang digunakan untuk mengidentifikasi dan mengukur sumber-sumber variasi dalam penilaian. Analisis ini membantu dalam memahami bagaimana setiap facet (peserta, penilai, dan kriteria) berkontribusi terhadap total variasi dalam hasil penilaian. Dengan demikian, Studi-G memungkinkan peneliti untuk mengevaluasi sejauh mana hasil penilaian konsisten dan dapat digeneralisasi.

Langkah ketiga adalah melakukan analisis data Studi-D berdasarkan temuan dari Studi-G. Studi-D adalah tahap lanjutan di mana peneliti menggunakan informasi dari Studi-G untuk mengevaluasi dan mengoptimalkan desain penilaian. Studi-D bertujuan untuk menentukan kombinasi terbaik dari jumlah penilai dan kriteria yang akan meningkatkan reliabilitas dan presisi penilaian. Dengan menggunakan temuan dari Studi-G, peneliti dapat melakukan simulasi dan prediksi untuk mengidentifikasi desain penilaian yang optimal, memastikan bahwa hasil penilaian lebih akurat dan andal.

HASIL PENELITIAN

Berdasarkan hasil analisis varians yang dilakukan dalam penelitian ini, ditemukan bahwa interaksi antara peserta, penilai, dan kriteria (PRQ) merupakan sumber utama variasi dalam hasil penilaian. Temuan ini menunjukkan bahwa kombinasi unik dari ketiga facet ini, peserta yang dinilai, penilai yang melakukan penilaian, dan kriteria yang digunakan untuk menilai, sangat mempengaruhi konsistensi hasil penilaian. Studi Generalizabilitas (Studi G) memperkuat temuan ini dengan menunjukkan bahwa meskipun ada upaya untuk menjaga konsistensi dalam proses penilaian, variabilitas yang dihasilkan dari interaksi ketiga facet tersebut tetap menjadi tantangan yang signifikan bagi reliabilitas penilaian.

Dalam konteks ini, variabilitas dalam penilaian dapat disebabkan oleh berbagai faktor, termasuk perbedaan interpretasi kriteria oleh penilai yang berbeda, variasi dalam kinerja peserta di bawah kondisi penilaian yang berbeda, serta bagaimana kriteria penilaian diterapkan dalam situasi yang berbeda. Untuk mengatasi masalah ini dan

meningkatkan reliabilitas serta presisi penilaian, analisis optimasi dilakukan. Hasil optimasi menunjukkan bahwa meningkatkan jumlah penilai dan jumlah kriteria yang digunakan dalam penilaian dapat secara signifikan memperbaiki reliabilitas penilaian.

Dengan menambah jumlah penilai, variasi yang disebabkan oleh perbedaan individu antar penilai dapat diminimalkan, sehingga penilaian menjadi lebih konsisten dan objektif. Demikian pula, dengan meningkatkan jumlah kriteria penilaian, penilaian menjadi lebih komprehensif dan dapat menangkap berbagai aspek kinerja peserta secara lebih akurat. Langkah-langkah ini akan membuat penilaian menjadi lebih dapat diandalkan dan kurang terpengaruh oleh variabilitas antar penilai atau kondisi penilaian, memastikan bahwa hasil penilaian mencerminkan kinerja yang sebenarnya dari peserta dalam berbagai kondisi pengajaran.

Tabel 1. Desain Observasi dan Estimasi (Observation and Estimation Designs)

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Person	P	15	INF	—
Rater	R	2	INF	—
Criteria	Q	5	INF	—

Tabel 2. Analisis Varian (Analysis of Variance)

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	242.16000	14	17.29714	1.01262	1.01262	1.01262	13.9	0.66528
R	14.72667	1	14.72667	0.08458	0.08458	0.08458	1.2	0.17371
Q	14.29333	4	3.57333	0.17738	0.17738	0.17738	0.0	0.11746
PR	83.57333	14	5.96952	0.13881	0.13881	0.13881	1.9	0.45634
PQ	362.70667	56	6.47690	0.60071	0.60071	0.60071	8.3	0.77560
RQ	30.77333	4	7.69333	0.16119	0.16119	0.16119	2.9	0.30323
PRQ	295.42667	56	5.27548	5.27548	5.27548	5.27548	72.5	0.97963
Total	**1043.66000**	**149**	—	—	—	—	**100 %**	—

Analisis varians pada tabel 2, menunjukkan bahwa variasi terbesar berasal dari interaksi antara peserta, penilai, dan kriteria (PRQ) dengan kontribusi sebesar 72.5% terhadap total varians. Varians dari peserta (P) memberikan kontribusi sebesar 13.9%, sedangkan varians dari penilai (R) dan kriteria (Q) sangat kecil, masing-masing sebesar 1.2% dan 0.0%. Hal ini menunjukkan bahwa interaksi antara ketiga facet ini sangat mempengaruhi hasil penilaian. Selanjutnya untuk memperoleh nilai yang maksimal maka dilakukan analisis lebih lanjut dengan studi-G seperti pada tabel 3 di bawah ini.

Tabel 3. Tabel Studi-G (Measurement Design P/RQ)

Source of Variance	Differentiation Variance	Source of Variance	Relative Error Variance	% Relative	Absolute Error Variance	% Absolute
P	1.01262	—	—	—	—	—
R	—	—	—	0.04226	5.4	
Q	—	—	(0.00000)	0.0	(0.00000)	0.0
PR	—	—	0.06940	9.7	0.06940	8.9
PQ	—	—	0.12014	16.8	0.12014	15.5
RQ	—	—	0.01612	2.1	0.01612	2.1
PRQ	—	—	0.52755	73.6	0.52755	68.0
Sum of Variances	**1.01262**	**—**	**0.71710**	**100%**	**0.77548**	**100%**
Standard Deviation	**1.00629**	**—**	**Relative SE = 0.84681**	**—**	**Absolute SE = 0.88061**	**—**

Dalam analisis studi-G, interaksi antara person (peserta), rater (penilai), dan kriteria (PRQ) pada tabel 3 menunjukkan kontribusi terbesar terhadap varians kesalahan, dengan persentase 73.6% untuk varians kesalahan relatif dan 68.0% untuk varians kesalahan absolut. Ini berarti bahwa sebagian besar variasi dalam hasil penilaian disebabkan oleh kombinasi unik dari ketiga facet ini. Koefisien generalizabilitas relatif sebesar 0.59 dan

koefisien generalizabilitas absolut sebesar 0.57 mengindikasikan bahwa reliabilitas penilaian berada pada tingkat moderat.

Koefisien generalizabilitas ini menggambarkan seberapa baik hasil penilaian dapat digeneralisasi ke berbagai situasi dan penilai yang berbeda. Nilai yang lebih tinggi menunjukkan reliabilitas yang lebih baik. Namun, nilai yang moderat dalam penelitian ini menunjukkan bahwa hasil penilaian masih dipengaruhi oleh variabilitas antar penilai dan situasi penilaian. Standard error relatif sebesar 0.84681 dan absolut sebesar 0.88061 menunjukkan adanya tingkat ketidakpastian dalam penilaian. Standard error ini mengukur sejauh mana hasil penilaian dapat diharapkan bervariasi dari nilai sebenarnya, sehingga nilai yang lebih tinggi mengindikasikan ketidakpastian yang lebih besar. Ini menegaskan bahwa meskipun penilaian berusaha konsisten, masih terdapat variabilitas yang signifikan dalam interaksi antara peserta, penilai, dan kriteria penilaian.

Agar nilai errornya menjadi lebih kecil makan dilakukan penambahan terhadapt jumlah ratel atau penilai dan jumlah kriteria, hasil analisis setelah penambahan tersebut adalah seperti pada tabel nilai analisis optimasi berikut ini.

Tabel 4. Nilai Analisis Optimasi

	G-study		Option 1	
	Lev.	Univ.	Lev.	Univ.
P	15	INF	15	INF
R	2	INF	3	INF
Q	5	INF	25	INF
Observ.	150		1125	
Coef_G rel.	0.58543		0.87805	
Rounded	0.59		0.88	
Coef_G abs.	0.56631		0.85556	
Rounded	0.57		0.86	
Rel. Err. Var.	0.71710		0.14064	
Rel. Std. Err. of M.	0.84681		0.37502	
Abs. Err. Var.	0.77548		0.17096	
Abs. Std. Err. of M.	0.88061		0.41348	

Analisis optimasi pada tabel 4, menunjukkan bahwa meningkatkan jumlah penilai dari 2 menjadi 3 dan jumlah kriteria dari 5 menjadi 25 dapat secara signifikan meningkatkan reliabilitas penilaian. Dalam studi ini, koefisien generalizabilitas relatif meningkat dari 0.59 menjadi 0.88, dan koefisien generalizabilitas absolut meningkat dari 0.57 menjadi 0.86 setelah peningkatan jumlah penilai dan kriteria. Koefisien generalizabilitas relatif dan absolut mengukur seberapa baik hasil penilaian dapat digeneralisasi ke berbagai situasi dan konteks yang berbeda. Peningkatan nilai koefisien ini menunjukkan bahwa penilaian menjadi lebih andal dan konsisten.

Selain itu, peningkatan jumlah penilai dan kriteria juga menurunkan varians kesalahan relatif dan absolut serta kesalahan standar. Varians kesalahan relatif dan absolut adalah ukuran variabilitas yang disebabkan oleh ketidakpastian dalam penilaian, sedangkan kesalahan standar menunjukkan sejauh mana hasil penilaian dapat bervariasi dari nilai sebenarnya. Penurunan dalam metrik ini menunjukkan bahwa hasil penilaian menjadi lebih presisi dan akurat. Dengan kata lain, peningkatan jumlah penilai dan kriteria mengurangi kemungkinan variasi dalam hasil penilaian yang disebabkan oleh perbedaan individu antar penilai atau kondisi penilaian yang berbeda. Ini menghasilkan penilaian yang lebih dapat diandalkan, memastikan bahwa hasil penilaian lebih mencerminkan kinerja sebenarnya dari peserta dalam berbagai situasi pengajaran.

PEMBAHASAN

Hasil analisis varians dalam penelitian ini menunjukkan bahwa interaksi antara person (peserta), rater (penilai), dan kriteria (PRQ) merupakan sumber terbesar variasi dalam penilaian, dengan kontribusi sebesar 72.5% terhadap total varians. Hal ini berarti bahwa sebagian besar variasi dalam hasil penilaian disebabkan oleh bagaimana ketiga elemen ini berinteraksi satu sama lain. Variasi yang disebabkan oleh person atau peserta hanya berkontribusi sebesar 13.9%, menunjukkan bahwa perbedaan individu antar peserta memberikan pengaruh yang lebih kecil terhadap hasil penilaian secara keseluruhan. Sebaliknya, variasi dari rater (penilai) dan kriteria (Q) sangat kecil, masing-masing hanya 1.2% dan 0.0%, yang mengindikasikan bahwa penilai dan kriteria yang digunakan secara individu tidak memberikan kontribusi besar terhadap total variasi dalam penilaian.

Interaksi antara person, rater, dan kriteria (PRQ) memiliki pengaruh signifikan terhadap hasil penilaian karena setiap kombinasi unik dari ketiga facet ini dapat menghasilkan hasil yang berbeda. Misalnya, seorang penilai mungkin memiliki standar yang berbeda dalam menilai kriteria tertentu dibandingkan dengan penilai lainnya, dan peserta mungkin merespons secara berbeda terhadap penilai dan kriteria yang berbeda pula. Ketidakkonsistenan ini lebih mungkin terjadi karena kombinasi unik dari peserta, penilai, dan kriteria yang digunakan dalam setiap situasi penilaian.

Dalam analisis studi-G, interaksi antara person, rater, dan kriteria (PRQ) juga menunjukkan kontribusi terbesar terhadap varians kesalahan relatif sebesar 73.6% dan varians kesalahan absolut sebesar 68.0%. Koefisien generalizabilitas relatif sebesar 0.59 dan koefisien generalizabilitas absolut sebesar 0.57 menunjukkan bahwa reliabilitas penilaian berada pada tingkat moderat. Koefisien generalizabilitas ini mengukur sejauh mana hasil penilaian dapat digeneralisasi ke berbagai situasi dan kondisi penilaian yang berbeda. Nilai koefisien yang lebih tinggi menunjukkan reliabilitas yang lebih baik, sedangkan nilai yang lebih rendah menunjukkan bahwa penilaian mungkin kurang konsisten (Tasdelen Teker and Guler 2019).

Standard error relatif sebesar 0.84681 dan absolut sebesar 0.88061 menunjukkan adanya ketidakpastian yang cukup signifikan dalam penilaian. Standard error ini mengukur sejauh mana hasil penilaian dapat bervariasi dari nilai sebenarnya, dengan nilai yang lebih tinggi menunjukkan ketidakpastian yang lebih besar (Cardinet et al. 2011; Kim, Malatesta, and Lee 2022). Angka-angka ini menegaskan bahwa meskipun ada upaya untuk membuat penilaian yang konsisten, variabilitas dalam interaksi facet masih menciptakan tantangan dalam mencapai reliabilitas yang lebih tinggi.

Hasil analisis optimasi memberikan pemahaman tentang bagaimana peningkatan jumlah penilai dan kriteria dapat meningkatkan reliabilitas penilaian secara signifikan. Dengan meningkatkan jumlah penilai dari 2 menjadi 3 dan jumlah kriteria dari 5 menjadi 25, koefisien generalizabilitas relatif meningkat dari 0.59 menjadi 0.88, dan koefisien generalizabilitas absolut meningkat dari 0.57 menjadi 0.86. Peningkatan ini menunjukkan bahwa dengan lebih banyak penilai dan kriteria, penilaian menjadi lebih konsisten dan andal. Selain itu, varians kesalahan relatif dan absolut serta kesalahan standar menurun, menunjukkan peningkatan presisi dalam hasil penilaian.

Penurunan varians kesalahan ini berarti bahwa hasil penilaian akan lebih dapat diandalkan dan lebih sedikit terpengaruh oleh variabilitas antar penilai atau kondisi penilaian (Uzun et al. 2018). Dengan lebih banyak penilai, perbedaan individu antar penilai dapat diminimalkan, sehingga hasil penilaian menjadi lebih objektif dan konsisten. Demikian pula, dengan lebih banyak kriteria penilaian, hasil penilaian menjadi lebih komprehensif dan akurat, mencakup berbagai aspek kinerja peserta.

Kesimpulannya, analisis varians dan optimasi menunjukkan bahwa interaksi antara person, rater, dan kriteria merupakan faktor utama dalam variabilitas penilaian. Peningkatan jumlah penilai dan kriteria penilaian adalah langkah efektif untuk meningkatkan reliabilitas dan presisi penilaian, memastikan bahwa hasil penilaian lebih konsisten dan andal. Langkah ini penting untuk mengurangi pengaruh variabilitas antar

penilai dan kondisi penilaian, sehingga penilaian lebih akurat mencerminkan kinerja peserta dalam berbagai situasi pengajaran.

SIMPULAN

Berdasarkan hasil analisis menggunakan Teori Generalisasi Dua Facet, penelitian ini menjawab bahwa sumber utama variasi dalam penilaian Internship Program di Teacher College Timor-Leste berasal dari interaksi antara peserta (person), penilai (rater), dan kriteria (PRQ) dengan kontribusi sebesar 72.5% terhadap total varians. Hal ini menunjukkan bahwa reliabilitas penilaian dipengaruhi secara signifikan oleh kombinasi unik dari ketiga facet tersebut. Temuan ini sekaligus menjawab tujuan pertama penelitian, yaitu mengidentifikasi sumber utama variasi hasil penilaian dalam konteks program magang. Selanjutnya, penelitian ini juga menemukan bahwa peningkatan jumlah penilai dan kriteria penilaian merupakan strategi yang efektif untuk meningkatkan reliabilitas hasil penilaian. Hasil analisis optimasi menunjukkan bahwa dengan menambah jumlah penilai dari dua menjadi tiga dan kriteria dari lima menjadi dua puluh lima, koefisien generalizabilitas relatif meningkat dari 0.59 menjadi 0.88, dan absolut dari 0.57 menjadi 0.86. Dengan demikian, penelitian ini menjawab tujuan kedua, yaitu memberikan rekomendasi konkret untuk meningkatkan reliabilitas penilaian melalui perbaikan desain penilaian, pelatihan penilai, dan evaluasi instrumen secara berkala agar hasil penilaian lebih akurat, konsisten, dan mencerminkan kinerja calon guru yang sesungguhnya.

DAFTAR PUSTAKA

1. Anwar, Sohail, and Frances Winsor. 2024. “Internship Development For A New Program: Preparation, Sponsor Development, And Intership Follow Up.”
2. Cardinet, Jean, Sandra Johnson, and Gianreto Pini. 2011. *Applying Generalizability Theory Using EduG*.
3. CHED. 2017. “Revised Guidelines for Student Intership Program in the Philippines (SIPP) for All Programs.” *Commission on Higher Education* (104).
4. Choi, Lee Jin, and Mi Yung Park. 2022. “Teaching Practicum During COVID-19: Pre-Service English Language Teachers’ Professional Identities and Motivation.” *SAGE Open* 12(3). doi: 10.1177/21582440221119472.
5. Clauser, Brian E. 2008. “A Review of the EDUG Software for Generalizability Analysis.” *International Journal of Testing* 8(3). doi: 10.1080/15305050802262357.
6. Edy, Duwi Leksono, Siti Malikha, Widiyanti, and Andika Bagus Nur Rahma Putra. 2019. “Analysis of the Competence of Expertise in the Intership Program in the Industrial Era 4.0 Vocational Education in Indonesia.” *International Journal of Innovation, Creativity and Change* 8(1).
7. Feiman-Nemser, Sharon. 2001. “From Preparation to Practice: Designing a Continuum to Strengthen and Sustain Teaching.” *Teachers College Record* 103(6). doi: 10.1111/0161-4681.00141.
8. Grossman, P. 2010. “Learning to Practice: The Design of Clinical Experience in Teacher Preparation.” *Policy Brief Series Document of AACTE & NEA* May.
9. Gugiu, Mihaiela Ristei, P. Cristian Gugiu, and Robert Baldus. 2012. “Utilizing Generalizability Theory to Investigate the Reliability of Grades Assigned to Undergraduate Research Papers.” *Journal of MultiDisciplinary Evaluation* 8(19). doi: 10.56645/jmde.v8i19.362.
10. Jiang, Zhehan, and William Skorupski. 2018. “A Bayesian Approach to Estimating Variance Components within a Multivariate Generalizability Theory Framework.” *Behavior Research Methods* 50(6). doi: 10.3758/s13428-017-0986-3.
11. Kim, Stella Y., Jaime L. Malatesta, and Won Chan Lee. 2022. “Generalizability Theory and Applications.” in *International Encyclopedia of Education: Fourth Edition*.
12. Koşar, Gülten. 2021. “Distance Teaching Practicum: Its Impact on Pre-Service EFL Teachers’ Preparedness for Teaching.” *IAFOR Journal of Education* 9(2). doi:

10.22492/ije.9.2.07.
13. Lee, Kwangmin, and Yafei Ye. 2023. "Investigating the Reliability of Foreign Language Classroom Anxiety Scale (FLCAS): An Application of Generalizability Theory." *Research Methods in Applied Linguistics* 2(1). doi: 10.1016/j.rmal.2022.100036.
14. Li, Qi, Zhilong Xie, and Guofang Zeng. 2023. "The Influence of Teaching Practicum on Foreign Language Teaching Anxiety Among Pre-Service EFL Teachers." *SAGE Open* 13(1). doi: 10.1177/21582440221149005.
15. Medvedev, Oleg, Quoc Truong, Alexander Merkin, Robert Borotkanics, Rita Krishnamurthi, and Valery Feigin. 2021. "Cross-Cultural Validation of the Stroke Riskometer Using Generalizability Theory." *Scientific Reports* 11(1). doi: 10.1038/s41598-021-98591-8.
16. Newmark, Charles S., and Tracey C. Hutchins. 1981. "Survey of Professional Education in Ethics in Clinical Psychology Intership Programs." *Journal of Clinical Psychology* 37(3). doi: 10.1002/1097-4679(198107)37:3<681::AID-JCLP2270370342>3.0.CO;2-O.
17. Tasdelen Teker, Gülşen, and Neşe Guler. 2019. "Thematic Content Analysis of Studies Using Generalizability Theory." *International Journal of Assessment Tools in Education* 6(2). doi: 10.21449/ijate.569996.
18. Uzun, N. Bilge, Mehtap Aktaş, Semih Aşiret, and Seha Yorulmaz. 2018. "Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students." *Asian Journal of Education and Training* 4(2). doi: 10.20448/journal.522.2018.42.85.90.

PROFIL SINGKAT

Mariano Dos Santos, adalah dosen di *Instituto Católico para a Formação de Professores (ICFP)*, Baucau, Timor-Leste. Bidang minatnya meliputi psikologi, pendidikan guru, evaluasi pembelajaran, dan pengukuran pendidikan. Saat ini ia terdaftar sebagai mahasiswa program doktor pada Prodi Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta. Ia aktif dalam penelitian terkait peningkatan kualitas pendidikan di Timor-Leste.

Ahmad adalah dosen di Universitas Bumigora Mataram, Indonesia. Bidang keahliannya mencakup pendidikan matematika, asesmen pendidikan, dan pembelajaran berbasis budaya lokal. Ia sedang menempuh studi doktoral pada Prodi Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta. Selain mengajar, ia aktif meneliti pengembangan instrumen dan evaluasi pembelajaran di sekolah.