Ketidaktepatan Penggunaan Validitas Butir dan Koefisien Reliabilitas dalam Penelitian Pendidikan dan Psikologi

Dali S. Naga

Abstract: Item validity is applied in educational and psychological research through item analysis to enhance the reliability of respondent scores. Recently, there are a couple of inappropriate ways of treating item validity, which neither ensures the reliability nor justifies the validity of the measurement. Similar cases also happen in the application of reliability coefficient.

Kata kunci: validitas butir, koefisien reliabilitas, kesalahan pengukuran, penelitian pendidikan.

Validitas butir digunakan pada analisis butir dalam rangka uji coba pengukuran untuk memperbaiki alat ukur. Melalui validitas butir, ada butir yang dapat dipertahankan di dalam alat ukur serta ada butir yang perlu dibuang, diperbaiki, atau diganti. Diharapkan melalui uji coba dan perbaikan yang berulang-ulang, kita akan memperoleh alat ukur yang baik dan dapat dipercaya.

Dalam bentuk statistika, validitas butir dinyatakan dalam bentuk koefisien korelasi di antara skor-satuan butir ke-i dengan skor responden A (skor total). Karena itu, koefisien korelasi ini dikenal juga sebagai koefisien korelasi butir-total, ρ_{iA} atau r_{iA}, seperti tampak pada Gambar 1.

Di dalam penelitian pendidikan dan mungkin juga psikologi, belakangan ini, muncul dua ketidaktepatan terhadap penggunaan validitas butir. Demikian

---

Dali S. Naga adalah dosen Universitas Negeri Jakarta.

pula telah muncul ketidaktepatan dalam penggunaan koefisien reliabilitas. Ketidaktepatan itu adalah sebagai berikut. Pertama, di dalam analisis butir pada uji coba pengukuran, butir dipertahankan di dalam alat ukur melalui penolakan hipotesis H_0 untuk r_{iA} pada taraf signifikansi α tertentu. Kedua, validitas butir digunakan untuk menentukan validitas pengukuran sebagai pengganti validitas isi, kriteria, atau konstruk. Ketiga, koefisien reliabilitas dianggap memadai jika hipotesis H_0 untuk koefisien reliabilitas itu berhasil ditolak pada taraf signifikansi α tertentu.

Tulisan ini bertujuan untuk menjelaskan mengapa tiga hal ini tidak tepat sehingga perlu dihindari di dalam penelitian pendidikan dan psikologi.

Res-pon-den	Butir											A
	1	2	3	.	.	.	i	j	.	.	.	N	
1							X						X
2							X						X
.							.						.
.							.						.
.							.						.
g							X						X
h							X						X
.							.						.
.							.						.
.							.						.
M							X						X

ρ_{iA} atau r_{iA}

Gambar 1 Koefisien korelasi butir-total untuk butir ke-i

PERANAN VALIDITAS BUTIR

Melalui koefisien korelasi butir-total, ρ_{iA} atau r_{iA} pada Gambar 1, validitas butir adalah korelasi di antara skor-satuan butir ke-i dengan skor responden A. Apa artinya kalau nilai koefisien korelasi butir-total adalah positif tinggi?

Skor responden A dapat kita susun dalam urutan peringkat (dari tinggi ke rendah atau sebaliknya). Skor-satuan pada butir ke-i dapat juga kita susun dalam urutan peringkat. Koefisien korelasi butir-total akan bernilai positif tinggi jika kedua peringkat itu mirip atau cukup konsisten. Dengan kata lain,

jika koefisien korelasi butir-total bernilai tinggi, maka skor tinggi pada butir ke-i berpasangan dengan skor tinggi pada responden A (yakni banyak responden yang menjawab betul). Demikian pula, skor rendah pada butir ke-i berpasangan dengan skor rendah pada skor responden A (yakni banyak responden yang menjawab salah).

Dengan demikian, butir itu memiliki daya untuk mengenal skor responden tinggi (melalui banyak jawaban benar) dan skor responden rendah (melalui banyak jawaban salah). Itulah sebabnya, koefisien korelasi butir-total (ρ_{iA} atau r_{iA}) atau validitas butir dikenal juga sebagai daya beda butir (validitas butir = daya beda butir = koefisien korelasi butir-total). Untuk menghindari pencampuradukan *validitas butir* dengan *validitas pengukuran*, penulis lebih menyukai istilah daya beda butir daripada istilah validitas butir.

Selanjutnya, apa dampak atau peranan dari koefisien korelasi butir-total di dalam penelitian? Kita mulai dengan memisalkan bahwa semua butir di dalam alat ukur memiliki validitas butir positif tinggi. Dalam hal ini, setiap skor-satuan butir, masing-masing, memiliki koefisien korelasi yang positif tinggi dengan skor responden. Peringkat skor-satuan pada setiap butir konsisten dengan peringkat skor responden. Akibatnya, peringkat skor-satuan di antara sesama butir juga saling konsisten. Karena itu, koefisien korelasi di antara butir atau interkorelasi butir (di antara butir ke-i dan ke-j) berupa ρ_{ij} atau r_{ij} juga bernilai positif tinggi.

Apa dampak atau peranan interkorelasi butir yang bernilai positif tinggi di dalam penelitian? Melalui hubungan statistika

$$\sigma_{ij} = \rho_{ij}\ \sigma_i\ \sigma_j$$

kita temukan bahwa interkorelasi ρ_{ij} yang bernilai positif tinggi menyebabkan kovariansi di antara butir juga bernilai positif tinggi. Hal ini dapat kita kaitkan dengan koefisien reliabilitas alpha Cronbach ρ_{α} dan koefisien reliabiltas Kuder-Richardson $\rho_{KR\text{-}20}$. Rumus koefisien reliabilitas ini dapat kita tulis sebagai berikut.

$$\rho_{\alpha} = \frac{N}{N-1}\frac{\sigma_A^2 - \sum \sigma_i^2}{\sigma_A^2} = \frac{N}{N-1}\frac{2\sum_{i<j}\sigma_{ij}}{\sigma_A^2}$$

$$\rho_{KR-20} = \frac{N}{N-1}\frac{\sigma_A^2 - \sum p_i q_i}{\sigma_A^2} = \frac{N}{N-1}\frac{2\sum_{i<j}\sigma_{ij}}{\sigma_A^2}$$

dengan N sebagai banyaknya butir dan σ_A^2 sebagai variansi pada skor responden A.

Dari rumus di atas tampak bahwa interkorelasi atau kovariansi butir σ_{ij} yang tinggi menyebabkan koefisien reliabiltas menjadi tinggi. Jadi, validitas butir atau daya beda butir atau koefisien korelasi butir-total yang positif tinggi berdampak kepada atau berperan pada peningkatan koefisien reliabilitas.

Selain melalui validitas butir, koefisien reliabilitas dapat juga ditingkatkan melalui perpanjangan alat ukur. Dengan memperpanjang dua paruhan setara (dengan koefisien korelasi paruh-paruh ρ_{pp}) pada alat ukur menjadi L bagian setara, maka melalui rumus koefisien reliabilitas Spearman-Brown

$$\rho_{SB} = \frac{L\rho_{pp}}{1+(L-1)\rho_{pp}}$$

koefisien reliabilitas dapat ditingkatkan.

Dengan demikian, koefisien reliabilitas pengukuran dapat ditingkatkan melalui validitas butir yang tinggi, dan perpanjangan alat ukur. Jadi, peran dan fungsi validitas butir atau daya beda butir atau koefisien korelasi butir-total adalah untuk peningkatan reliabilitas pengukuran.

KETIDAKTEPATAN PENGGUNAAN VALIDITAS BUTIR DAN RELIABILITAS

Pertama, prosedur uji hipotesis terhadap koefisien korelasi butir-total r_{iA} menghasilkan statistik

$$t = \frac{r_{iA}\sqrt{n-2}}{\sqrt{1-r_{iA}^2}}$$

dengan n sebagai banyaknya responden dan ukuran sebesar 10 kali jumlah butir atau minimal sebesar 5 kali jumlah butir di dalam alat ukur (Nunnally, 1970: 214-215).

Kalau nilai t cukup besar (melebihi t_{tabel}) maka pada taraf signifikansi α tertentu, hipotesis H_0 dapat ditolak. Memang benar bahwa nilai t dapat diperbesar melalui r_{iA} yang besar atau koefisien reliabilitas yang tinggi. Karena itu, penolakan H_0 menunjukkan bahwa butir itu layak dipertahankan di dalam alat ukur.

Meskipun demikian, masih ada cara lain untuk memperbesar nilai t. Dari rumus tampak bahwa nilai t dapat juga diperbesar melalui peningkatan n atau peningkatan banyaknya responden. Walaupun nilai r_{iA} kecil, kalau n cukup besar, maka nilai t menjadi cukup besar sehingga mampu menolak H_0.

Variabel n yang besar dengan nilai r_{iA} yang kecil yang mampu menolak hipotesis H_0 menghasilkan koefisien reliabiltas yang rendah.

Di sinilah terletak ketidaktepatannya. Penolakan hipotesis H_0 belum dapat menjamin peningkatan koefisien reliabilitas karena n yang besar dengan r_{iA} yang kecil juga mampu menolak hipotesis H_0. Penolakan hipotesis H_0 dengan n yang cukup besar tetapi dengan nilai r_{iA} yang cukup kecil mampu mempertahankan butir yang tidak baik di dalam alat ukur. Dengan kata lain, tidak jelas bagi kita apakah tertolaknya hipotesis H_0 itu karena r_{iA} yang tinggi ataukah karena n yang besar.

Sebenarnya keberatan terhadap pensampelan (sampling) responden seperti ini telah dikemukan oleh Nunnally (1970: 15). Menurut Nunnally, pengkaji psikologi sering terjebak pada anggapan bahwa reliabilitas suatu ujian meningkat dengan banyaknya orang yang digunakan di dalam studi reliabilitas. Selanjutnya Nunnally juga menyatakan bahwa "perkiraan reliabilitas yang diperoleh pada suatu studi adalah independen terhadap banyaknya orang di dalam studi melainkan, di setiap studi, reliabilitas berhubungan dengan banyaknya butir di dalam ujian," seperti tampak pada rumus koefisien reliabilitas Spearman-Brown di atas.

Hal ini telah kita lihat pada uraian di atas. Karena itu, pada sejumlah bacaan, kriteria untuk mempertahankan butir di dalam alat ukur ditentukan oleh nilai koefisien korelasi butir-total. Kriteria empirik mencakup nilai 0,20 (Aiken, 1997: 65; Crocker & Algina, 1986: 324; Nunnally, 1970: 202) atau nilai 0,25 (Henning, 1987: 53). Sekali lagi, kriteria untuk mempertahankan butir di dalam alat ukur bukan ditentukan melalui penolakan hipotesis H_0.

Kedua, validitas butir atau daya beda butir atau koefisien korelasi butir-total hanya berbicara tentang hubungan di antara skor-satuan pada butir dengan skor responden. Apapun yang diungkapkan oleh skor-satuan pada butir dan skor responden tidak menjadi masalah. Selama korelasi di antara mereka bernilai positif tinggi, selama itu pula validitas butir adalah tinggi. Validitas butir akan tetap tinggi sekalipun skor responden (dan skor-satuan pada butir) tidak mengukur sasaran yang hendak diukur.

Di sinilah letak ketidaktepatannya. Validitas butir melalui koefisien korelasi butir-total tidak mampu menjelaskan apakah skor-satuan pada butir dan skor responden telah mengukur apa yang memang hendak diukur. Validitas butir tidak dapat menjamin apakah pengukuran telah mengukur apa yang seharusnya diukur.

Validitas pengukuran perlu dilakukan melalui validitas yang telah kita kenal (validitas isi, kriteria, konstruk) dengan prosedur yang berkaitan dengan validitas pengukuran. Mereka tidak dapat digantikan dengan validitas butir.

Ketiga, pada dasarnya, koefisien reliabilitas adalah koefisien korelasi terhadap pengukuran itu sendiri (Naga, 1997) baik pada butir yang sama maupun pada butir yang setara. Di sini, koefisien reliabilitas itu (ukur-ukur ulang, ukur-ukur setara, Spearman-Brown, alpha Cronbach, dan Kuder-Richardson) kita nyatakan dengan ρAA atau r_{AA}. Seperti halnya pada ketidaktepaan pertama, statistik r_{AA} ini adalah

$$t = \frac{r_{AA}\sqrt{n-2}}{\sqrt{1-r_{AA}^2}}$$

dengan n sebagai banyaknya responden, berukuran sebesar 5 sampai 10 kali jumlah butir di dalam uji coba pengukuran. Tampak dari statistik itu, jika nilai t cukup besar sehingga melampaui t_{tabel}, maka hipotesis H_0 dapat ditolak. Memang benar bahwa nilai koefisien reliabilitas r_{AA} yang tinggi dapat meningkatkan nilai t. Namun masih ada cara lain untuk meningkatkan nilai t. Sekalipun nilai r_{AA} kecil, jika nilai n cukup besar, maka nilai t dapat juga ditingkatkan sehingga mampu menolak hipotesis H_0.

Di sinilah letak ketidaktepatannya. Penolakan hipotesis H_0 tidak selalu menjamin koefisien reliabilitas yang tinggi. Dengan n yang cukup besar serta koefisien reliabilitas yang rendah pun, hipotesis H_0 mampu ditolak. Padahal koefisien reliabilitas yang rendah tidak kita kehendaki di dalam penelitian. Dengan kata lain, tidak jelas bagi kita apakah tertolaknya hipotesis H_0 karena r_{AA} yang tinggi ataukah karena n yang besar.

Sesungguhnya hal ini telah juga dikemukan oleh Nunnally (1970: 15) bahwa "di dalam studi tentang reliabilitas dari suatu pengukuran baru, diperlukan penentuan berapa reliabilitasnya; hanya pernyataan bahwa koefisien reliabilitas berbeda secara signifikan terhadap nol adalah hampir tidak berguna."

Untuk mengatasi ketidaktepatan pada validitas butir ini, kita menggunakan kriteria empirik yang telah dikemukakan oleh sejumlah penulis seperti yang telah disinggung di atas. Mereka mencakup banyaknya responden uji coba dan ukuran daya beda butir. Sekalipun telah dikemukakan di atas, angka-angka itu ditampilkan sekali lagi dalam bentuk kriteria empirik berikut ini.

Dalam hal banyaknya responden, Nunnally (1970: 214-215) menyatakan bahwa ukuran responden pada uji coba adalah sebesar sepuluh kali jumlah butir.

Jadi, untuk uji coba alat ukur 50 butir, misalnya, diperlukan 10 x 50 = 500 responden. Namun apabila uji coba itu akan melibatkan banyak sekali responden, minimal ukuran responden adalah lima kali jumlah butir. Jadi, untuk uji coba alat ukur 100 butir, minimal diperlukan 5 x 100 = 500 responden.

Crocker dan Algina (1986: 322) membahas ukuran yang dikemukakan oleh Nunnally serta menambahkan bahwa demi kestabilan informasi, minimal diperlukan 200 responden. Jadi, sekalipun alat ukur mengandung hanya 20 butir, minimal diperlukan 200 responden. Lebih dari itu, kriteria Nunnally digunakan untuk menentukan jumlah responden selanjutnya.

Sekiranya kita menghendaki kestabilan yang lebih tinggi, kita dapat menggunakan kriteria yang dikemukakan oleh Davis (1966: 283) untuk kelompok skor tinggi dan kelompok skor rendah yang biasa digunakan pada daya beda butir. Davis menganjurkan 100 responden untuk masing-masing 27% kelompok skor tinggi dan kelompok skor rendah sehingga keseluruhannya mencakup minimal 371 responden atau dapat kita bulatkan menjadi 400 responden.

Ada beberapa penulis mengemukakan kriteria empirik untuk menentukan batas validitas butir dalam mempertahankan butir di dalam alat ukur. Crocker dan Algina (1986: 324) mengemukakan angka minimum 0,2. Nunnally (1970: 202) mengemukakan angka minimum 0,2. Aiken (1994: 65) mengemukakan angka minimum 0,2. Mehrens dan Lehmans (1991: 167) mengemukakan angka minimum 0,2. Henning (1987: 53) mengemukakan angka minimum 0,25. Tampak di sini bahwa mayoritas penulis buku mengemukakan kriteria empirik sebesar 0,2.

PENUTUP

Tulisan ini menunjukkan dan menjelaskan ketidaktepatan yang terjadi di sekitar penggunaan validitas butir dan koefisien reliabilitas. Penelitian pendidikan dan psikologi berikut laporan hasil penelitian hendaknya menghindari ketidaktepatan ini. Pengujian hipotesis untuk r_{iA} pada taraf signifikasi tertentu tidak menjamin kelayakan reliabilitas pada pengukuran. Validitas butir tidak juga dapat menjamin validitas pengukuran sehingga tidak dapat digunakan sebagai pengganti validitas pengukuran (isi, kriteria, atau konstruk). Demikian pula pengujian hipotesis untuk r_{AA} pada taraf signifikansi tertentu tidak menjamin kelayakan reliabilitas pada pengukuran di dalam penelitian.

Khusus mengenai validitas butir atau daya beda butir, sejumlah penulis mengemukakan kriteria empirik berupa jumlah responden di atas 200 (atau 400)

untuk selanjutnya bertambah dengan minimal lima kali jumlah butir. Sejumlah penulis juga mengemukakan kriteria empirik sebesar 0,2 untuk validitas butir atau daya beda butir atau koefisien korelasi butir-total. Kriteria empirik ini dapat digunakan untuk menghindari ketidaktepatan yang sedang terjadi sekarang ini di dalam penelitian pendidikan dan psikologi.

DAFTAR RUJUKAN

Aiken, L.R. 1997. *Psychological Testing and Assessment.* Boston: Allyn and Bacon.

Crocker, L. & Algina, J. 1986. *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

Davis, F.B. 1966. *Item Selection Technique: Educational Measurement.* Washington, D.C.: American Council on Education.

Henning, G. 1987. *A Guide to Language Testing: Development, Evaluation, Research.* Cambridge: Newbury House Publishers.

Mehrens, W.A. & Irvin, J.L. 1991. *Measurement and Evaluation in Education and Psychology.* Fort Worth: Hartcourt Brace College Publishers.

Naga, D.S. 1997. The Misuses of Reliability Coefficient and Sampling Variance in Educational Research. *The Journal of Education*, 4 (Special Edition): 305-309.

Nunnally Jr., J.C. 1970. *Introduction to Psychological Measurement.* New York: McGraw-Hill Book Company.