jurnal βeta kimia
e-ISSN: 2807-7938 (online) dan p-ISSN: 2807-7962 (print)
Volume 4, Nomor 2, November 2024
http://ejurnal.undana.ac.id/index.php/jbk

Konsistensi dan Dependabilitas Penilaian Microteaching pada Program Studi Pendidikan Kimia : Teori Generalisabilitas

(*Consistency and Dependability of Microteaching Assessment in the Chemistry Education Study Program: Generalisability Theory*)

Dewi Lestarani[1,2]*, Raden Rosnawati[1], Antonius Umbu Anarato[1], Arvinda C. Lalang[2], Heru Christianto[2], Maria Ulfah[3], Winioliski L.O. Rohi Bire[4], Adrian R. Lalang[5]

[1]*Universitas Negeri Yogyakarta: Jl. Colombo No.1, Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia*
[2]*Program Studi Pendidikan Kimia, Universitas Nusa Cendana,Indonesia:Jln. Adisucipto Penfui, Kupang, NTT, Indonesia*
[3]*Program Studi Pendidikan Kimia, Universitas Tanjungpura, Kota Pontianak , Kalimantan Barat , Indonesia*
[4]*Program Studi Farmasi, Jl. Manafe No.17, Kayu Putih, Universitas Citra Bangsa, Kota Kupang, NTT*
[5]*Program Studi Statistika, Jl. Ir. Soekarno, Fontein, Universitas San Pedro, Kota Kupang, NTT*
**e-mail korespondensi: dewilestarani.2023@student.uny.ac.id*

Info Artikel:
Dikirim:
14 Oktober 2024
Revisi:
14 November 2024
Diterima:
20 November 2024
Kata Kunci:
Teori Generalisabilitas, Konsistensi, Dependabilitas, Penilaian
Keywords:
Generalisability, Consistency, Dependability, Assessment Theory

Lisensi:
Attribution-Share Alike 4.0 International (CC-BY-SA 4.0)
OPEN ACCESS

Abstrak-Subjektivitas dan kurang konsistennya penilai/rater dalam proses penyekoran merupakan kritik yang umum ditujukan pada penilaian dalam pembelajaran. Oleh karena itu, artikel ini menyajikan hasil reliabilitas penilaian pada matakuliah Microteaching menggunakan tori generalisabilitas dengan desain px (i:r). Data yang dikumpulkan merupakan nilai ujian mahasiswa baik ujian pada nilai pengamatan di kelas, ujian tengah semester, ujian akhir semester. Responden berjumlah 30 orang, dengan 5 jenis tes dan dinilai oleh 3 rater yang berbeda. Analisis data pada pada artikel ini menggunakan program R untuk menghitung relative error varians; mendapatkan koefisien generalizabilitas dan *coefficient dependability* dari hasil tes secara empirik. Hasil penelitian didapatkan nilai eror terbesar terdapat pada person dan rater 1 dengan nilai 28.6 % dan 18%, nilai koefisien generalizabilitas sebesar 0.93 dan nilai koefisien *dependability* sebesar 0.69. Sehingga berdasarkan teori D maka dilakukan modifikasi dengan menambahkan jumlah rater agar nilai koefisien *dependability* menjadi › 0.7.

Abstract-Subjectivity and inconsistency among raters in the scoring process are common criticisms directed at assessment in education. Therefore, this article presents the results of reliability analysis in the Microteaching course assessment using Generalizability Theory with a px (i:r) design. The data collected includes students' exam scores from classroom observation assessments, midterm exams, and final exams. The respondents consisted of 30 individuals, evaluated through five types of tests by three different raters. Data analysis in this article employs the R program to calculate the relative error variance, as well as to obtain the generalizability coefficient and dependability coefficient from the test results empirically. The findings indicate that the largest error values are attributed to persons and Rater 1, with values of 28.6% and 18%, respectively. The generalizability coefficient is 0.93, and the dependability coefficient is 0.69. Based on D-theory, modifications were made by increasing the number of raters to ensure the dependability coefficient exceeds 0.7.

PENDAHULUAN

Penilaian merupakan proses yang sangat penting dalam pendidikan karena penilaian memiliki peran utama untuk pengembangan kualitas pendidikan [1]. Penilaian dapat dikatakan sebagai jembatan antara mengajar dan belajar. Melalui penilaian, seorang pendidik dapat memprediksi dan mengetahui sesuai atau tidaknya hasil dengan tujuan pembelajaran yang diharapkan [2]. Tujuan dilakukannya penilaian dalam pendidikan adalah memberikan informasi, meningkatkan progam dan kualitas pembelajaran yang sedang berlangsung.

Kementerian Pendidikan dan Kebudayaan pada Pedoman dan Penilaian Gerakan Literasi Nasional Tahun 2017 disebutkan bahwa proses penilaian dilakukan dengan menggunakan berbagai macam metode yang relevan dan cocok, yaitu pengamatan, dokumentasi, wawancara dengan pemangku kepentingan, serta telaah data sekunder dari berbagai macam lembaga yang relevan. Penggunaan pedoman penilaian dan evaluasi gerakan literasi nasional merupakan usaha untuk mewujudkan kebutuhan pendidikan di abad 21. Pedoman tersebut digunakan sebagai rujukan untuk mengetahui konten pengetahuan. Oleh karena itu, penilaian ini dapat juga digunakan untuk melihat kinerja peserta didik dalam literasi, berhitung, sains, dan pemahaman sosial dalam konteks berpikir tingkat tinggi, berkolaborasi, dan dalam hal penggunaan teknologi [3]. Salah satu jenis pendekatan penilaian adalah penilaian sumatif. Penilaian sumatif digunakan untuk mengevaluasi topik atau materi yang telah dipelajari di kelas.

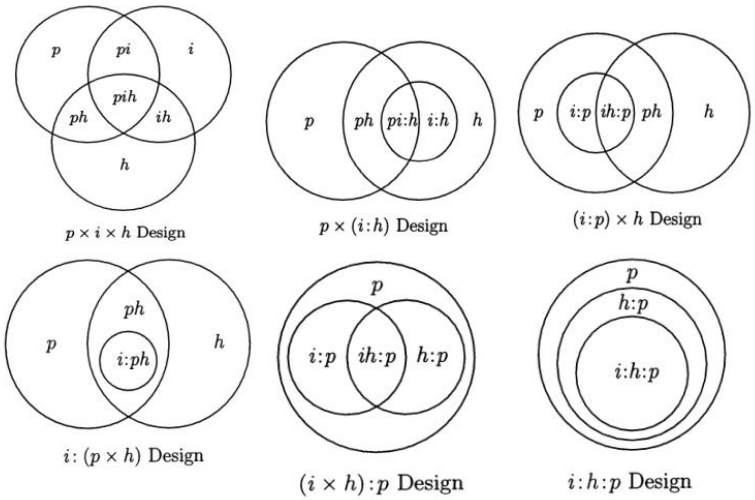
Penilaian sumatif yang dilakukan oleh pendidik seharusnya disesuaikan dengan proses pembelajaran yang dilakukan di kelas. Proses penilaian patut mempertimbangkan subjektivitas penilai karena memengaruhi hasil penilaian, sehingga diperlukan evaluasi yang lebih objektif untuk memastikan keadilan dan akurasi [4], [5]. Kritik yang sering muncul dalam penilaian adalah pengaruh subjektivitas penilai terkait proses penilaian yang cenderung lebih tinggi dibandingkan dengan penilai yang lain [6], [7]. Pelaksanakan penilaian sering terdapat beberapa variabel yang tidak relevan seperti suasana hati dan kondisi sekitar mempengaruhi proses penilaian [8]. Dengan kata lain, kemampuan penilai dalam memahami dan menerapkan rubrik penilaian serta tingkat subjektivitas penilai sangat berpengaruh dalam memberikan penilaian. Dua hal yang dapat dilakukan untuk meminimalisir efek subjektivitas dalam penilaian [9]. Pertama, mengembangkan mekanisme penilaian yang jelas yang berisi deskripsi tugas yang harus dikerjakan oleh siswa dan guru. Deskripsi tersebut berisi kriteria keterampilan dan pengetahuan yang akan dinilai, yang dituangkan dalam sebuah rubrik. Kedua, memberikan pelatihan kepada penilai tentang bagaimana menggunakan rubrik untuk membuat keputusan tentang tes yang diberikan kepada siswa.

Efektivitas penilaian tergantung pada kualitas dan koordinasi antara penilai dan rubrik. Dengan kata lain, peran rubrik dalam dalam konteks penilaian sangat penting [10]. Rubrik akan menjadi alat yang sangat penting dalam mentransfer apa yang apa yang telah dilakukan atau dihasilkan siswa ke dalam bentuk penilaian [11]. Oleh karena itu, untuk mendapatkan hasil yang valid dan reliabel, rubrik harus menyediakan informasi yang cukup untuk membantu penilai menilai tes yang telah berhasil diselesaikan oleh siswa [12].

Penilai seringkali tidak konsisten dalam menggunakan rubrik penilaian [13]. Hal ini disebabkan oleh kurangnya pengalaman penilai dalam menggunakan serta kualitas rubrik [10]. Selain itu, ketidakkonsistenan penggunaan rubrik juga terjadi karena kurangnya pemahaman tentang konstruk atau aspek rubrik. Penilai mencoba menggunakan berbagai macam rentang ketika menilai hasil [14]. Namun, meskipun penilai memiliki pelatihan dan pengalaman mengajar yang sama, perbedaan dalam interpretasi individu dan subjektivitas dapat menyebabkan evaluasi tugas siswa yang berbeda [15]. Perbedaan dalam penilaian disebabkan oleh cara penilai penilai dalam memahami dan menerapkan rubrik penilaian serta tingkat subjektivitas dalam memberikan penilaian [16]. Hal ini menunjukkan pentingnya instrumen penilaian beserta rubrik yang baik. Dengan demikian, penilai yang menggunakan penilaian tes dapat memberikan nilai yang relatif sama. Masalah mengenai ketidakkonsistenan penilai dalam memahami rubrik berimplikasi pada perbedaan yang mencolok pada hasil skor yang yang diberikan pada lembar penilaian [17]. Dampaknya, hasil penilaian yang diterima yang diterima oleh siswa menjadi bias. Kondisi ini menjadi masalah serius dan mendorong dilakukannya pengujian empiris terhadap validitas dan reliabilitas instrumen penilaian penilai yang digunakan dalam menilai tes mahasiswa.

Keterbaruan penelitian ini terletak pada penggunaan pendekatan analitis modern (*Generalizability Theory*), fokus pada varians kesalahan dan modifikasi sistem penilaian, yang memberikan kontribusi signifikan terhadap praktik penilaian yang lebih reliabel dan dapat diandalkan. Penyempurnaan konsep reliabilitas dalam teori klasik adalah generalizabilitas yang merupakan suatu cara untuk meningkatkan akurasi interpretasi tes [18]. Teori generalizabilitas (teori G) memberikan suatu kerangka kerja untuk mengkonseptualisasi, menginvestigasi, dan mendesain pengamatan yang reliabel sehingga dapat memperkirakan sumber kesalahan pengukuran [19]. Teori G mempertimbangkan sumber-sumber varians sistematik yang berbeda dalam pengukuran dan menggambarkan cara-cara mengestimasi banyak varians yang disumbangkan oleh sumber-sumber ini [20].

Pengamatan (skor tes teruji) dalam teori G dilihat sebagai sampel dari populasi pengamatan yang dapat diterima. Populasi menggambarkan kondisi yang teruji dapat diamati atau dites, yang menimbulkan hasil yang ekivalen pada beberapa tingkatan spesifik. Teori G menekankan bahwa keberadaan populasi yang berlainan dan menjadi tanggung jawab penyusun tes untuk menetapkan batasan berlakunya hasil tes tersebut. Kondisi spesifik yang dipertimbangkan dalam tes biasa disebut facets atau dimension. Sebagai contoh adalah ukuran sampel pengambil tes, banyaknya item tes, bentuk tes, jumlah rater dan sebagainya. Kondisi ini dispesifikasi, dan pengaruhnya dapat diuji. Enam model desain dua facet digambarkan dalam diagram Venn seperti gambar di bawah ini:



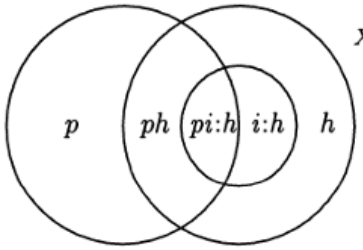
Gambar 1. Diagram Venn 6 models design two-facet [18]

Tulisan ini akan mengkaji beberapa hal sebagai berikut: (1) Membahas perhitungan *relative error varians*, (2) Mengestimasi koefisien generalizabilitas dari hasil tes secara empirik, (3) Membahas hasil perhitungan *coefficient dependability* dengan desain p x (i:r) berdasarkan Brennan (2001) [18].

METODE PENELITIAN

Pendekatan yang digunakan dalam tulisan ini yaitu pendekatan kuantitatif deskriptif pada mata kuliah *Microteaching*. Penelitian ini dilakukan pada seluruh tes pada mata kuliah Microteaching Program Studi Pendidikan Kimia, Universitas Nusa Cendana yang melibatkan 30 mahasiswa dengan 5 tes yang sama. Rater dalam artikel ini merupakan 3 orang dosen dimana setiap rater diberikan lembar penilaian yang sama berisikan 10 item untuk 5 aspek pengukuran. Instrumen yang sama namun dinilai oleh 3 rater yang berbeda. Penilaian yang dilakukan oleh rater akan dianalisis untuk menentukan koefisien generalizabilitas tes. Penentuan koefisien

generalizabilitas tes dilakukan dengan model 2 facet sehingga menggunakan analisis varian model 3 jalur, pada masing-masing mahasiswa, item dan rater. Desain yang digunakan p x (i:h), dan dimodifikasi atau diubah berdasarkan kebutuhan peneliti dengan p x (i:r). Dimana *p* itu *person* (mahasiswa) yang diuji pada mata kuliah Microteaching, *i* adalah item pertanyaan, *r* adalah dosen penilai matakuliah. Analisis data dengan bantuan program R. Gambar 2 merupakan diagram *Generalizability Theory Two Facet* dengan desain p x (i:r).



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

$$X_{pih} = \mu + \nu_p + \nu_h + \nu_{i:h} + \nu_{ph} + \nu_{pi:h}$$

Gambar 2. Desain p x (i:r) [21]

HASIL DAN PEMBAHASAN

Terdapat 30 mahasiswa matakuliah Microteaching pada Program Studi Pendidikan Kimia, Universitas Nusa Cendana yang dinilai oleh tiga raters pada lima tes dengan penilaian yang sama, Data yang peroleh dianalisis berbantuan program R. *Output* dari analisis varians penilaian dengan dua aspek atau *two-facet design* ditampilkan pada Gambar 3.

```
> twofacet
         Source Est.Variance Percent.Variance
1 person.raters       1.0927              5.5%
2        person       5.7084             28.6%
3   item.raters       0.5140              2.6%
4        raters       2.7433             13.7%
5     raters.1       3.5945               18%
```

Gambar 3. Nilai eror pada tiap varian

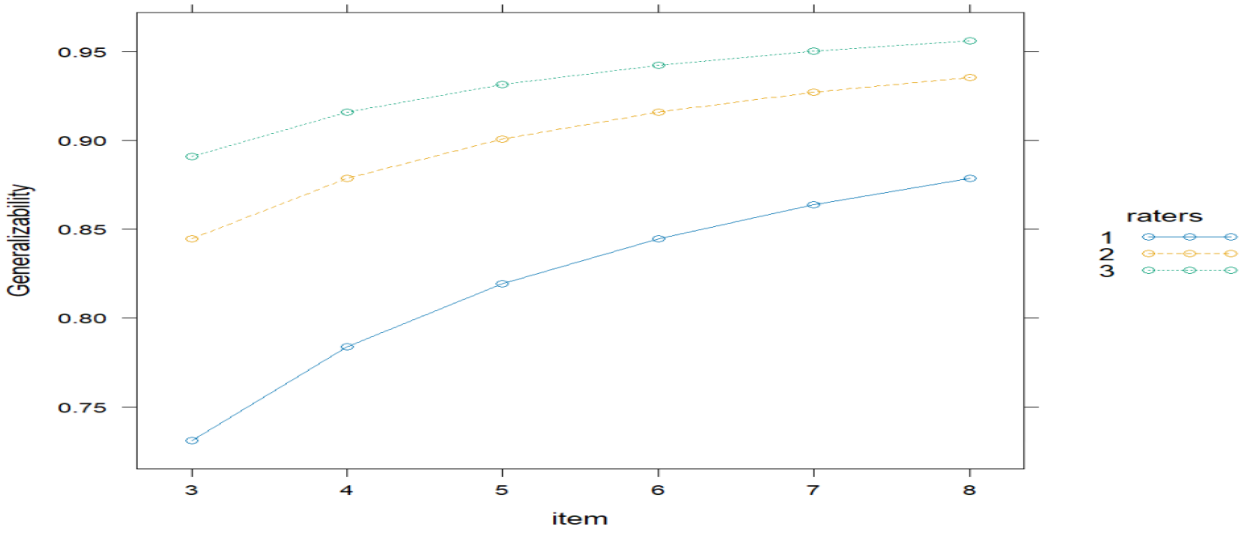
Berdasarkan Gambar 3 di atas, 5 aspek yang memberikan varians pada penilaian Microteaching yaitu, person sebesar 28,6 %, raters.1 sebesar 18%, raters sebesar 13,7%, person.raters sebesar 5,5% dan item.raters sebesar 2,6%. Nilai varians terbesar terdapat pada person, ratersl dan raters.

Varians pada *person* memiliki nilai yang lebih tinggi (5.7084 atau 28.6%) karena ini menunjukkan perbedaan antar individu yang dinilai. Dalam konteks penilaian atau pengukuran, varians pada *person* yang tinggi mengindikasikan adanya variasi yang signifikan antara individu-individu yang diuji atau dinilai [21], [22]. Hal ini dipengaruhi oleh perbedaan karakteristik individu, dimana setiap individu memiliki keahlian, kualitas, tingkat kompetensi, keterampilan yang berbeda akan menghasilkan skor yang berbeda. Varians *person* yang tinggi juga bisa menunjukkan bahwa instrumen penilaian dapat diandalkan dalam membedakan secara efektif antara individu. Dalam analisis varians, skor sesungguhnya (*true score*) dari individu biasanya mencerminkan kemampuan atau atribut aktual mahasiswa [23]. Hal ini menunjukkan bahwa kemampuan mahasiswa dengan menggunakan tes ini tidak menunjukkan skor tampak karena

memiliki eror yang tinggi, dan untuk menurunkan nilai eror ini maka perlunya ada persiapan bagi mahasiswa untuk mengikuti tes.

Nilai varians tertinggi kedua adalah pada *raters*.1 dengan nilai eror sebesar 18%, menunjukkan sumber variasi tambahan yang terkait dengan perbedaan antar rater. Varians *raters*.1 bisa mengindikasikan adanya ketidakkonsistenan atau perbedaan yang belum dijelaskan antara rater. Misalnya, meskipun ada instruksi atau pedoman penilaian yang sama, setiap rater mungkin memiliki kecenderungan atau bias tertentu yang mempengaruhi konsistensi penilaian mereka. Varians ini dapat muncul ketika rater tidak selalu memberikan skor yang sama untuk situasi atau objek yang mirip karena interpretasi atau standar pribadi yang sedikit berbeda. Hal ini dibuktikan pada lembar penilaian rater 1, memiliki skala penilaian yang berbeda dibanding 2 rater lainnya, dimana rater 1 menilai mahasiswa lebih rendah kepada beberapa mahasiswa.

Selanjutnya, varians pada *raters* (dengan nilai 2.7433 atau 13.7%) menggambarkan perbedaan atau variasi antar rater (penilai) dalam memberikan skor atau penilaian. Setiap rater mungkin memiliki kecenderungan atau bias pribadi yang mempengaruhi cara mereka menilai. Varians ini muncul karena adanya perbedaan cara setiap rater mengevaluasi atau menilai individu atau objek yang sama. Meskipun ada pedoman atau kriteria penilaian yang sama, setiap rater bisa memiliki standar interpretasi yang berbeda. Variasi ini menunjukkan bahwa setiap rater memiliki standar atau ambang batas penilaian yang tidak seragam. Variasi ini penting untuk diperhatikan karena tingginya varians pada raters dapat menurunkan reliabilitas penilaian secara keseluruhan. Dalam konteks penilaian atau pengukuran, diupayakan agar perbedaan antar rater diminimalkan, misalnya dengan memberikan pelatihan yang lebih terstandar atau menggunakan pedoman penilaian yang lebih jelas, sehingga setiap rater dapat memberikan skor yang lebih seragam. Pembuktian nilai rater ini terlihat jelas pada plot yang ditampilkan pada Gambar 3.



Gambar 4. Plot data Generalizability 3 rater

Gambar 4 menunjukkan plot hubungan antara jumlah item dan tingkat *generalizability* (keandalan generalisasi) dengan jumlah penilai (raters) yang berbeda. Berdasarkan gambar di atas, diperoleh adanya perbedaan yang cukup signifikan antara rater 1 dengan rater 2 dan rater 3. Hal ini menunjukkan bahwa ada perbedaan penilaian yang diberikan rater 1 terhadap mahasiswa dengan jumlah item tes yang sama. Adanya subjektivitas penilaian pada raters 1, sedangkan pada rater 2 dan rater 3, selisih nilai *generalizability* keduanya tidak jauh.

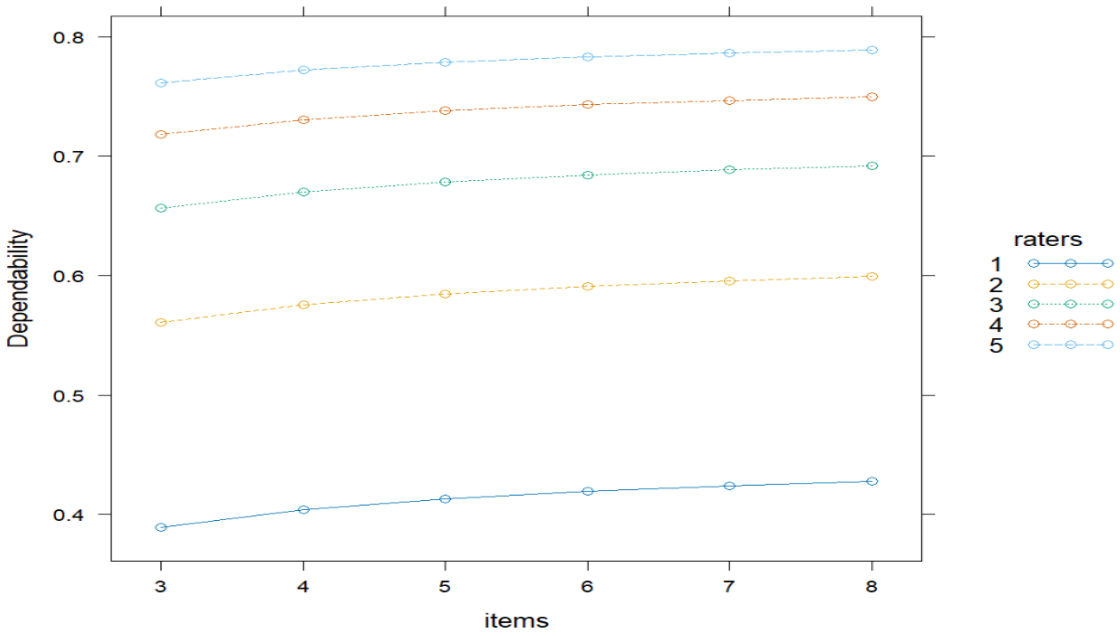
Selanjutnya, estimasi koefisien generalizabilitas dan koefisien *dependability* untuk memahami keandalan yang dapat digeneralisasi ke berbagai konteks (generalizability) dan keandalan dalam

kondisi tertentu (*dependability*) [24], [25]. Berikut estimasi koefisien generalizabilitas dan *dependability*.

```
The generalizability coefficient is: 0.9314737.
The dependability coefficient is: 0.6898185.
```

Gambar 5. Nilai Koefisien G dan Koefisien D

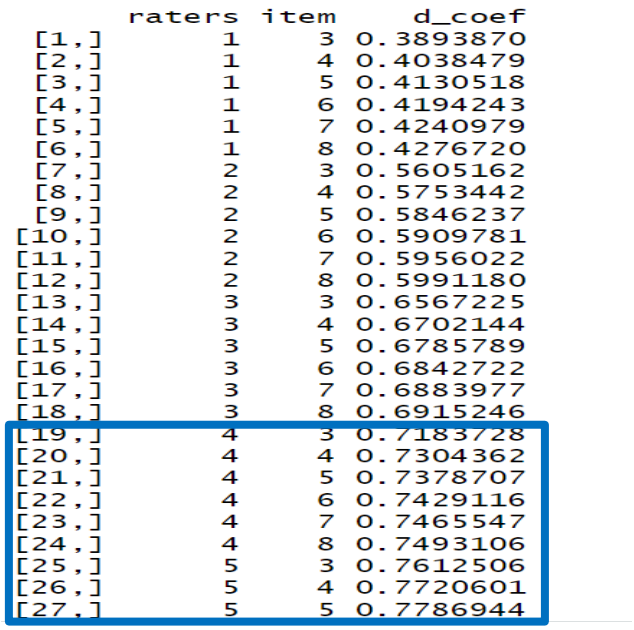
Berdasarkan Gambar 5, nilai koefisien *generalizability* sebesar 0.93 menunjukkan tingkat keandalan atau konsistensi dalam pengukuran yang dapat digeneralisasi ke berbagai kondisi atau situasi. Nilai ini mendekati 1, yang berarti hasil pengukuran sangat dapat diandalkan dan stabil ketika diaplikasikan ke sampel yang lebih luas atau pada situasi berbeda. Pada gambar 5, estimasi koefisien *dependability* sebesar 0.69 menunjukkan tingkat keandalan atau konsistensi pengukuran dalam kondisi atau situasi spesifik. Nilai ini lebih rendah daripada koefisien *generalizability*, yang berarti keandalan dalam kondisi tertentu lebih rendah dibandingkan keandalan yang dapat digeneralisasi. Ada perbedaan nilai yang cukup besar dikarenakan koefisien D menekankan estimasi, penggunaan, dan interpretasi dari varians komponen untuk membuat keputusan, dengan prosedur pengukuran yang baik. Sehingga perlu diperhatikan bahwa prosedur penilaian keandalan dilakukan dengan langkah mencari harga-harga varians yang dibentuk oleh facet. Studi D digunakan untuk mengidentifikasi jumlah kondisi yang optimal dari setiap aspek untuk memaksimalkan keandalan. Gambar 5 juga menunjukkan nilai reliabel berdasarkan koefisen *dependability* tidak reliabel karena < 0.7. Untuk meningkatkan nilai reliabilitas maka dilakukan modifikasi variabel terkait yaitu pada jumlah rater. Hasil modifikasi dapat dilihat pada Gambar 6 di bawah ini.



Gambar 6. Plot modifikasi untuk Koefisien Dependability

Berdasarkan Gambar 6, nilai koefisien *dependability* meningkat di atas 0.7 dengan bertambahnya jumlah rater menjadi 4 dan 5 rater. Plot ini menunjukkan pentingnya kombinasi jumlah rater dalam meningkatkan *dependability*. Sehingga dapat dapat disimpulkan penambahkan

rater meningkatkan keandalan penilaian. Hal ini dibuktikan berdasarkan data pada Gambar 7 di bawah ini.



	raters	item	d_coef
[1,]	1	3	0.3893870
[2,]	1	4	0.4038479
[3,]	1	5	0.4130518
[4,]	1	6	0.4194243
[5,]	1	7	0.4240979
[6,]	1	8	0.4276720
[7,]	2	3	0.5605162
[8,]	2	4	0.5753442
[9,]	2	5	0.5846237
[10,]	2	6	0.5909781
[11,]	2	7	0.5956022
[12,]	2	8	0.5991180
[13,]	3	3	0.6567225
[14,]	3	4	0.6702144
[15,]	3	5	0.6785789
[16,]	3	6	0.6842722
[17,]	3	7	0.6883977
[18,]	3	8	0.6915246
[19,]	4	3	0.7183728
[20,]	4	4	0.7304362
[21,]	4	5	0.7378707
[22,]	4	6	0.7429116
[23,]	4	7	0.7465547
[24,]	4	8	0.7493106
[25,]	5	3	0.7612506
[26,]	5	4	0.7720601
[27,]	5	5	0.7786944

Gambar 7. Nilai Koefisien *Dependability* Tiap Rater dan Item

Berdasarkan data pada Gambar 7, diperoleh nilai koefisien *dependability* (koefisien D), meningkat secara bertahap seiring bertambahnya jumlah raters dan item. Hal ini mencerminkan bahwa lebih banyak raters dan item menghasilkan keandalan yang lebih tinggi. Semakin banyak raters dan item, semakin tinggi nilai d_coef, yang berarti evaluasi menjadi lebih andal. Hal ini sejalan dengan hasil plot yang sebelumnya ditunjukkan pada Gambar 6.

KESIMPULAN

Nilai ujian matakuliah *Microteaching* pada Program Studi Pendidikan Kimia, Universitas Nusa Cendana telah diperiksa menggunakan teori generalisasi dengan desain p x (i:r). Berdasarkan hasil dan pembahasan yang telah dilakukan maka disimpulkan beberapa hal berikut: nilai eror terbesar terdapat pada person dan rater 1 dengan nilai 28.6 % dan 18%, Koefisien generalizabilitas dari hasil tes seca empirik sebesar 0.93 dan nilai koefisien dependability sebesar 0.69. Sehingga berdasarkan teori D maka dilakukan modifikasi dengan menambahkan jumlah rater agar nilai koefisien *dependability* menjadi › 0.7.

DAFTAR PUSTAKA

[1] L. Lubna, "Isu-Isu Pendidikan Di Indonesia: Inovasi Kurikulum Dan Peningkatan Profesionalitas Guru," *Society*, vol. 5, no. 2, pp. 15–25, Oct. 2014, https://doi.org/10.20414/society.v5i2.1455.

[2] T. Andayani and F. Madani, "Peran Penilaian Pembelajaran Dalam Meningkatkan Prestasi Siswa di Pendidikan Dasar," *Jurnal Educatio FKIP UNMA*, vol. 9, no. 2, pp. 924–930, Jun. 2023, https://doi.org/10.31949/educatio.v9i2.4402.

[3] V. S. Damaianti, Y. Abidin, and R. Rahma, "Higher order thinking skills-based reading literacy assessment instrument: An Indonesian context," *Indonesian Journal of Applied Linguistics*, vol. 10, no. 2, pp. 513–525, Oct. 2020, https://doi.org/10.17509/ijal.v10i2.28600.

[4] J. L. Ferreira Neto, L. G. M. F. Duarte, and C. M. F. Penido, "Avaliação E Processos De Subjetivação Na Atenção Básica À Saúde: Avaliação E Subjetivação," *Psicol Estud*, vol. 27, Feb. 2022, https://doi.org/10.4025/psicolestud.v27i0.48663.

[5] S. Salsabila, "Sistem Pendukung Keputusan Penilaian Tenaga Kependidikan Dengan Metode Fuzzy Analytic Hierarchy Process," *Journal of Information Technology*, vol. 4, no. 2, pp. 01–06, May 2023, https://doi.org/10.47292/joint.v4i2.73.

[6] N. K. Park, M. Y. Chun, and J. Lee, "Revisiting Individual Creativity Assessment: Triangulation in Subjective and Objective Assessment Methods," *Creat Res J*, vol. 28, no. 1, pp. 1–10, Jan. 2016, https://doi.org/10.1080/10400419.2016.1125259.

[7] K. Grint, "What's Wrong With Performance Appraisals? A Critique and A Suggestion," *Human Resource Management Journal*, vol. 3, no. 3, pp. 61–77, Mar. 1993, https://doi.org/10.1111/j.1748-8583.1993.tb00316.x.

[8] W. Priatna and R. Purnomo, "Implementasi Fuzzy Inference System Metode Sugeno Pada Aplikasi Penilaian Kinerja Dosen," *Techno.Com*, vol. 19, no. 3, pp. 245–261, Aug. 2020, https://doi.org/10.33633/tc.v19i3.3638.

[9] B. Bulut, H. Ulu, and A. Kan, "Multimodal Literacy Scale: A Study of Validity and Reliability," *Egitim Arastirmalari - Eurasian Journal of Educational Research*, vol. 15, no. 61, pp. 45–60, 2015, https://doi.org/10.14689/ejer.2015.61.3.

[10] A. Kan and O. Bulut, "Crossed random-effect modeling: Examining the effects of teacher experience and rubric use in performance assessments," *Eurasian Journal of Educational Research*, no. 57, pp. 1–28, Dec. 2014, https://doi.org/10.14689/ejer.2014.57.4.

[11] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educ Res Rev*, vol. 2, no. 2, pp. 130–144, Jan. 2007, https://doi.org/10.1016/j.edurev.2007.05.002.

[12] J. Stuhlmann, C. Daniel, A. Dellinger, R. Kenton, and T. Powers, "A Generalizability Study Of The Effects Of Training On Teachers' Abilities To Rate Children's Writing Using A Rubric," *Read Psychol*, vol. 20, no. 2, pp. 107–127, Jun. 1999, https://doi.org/10.1080/027027199278439.

[13] R. Smit, P. Bachmann, V. Blum, T. Birri, and K. Hess, "Effects of a rubric for mathematical reasoning on teaching and learning in primary school," *Instr Sci*, vol. 45, no. 5, pp. 603–622, Oct. 2017, https://doi.org/10.1007/s11251-017-9416-2.

[14] W. D. Shafer, G. Swanson, N. Bene, and G. Newberry, "Effects of Teacher Knowledge of Rubrics on Student Achievement in Four Content Areas," *Applied Measurement in Education*, vol. 14, no. 2, pp. 151–170, Apr. 2001, https://doi.org/10.1207/S15324818AME1402_3.

[15] T. Lumley, "Perceptions of Language-trained Raters and Occupational Experts in a Test of Occupational English Language Profficiency," *English for Specific Purposes*, vol. 17, no. 4, pp. 347–367, Oct. 1998, https://doi.org/10.1016/S0889-4906(97)00016-1.

[16] T. Eckes, "Rater types in writing performance assessments: A classification approach to rater variability," *Language Testing*, vol. 25, no. 2, pp. 155–185, Apr. 2008, https://doi.org/10.1177/0265532207086780.

[17] H. L. Andrade and Y. Du, "Student Perspectives on Rubric-Referenced Assessment Student Perspectives on Rubric-Referenced Assessment," 2005. [Online]. Available: https://scholarsarchive.library.albany.edu/edpsych_fac_scholarhttps://scholarsarchive.library.albany.edu/edpsych_fac_scholar/2

[18] R. L. Brennan, "Generalizability Theory and Classical Test Theory," *Applied Measurement in Education*, vol. 24, no. 1, pp. 1–21, Dec. 2010, https://doi.org/10.1080/08957347.2011.532417.

[19] N. M. Webb and R. J. Shavelson, "Generalizability Theory: Overview," in *Encyclopedia of Statistics in Behavioral Science*, Wiley, 2005. https://doi.org/10.1002/0470013192.bsa703.

[20] M. Sudaryanto, K. Saddhono, and Lina, "Applying Item Responses Theory For Measuring Student's Ability In Academic Speaking," *Humanities & Social Sciences Reviews*, vol. 8, no. 2, pp. 305–312, Mar. 2020, https://doi.org/10.18510/hssr.2020.8234.

[21] M. H. Zubaidillah, "Prinsip dan Alat Evaluasi dalam Pendidikan," Jul. 17, 2018. https://doi.org/10.31219/osf.io/4tgfm.

[22] F. Razi, "Konsep Dasar Evaluasi Pembelajaran," Mar. 12, 2021. https://doi.org/10.31219/osf.io/nmua2.

[23] L. G. Otaya, H. Anwar, and R. T. Husain, "Estimating the Students' Skill in Reciting and Writing Alqur'an at Faculty of Tarbiyah and Teacher Training IAIN Sultan Amai Gorontalo," *Nadwa*, vol. 1, no. 1, p. 75, Aug. 2019, https://doi.org/10.21580/nw.2019.1.1.3590.

[24] J. Sánchez-Meca, J. A. López-Pina, and J. A. López-López, "Una revisión de los estudios meta-analíticos de generalización de la fiabilidad," *Escritos de Psicología - Psychological Writings*, vol. 2, no. 1, pp. 110–121, Dec. 2008, https://doi.org/10.24310/espsiescpsi.v2i1.13365.

[25] P. E. Clayson, K. A. Carbine, S. Baldwin, J. A. Olsen, and M. J. Larson, "Using Generalizability Theory and the ERP Reliability Analysis (ERA) Toolbox for Assessing Test-Retest Reliability of ERP Scores Part 1: Algorithms, Framework, and Implementation," Jul. 07, 2020. https://doi.org/10.31234/osf.io/kcven.

[26] Brennan, R. L. (2001). Statistics for Social Science and Public Policy: Generalizability Theory. Springer Verlag