# KONSISTENSI PARAMETER TES

***Rustam***
***(Universitas Terbuka)***

---

***Abstrak***

*The study was amied at findding out the consistency of test parameters in which sample size of the same population change. The population of this study was student's responses to the examination of the PPDG 2332 course which was given at the examination period of 96.2 at Universitas Terbuka. By proportional cluster random sampling, 3897 students were assigned to be samples of this study.*

*The study finding showed that the test parameters (the reliability coefficient, the index of difficulty level, and the index of discrimination), and the Standard Error of Measurement (SEM) will change if the sample size changes. All test parameters except the index of difficulty level, had a positive correlation with the sample size. In the other words, all test parameters, except the increase of the sample size. The greater the sample sizes, the smaller the index of the test difficulty becomes; or the greater the sample size is, the higher the level of test difficulty becomes.*

---

## Pendahuluan

Salah satu pendekatan di dalam pengukuran pendidikan adalah teori pengukuran klasik. Pengukuran klasik memiliki ciri khusus yang ditunjukan oleh kenyataan bahwa karakteristik butir tes (perangkat tes) atau alat ukur psikologi lainnya tidak dapat dipisahkan dari kelompok peserta yang merespons tes tersebut.

Artinya apabila butir tes yang sama direspons oleh kelompok peserta yang berbeda, maka parameter kelompok butir tes itu pada umumnya berubah (Naga, 1992). Dengan kata lain, tingkat kesukaran, daya pembeda dan koefisien reliabilitas tes itu berubah semata-mata karena mereka direspons oleh kelompok yang berbeda. Untuk butir yang sama, jika kelompok peserta yang meresponsnya berbeda, parameter butir tes dan perangkat tesnya akan berbeda pula.

Demikian pula, jika kelompok peserta yang merespons kelompok butir tes tertentu berbeda, maka ciri atau parameter kelompok tes tersebut pada umumnya berubah. Hasil analisis tes yang direfleksikan oleh parameter peserta dan butir tes, sangat bergantung kepada butir tes atau kelompok butir tes dan peserta yang merespon tes tersebut.

Teori pengukuran klasik masih banyak digunakan oleh lembaga atau perorangan dalam mengambil keputusan yang terkait dengan tes atau alat ukur psikologi pada umumnya. Namun, para pengguna jarang sekali melihat ciri teori pengukuran kalik ini, sehingga pengambilan keputusan hanya berdasarkan kepada kriteria, misalnya: koefisien tingkat kesukaran (P) dikatakan baik apabila $0{,}30 \leq P \leq 0{,}80$, koefisien reliabilitas (r) dikatakan baik apabila $r \geq 0{,}6$ dan seterusnya, tanpa memperhatikan ukuran sampel baik butir tes maupun peserta tes.

Sebaiknya, para pengguna dalam mengambil keputusan tidak hanya memperhatikan kriteria parameter tes yang telah disepakati, tetapi juga perlu memperhatikan ukuran sampel butir tes dan juga ukuran sampel peserta tes yang merespons tes yang diberikan. Untuk sampel butir tes, di samping ukurannya, dimensi butir dalam tes juga berpengaruh dalam hasil kalibrasi (Allen dan Yen, 1979).

Seperti telah diuraikan, bahwa hasil analisis tes berupa parameter tes sangat bergantung kepada peserta yang merespon tes tersebut, bagaimanakah konsistensi parameter tes, bila ukuran sampel yang merespons (peserta) tes tersebut berubah?

## Tujuan dan Kegunaan Penelitian

Tujuan penelitian ini adalah untuk mengetahui konsistensi parameter tes, bila ukuran sampel yang merespons tes tersebut berubah atau berbeda, walaupun beasal dari populasi yang sama.

Dari hasil penelitian, diharapkan para pengguna teori tes klasik, khususnya staf edukatif yang berkecimpung dalam pengembangan tes, akan mendapatkan gambaran tentang hubungan antara ukuran sampel yang merespons dan parameter tes. Gambaran tersebut dapat membantu untuk mengambil keputusan yang terkait dengan analisis butir secara lebih tepat dan akurat.

## Metode Penelitian

## Data, Populasi, dan Sampel

Data penelitian ini berbentuk respons terhadap perangkat tes matakuliah Pendidikan Matematika 2 (PPDG 2332) masa uji 96.2, yang terdiri dari 60 butir soal. Tes matakuliah PPDG 2332 dan pada masa uji tersebut secara kualitatif (hasil telaah) telah cukup baik (Rustam, 1998).

Populasi dalam penelitian ini adalah seluruh respons mahasiswa terhadap tes matakuliah PPDG 2332 masa uji 96.2, yang tersebar di seluruh wilayah Indonesia. Sementara sampel diambil melalui teknik *Proportional Cluster Random Sampling*, dengan tahapan seperti berikut. Tahap awal, menentukan sampel wilayah berdasarkan kawasan Indonesia (kawasan Barat, Tengan, dan Timur). Setiap kawasan diwakili oleh tiga Unit Program Belajar Jarak Jauh (UPBJJ). Ketiga UPBJJ tersebut dipilih berdasarkan kategori UPBJJ besar, sedang dan kecil, yaitu kawasan Barat di wakili oleh UPBJJ Semarang, Bandar lampung, dan Bengkulu; kawasan Tengah diwakili UPBJJ Denpasar, Pontianak dan Palangkaraya; kawasan Timur di wakili Oleh UPBJJ Ujung Pandang, Ambon dan Kendari. Ukuran sampel setiap kawasan sebanding. Selanjutnya, sampel dalam setiap kawasan diambil secara acak yaitu masing-masing kawasan ukuran sampel sebesar 1299LJU dan proporsional. Berdasarkan teknik tersebut diperoleh ukuran sampel sebesar n = 3897 mahasiswa.

## Teknik Analisis Data.

Sebelum data dianalisis, ukuran sampel sebesar n = 3897 LJU (Lembar Jawaban Ujian) kemudian, diacak kembali dengan ukuran sampel: 50, 100, 150,..., 1500, 1600, 1700,..., 2500, 2650, 2800,..., 3897. Selanjutnya setiap ukuran sampel data dianalisis menggunakan program Iteman dari MicroCAT versi 3.00 tahun 1989. Analisis ini

digunakan untuk menentukan parameter butir dan tes. Tahap berikutnya, untuk menjawab masalah penelitian, parameter tes dikorelasikan kepada ukuran sampel dengan menggunakan uji korelasi Product Moment dari Pearson.

## Pembahasan

Penelitian yang bersifat eksploratif ini merupakan penelitian yang memaparkan hasil analisis sebagaimana adanya dan mencoba menghubungkan hasil analisis dengan teori yang ada.

Dengan demikian, hasil analisis berupa koefisien korelasi dalam tingkat signifikansi yang digunakan bukan untuk menguji hipotesis, tetapi alat untuk melihat kecenderungan dari data yang dianalisis.

Dilaihat dari butir per butir, dari 60 butir tes tersebut secara empiris tidak satu butir tes ditolak sabagai alat ukur. Tidak satu butirpun dari tes tersebut yang memiliki indeks daya beda yang negatif dan indeks tingkat kesukaran yang nol atau satu. Dengan demikian ke-60 butir soal tersebut dapat dikatakan sebagai alat ukur yang baik.

Namun demikian, kalau dilihat dari sudut kriteria butir soal yang baik menurut Pusisjian Depdikbud (1991), secara empiris tes ini memiliki butir tes yang baik sebanyak 38 butir, sisanya sebanyak 22 butir tes masih perlu direvisi. Ke-22 butir tes tersebut masih termasuk kategori sangat mudah, sangat sukar, mimiliki daya beda yang sangat rendah, dan atau keduanya (lihat Tabel 1).

**Tabel 1. Sebaran Butir Berdasarkan Kriteria**

| No | Kriteria | Σ Butir | Keputusan |
|---|---|---|---|
| 1. | 0.00<0,30 & 0,00<$r_{bis}$ <0,30 | 1 | Revisi |
| 2. | 0,30≤ P≤ 0,80 & $r_{bis}$ <,30 | 1 | Revisi |
| 3. | 0,80 < P< 1,00 & $r_{bis}$ < 0,30 | 0 | Revisi |
| 4. | 0,00 < P <0,30 & $r_{bis}$ ≥ 0,30 | 1 | Revisi |
| 5. | 0,30 ≤ P ≤ 0,80 & $r_{bis}$ ≥ 0,30 | 38 | Revisi |
| 6. | 0,80 < P < 1,00 & $r_{bis}$ ≥ 0,30 | 19 | Revisi |
| JUMLAH | | 60 | |

Keterangan: P = Indeks tingkat kesukaran butir tes

$r_{bis}$ = Indeks daya beda butir tes

Secara keseluruhan, tes PPDG 2332 memiliki konsistensi interal yang tinggi, ditunjukan oleh koefisien reliabilitas yang besarnya 0,911. Koefisien reliabilitas yang diperoleh menunjukan hasil estimasi skor amatan terhadap skor amatan terhadap skor sebenarnya yang dapat dipercaya, karena varians errornya semakin kecil. Disamping itu, sesuai

dengan pendapat Gronlund (1985), bahwa koefisien reliabilitas untuk tes buatan guru sebesar 0,6 sudah cukup memadai, sementara tes PPDG 2332 memiliki koefisien reliabilitas yang jauh lebih besar dari yang disebutkan oleh Gronlund. Selain itu, tes ini memiliki kategori tingkat kesukaran sedang dan menengah pada kategori tingkat kesukaran mudah (P=0,659), dan tingkat daya beda culup tinggi dengan koefisien biserial sebesar 0,551.

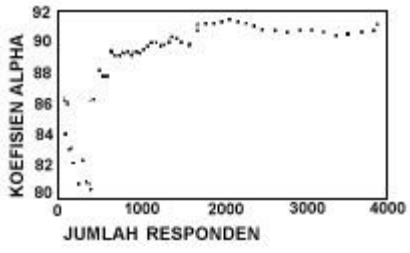
Dalam teori tes klasik berlaku kesalahan pengukuran standar yang sama untuk semua peserta tes. Hal tersebut berlaku dalam kasus ini, sehingga semua peserta tes memiliki kesalahan pengukuran standar sebesar 3,130 (skala 0-60). Dengan demikian, skor sebenarnya untuk setiap peserta tes tidak dapat dipastikan seperti skor yang diperoleh (skor amatan) dari merespons tes tersebut. Oleh karena skor yang diperoleh setiap peserta tes tidak pasti, maka skor sebenarnya berada dalam rentang tertentu, dan letaknya perlu diestimasi.

Jika diambil taraf kepercayaan 90% atau taraf signifikansi 10%, estimasi interval skor sebenarnya sebesar $X.Z_{a/2.}$ SEM $\leq$ T $\leq$ X. + $Z_{a/2}$ . SEM (Keterangan: T = skor sebenarnya; X = skor yang diperoleh; $Z_{a/2}$ = nilai kritis devisi standar normal; SEM = kesalahan pengukuran standar), maka harga rata-rata skor sebenarnya berada dalam rentang: 34,41 $\leq$ T $\leq$ 44,73. Berdasarkan rentang skor tersebut suatu skor sebenarnya tidak tinggal atau mutlak seperti yang diperoleh peserta tes, tetapi skor peserta tes tersebut berada dalam suatu rentang harga.

Konsistensi parameter tes mengukur derajat perubahan parameter tes yang diduga dipengaruhi oleh perubahan ukuran sampel (banyaknya responden) yang merespon tes tersebut. Dengan demikian, jika ukuran sampel diubah, apakah parameter tes akan mengalami perubahan ? Bagaimana kecenderungan perubahan parameter tes tersebut?.

Berdasarkan hasil analisis, antara koefisien Alpha (koefisien reliabilitas) dan jumlah responden terdapat hubungan positif (0,66) yang signifikan (P=0,00). Hal ini berarti, bahwa apabila jumlah responden bertambah atau semakin besar, maka koefisien Alpha akan semakin besar. Dilihat dari formula dasar reliabilitas, yaitu $r_{xx} = 1 - S_e{}^2/s_x{}^2$, dimana $s_e^2$ adalah varians eror dan $s_x^2$ adalah varians skor amatan (Allen dan Yen, 1979), maka sangat mungkin terjadi hubungan positif seperti hasil analisis. Merujuk formula tersebut, apabila peserta tes atau jumlah responden diperbesar, varians skor amatan akan cenderung membesar, dan selanjutnya koefisien reliabilitas juga akan membesar.

Hubungan positif antara koefisien reliabilitas dan jumlah responden, juga terlihat jelas pada scatter plot (lihatGambar1).




Gambar 1: *Scatter Plot* antara Koefisien Alpha dan Jumlah Responden

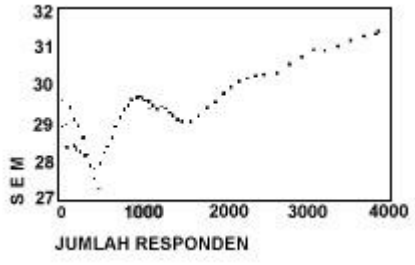
Bila diamati, pada delapan ukuran sampel pertama, tidak terlihat adanya kecenderungan koefisien reliabilitas. Namun, setelah diatas delapan ukuran sampel pertama kecenderungan positif terlihat dengan jelas. Semakin besar. Secara keseluruhan koefisien reliabilitas tidak meningkat secara ekstrim tetapi secara bertahap dalam rentang 0,81 sampai dengan 0,92. Dari koefisien korelasi dan scatter plot terlihat bahwa semakin besar. Dengan demikian koefisien reliabilitas tidak konsisten atau koefisien reliabilitas akan berubah bila jumlah responden yang merespons tes tersebut juga bertambah.

Berikutnya, antara kesalahan pengukuran standar (SEM) dan jumlah responden juga terdapat hubungan yang positif, dengan koefisien korelasi antara keduanya sebesar 0,896, dan peluang kesaalahan (P) sebesar 0,00. Ini berarti bahwa apabila ukuran sampel yang merespons tes tersebut bertambah, maka harga SEM juga akan bertambah.

Bila formula SEM = $s_x\sqrt{\quad 1 - r_{xx}}$, dikaitkan dengan hubungan korelasi yang positif antara koefisien reliabilitas dan jumlah responden (semakin banyak responden, semakin besar koefisien reliabilitas), makaterlihat bahwa korelasi positif antara jumlah responden dan SEM bukan disebabkan oleh kenaikan koefisien reliabilitas, tetapi semata-mata dikarenakan kenaikan varians skor peserta tes.

Formula SEM menunjukkan bahwa jika Sx = konstan, maka semakin besar koefisien reliabilitas akan menyebabkan semakin kecil harga SEM. Namun kenyataannya, koefisien reliabilitas semakin besar, harga SEM pun semakin besar pula. Keduanya seolah-olah bertentangan. Dengan demikian dapat disimpulkan bahwa koefisien reliabilitas bukan yang menyebabkan kenaikan harga SEM. tetapi unsur lain, yaitu varians skor peserta tes. Kondisi ini sangat mungkin terjadi, karena dengan bertambahnya ukuran sampel data cenderung semakin luas sebenarnya. Pada akhirnya apabila semakin besar varians skor peserta tes, semakin besar pula harga SEM.

Hubungan antara SEM dan jumlah responden tergambar secara detail dan jelas pada *scatter plot* (lihat Gambar 2). Semakin banyak peserta yang merespons tes tersebut semakin besar pula harga SEM, secara keseluruhan peningkatan harga SEM relatif teratur seiring dengan peningkatan jumlah peserta yang merespons tes tersebut. Perubahan atau peningkatan harga SEM tidak mencolok tidak mencolok, yakni dalam rentang 2,5 sampai dengan 3,2 dalam skala tes (0-60).




Gambar 2: *Scatter Plot* antara SEM dan Jumlah Responden

Baik hasil analisis yang berupa koefisien korelasi maupun scatter plot menunjukkan bahwa apabila jumlah responden bertambah, maka harga SEM pun bertambah besar. Hal ini berarti bahwa harga SEM tidak konsisten, harganya akan berubah kalau jumlah responden juga berubah, walaupun respons yang dianalsis berasal dari sampel pada populasi yang sama.

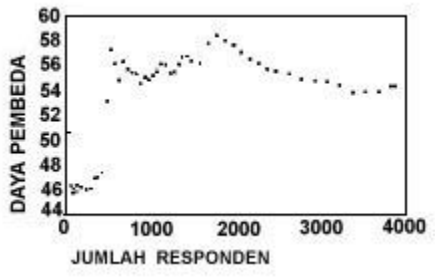
Selanjutnya, antara indeks tingkat kesukaran dan jumlah responden terdapat hubungan yang negatif, ditunjukkan oleh koefisien korelasi yang negatif sebesar -0,847 dengan peluang kesalahan P=0,00. Dengan demikian hubungan negatif keduanya sangat signifikan. Semakin banyak responden atau semakin besar ukuran sampel, akan terjadi semakin kecil indeks tingkat kesukaran (semakin tinggi taraf kesukaran) atau semakin sukar tes tersebut.

Menurut Suryabrata (1989), pada umumnya reliabilitas dan varians skor suatu tes meningkat apabila variansi taraf kesukaran butir tes menurun (indeks tingkat kesukaran meningkat). Sekilas pendapat tersebut bertentangan dengan hasil analisis yang menunjukan koefisien reliabilitas dan varians skor meningkat dan taraf kesukarn semakin meningkat (indeks tingkat kesukran semakin kecil). Tetapi bila dicermati lebih lanjut tidak demikian adanya. Varians taraf kesukaran tes cenderung kecil, karena indeks tingkat kesukarannya berada disekitar rata-rata (0,65), dan butir yang indeks tingkat kesukaran dibawah 0,50 hanya sebanyak 9 butir (15%) dari jumlah butir. Dengan demikian, dapat disimpulkan bahwa indeks tingkat kesukaran tidak konsisten terhadap perubahan ukuran sampel Bila ukuran sampel peserta tes berubah, maka indeks tingkat kesukaran juga akan berubah, walaupun penambahan ukuran sampel tersebut tetap dalam satu populasi.

Ternyata, antara indeks daya beda dan jumlah responden terdapat hubungan yang positif, yang ditunjukan oleh koefisien korelasi antara keduanya sebesar 0,415 dengan peluang kesalahan P = 0,00. Walaupun hubungan keduanya ditandai oleh koefisien korelasi tidak perlu besar, yakni hanya sebesar 0,415, namun cukup signifikan. Ini berarti bahwa apabila jumlah responden bertambah, maka indeks daya beda juga bertambah. Tetapi, penambahan indeks daya beda relatif tidak sebanding dengan penambahan jumlah responden atau indeks daya beda berfluktuasi tajam pada setiap pertambahan responden.

Hubungan antara indeks daya beda dan jumlah responden juga tergambar pada scatter plot secara jelas (lihat Gambar 3).

Pada 10 ukuran sampel pertama terjadi kenaikan indeks daya beda yang cukup tajam. Kemudian, dari ukuran sampel ke-17 sampai ukuran sampel ke- 31 juga terjadi kenaikan indeks daya beda, dan untuk ukuran sampel selanjutnya relatif terjadi penurunan indeks daya beda. Kenaikan maupun penurunan indeks daya beda mulai dari 10 ukuran sampel pertama berbeda dalam rentang indeks daya beda 0,54 sampaidengan 0,58.




Gambar 3: *Scatter Plot* antara koefisien Alpha dan Jumlah Responden

Hasil analisis sebelumnya yang menunjukkan bahwa indeks kesukaran tes berada dalam kategori sedang atau P di sekitar 0,5. Jika dikaitkan dengan formula $r_{bis}$ hubungan antara indeks daya beda dan jumlah responden menjadi semakin jelas. Indeks tingkat kesukaran setiap butir dalam tes tersebut relatif berada pada harga P = 0,5; sehingga P ( 1-P) relatif berada pada 0,25 atau mendekati konstan pada 0,25. Sementara varians skor peserta tes meningkat sesuai dengan peningkatan ukuran sampel. Karena ada dua komponen ($S_t$ dan P(1-P) dalam penentuan harga $r_{bis}$ relatif konstan untuk setiap butir. Dengan demikian, sangat mungkin terjadi hubungan yang relatif tidak besar antara indeks daya beda dan ukuran sampel.

Indeks daya beda mempunyai hubungan positif yang signifikan terhadap jumlah responden. Ini berarti bahwa apabila jumlah responden bertambah, indeks daya beda juga akan bertambah besar. Dengan demikian dapat disimpulkan bahwa indeks dayabeda tidak konsisten, harga indeks daya beda akan semakin besar jika banyak peserta tes bertambah, walaupun penambahan peserta tes dalam populasi yang sama.

## Kesimpulan

Respons yang digunakan dalam penelitian diperoleh dari respons peserta tes yang berasal dari tes yang baik secara empiris, dengan taraf kesukaran sedang dan daya beda yang cukup tinggi.

Parameter tes (indeks reliabilitas, indeks tingkat kesukaran, dan indeks daya beda) dan kesalahan pengukuran standar (SEM) akan berubah (tidak konsisten) bila ukuran sampel dan merespons tersebut berubah. Seluruh parameter tes memiliki korelasi positif terhadap jumlah responden yang merespons tes tersebut, kecuali parameter indeks tingkat kesukaran berkorelasi negatif terhadap jumlah responden. Dengan demikian, semua parameter tes akan mengalami kenaikan seiring dengan kenaikkan ukuran sampel (jumlah responden), kecuali indeks tingkat kesukaran. Semakin besar ukuran sampel, maka akan semakin kecil indeks tingkat kesukaran tes atau semakin besar ukuran sampel maka akan semakin tinggi taraf kesukaran tes tersebut.

Penelitian ini merupakan penelitian awal dan memiliki keterbatasan, yaitu variasi skor atau kemampuan mahasiswa yang menjadi sampel dalam penelitian ini tidak dikontrol dan respons peserta tes hanya diambil dari satu matakuliah. Oleh karena itu diharapkan ada penelitian lanjutan dengan mengontrol variasi skor dan respons peserta tes diambil dari beberapa matakuliah.

## Daftar pustaka


Anastasi, A. (1988). *Psychological testing*. New York: Macmillan publishing Company.

Allen, M.J. and Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Azwar, S (1992). *Reliabilitas dan validitas*. Yogyakarta: Sigma Alpha.

. (1996). *Tes prestasi fungsi dan pengembangan pengukuran prestasi belajar*. Yogyakarta: Pustaka Belajar.

Croker, L. and Algina, J. (1986). *Introduction to classical and Modern test theory*. New

York: Holt, Rinehart and Winston, Inc.

Depdikbud. (1995). Bank soal. *Buletin pengujian dan penilaian)*, halaman 3-7. Jakarta: Pusjian, Balitbang, Depdikbud.

Gronlund, N.E ( 1985). *Measurement and evaluation in teaching. New York*: Macmillan Publishing Company.

Hopkins, K.D., Stanley, J.C and Hopkins, B.R. (1990). *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice Hall Inc.

Masrun. (1975). *Analisis Item*. Yogyakarta: Fakultas Psikologi UGM.

Naga, D.S (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma.

Rustam. (1998). Karakteristik Tes Program Penyetaraan D2 PGSD Universitas Terbuka: Implementasi Model Rasch. *Jurnal Penelitian dan Evaluasi*, 93-104. Tahun I, PPS IKIP Yogyakarta.

Suryabrata, S. (1987). *Pengembangan tes hasil belajar*. Jakarta: CV Raja Wali.

Zainul, A . dan Nasution, N. (1993). *Penilaian hasil belajar*. Jakarta: PAU, PPAI, DIKTI, Depdikbud.