Melibatkan Rater dalam Pengembangan Alat Ukur

Wahyu Widhiarso | wahyu_psy@ugm.ac.id
Fakultas Psikologi Universitas Gadjah Mada

1 Pengantar

Tulisan berikut akan membahas masalah bagaimana melibatkan rater dalam pengembangan alat ukur dan koefisien psikometris apa saja yang terkait dengannya.

Mengapa Melibatkan rater ? Ada banyak pertimbangan mengapa melibatkan rater. Berikut ini dua alasan mengapa peneliti melibatkan rater.

a. Meningkatkan kualitas alat ukur yang dikembangkan
   Melibatkan pakar dalam menilai butir-butir yang kita tulis akan memastikan bahwa butir yang kita buat relevan dengan apa yang kita ukur dan mewakili keseluruhan domain ukur. Misalnya, meminta praktisi di bidang marketing untuk mengevaluasi butir skala kepuasan konsumen akan memastikan bahwa butir-butir yang kita tulis mewakili indikator-indikator konsumen yang puas.
b. Jenis alat ukur yang dikembangkan
   Jika *self report* adalah instrument yang diisi sendiri oleh responden, maka instrumen observasi menggunakan rater untuk memberikan penilaian.

2 Aplikasi

Properti psikometris yang biasa dipakai untuk mengevaluasi alat ukur adalah reliabilitas dan validitas. Pelibatan rater dalam pengembangan alat ukur membantu kita untuk mengevaluasi alat ukur yang kita kembangkan. Fungsi rater tergantung pada kebutuhan kita, rater sebagai penilai instrument yang kita kembangkan ataukah rater sebagai pemberi skor instrument observasi. Penilaian rater terhadap instrument biasanya dinamakan dengan *judgement professional* karena mereka memiliki kapabilitas dalam hal konstrak yang kita ukur. Rater yang bertugas memberikan skor tidak harus profesional di bidang itu, tetapi bisa juga individu yang terlatih untuk mengobservasi dalam bidang yang kita ukur.

2.1 Studi Validitas

2.1.1 Indeks Rasio Validitas Isi (CVR)

Dalam pendekatan ini, sebuah panel *subjek-matter experts* diminta untuk menunjukkan apakah suatu aitem pengukuran dalam satu skala lainnya adalah “penting” sebagai bentuk operasionalisasi bangunan teori. Masukan panel ini kemudian digunakan untuk menghitung CVR untuk setiap item engan calon dalam instrumen pengukuran. Untuk mengukur CVR, sejumlah ahli (panel) diminta untuk memeriksa setiap item pada instrumen pengukuran. Penyekoran terdiri dari tiga alternatif, yaitu aitem tertentu adalah relevan, kurang relevan atau tidak relevan dengan domain yang diukur. Penyekoran ini dilakukan terhadap semua aitem. Prosedur menghitung CVR bisa dilihat di Widhiarso (2010)

Kategorisasi Nilai CVR Lawshe (1975) menyajikan sebuah tabel CVR nilai minimum berdasarkan uji signifikansi satu ekor dengan p =. 05. Karena nilai CVR tergantung pada jumlah panel maka nilai CVR tergantung pada jumlah panel yang digunakan . Sebagai contoh, Lawshe menyimpulkan bahwa nilai CVR dari 0,29 akan baik-baik untuk 40 panelis yang digunakan, sebuah CVR dari 0,51 akan cukup dengan 14 panelis, tapi CVR minimal 0,99 akan diperlukan dengan tujuh atau lebih sedikit Panelis. Jelas, berikut rekomendasi

Lawshe yang ketat akan membutuhkan sejumlah besar rater. Perhatikan bahwa, dalam prakteknya, positif CVR nilai-nilai yang lebih rendah dalam besarnya dari yang dibutuhkan dengan menggunakan kriteria Lawshe yang kadang-kadang digunakan sebagai dasar untuk berdebat untuk bukti validitas isi ketika sejumlah relatif kecil rater digunakan untuk memberikan peringkat.

2.2 Studi Reliabilitas

Studi reliabilitas yang melibatkan rater biasanya dinamakan dengan kesepakatan antar rater (*inter rater agreement*) atau reliabilitas antar rater (*inter-rater reliability*). Jika pada kasus *self-report* reliabilitas ditunjukkan dengan konsistensi internal yang terlihat dari antara satu butir dan butir lainnya memiliki korelasi yang tinggi, maka dalam kasus reliabilitas antar rater yang diuji konsistensinya adalah raternya. Jadi posisi butir digantikan dengan posisi orang (rater).

Rater-rater yang memiliki kesepakatan tinggi terlihat dari posisi subjek yang diobservasi. Jika urutan skor subjek dari Rater A dan B hampir sama maka kedua rater memiliki kesepakatan yang tinggi (Ebel & Frisbie, 1991). Hal ini dikarenakan kesepatakan dioperasionalkan dalam bentuk korelasi. Dari 10 siswa yang diobservasi oleh dua orang rater. Jika penilaian rater A terhadap siswa P paling tinggi dibanding dengan siswa lainnya, dan Rater B juga demikian maka kedua rater dapat dikatakan konsisten. Dalam hal ini masalah apakah Rater A pelit memberikan skor sedangkan rater B tergolong murah dalam memberikan skor tidak mempengaruhi.

2.2.1 Koefisien Kappa dari Cohen

Cohen (1960) mengembangkan koefisien untuk mengukur kesepakatan antar rater yang kemudian dikenal dengan koefisien kappa.

Penggunaan Koefisien kappa tepat digunakan ketika (a) rater yang dipakai tidak banyak. Biasanya satu subjek dinilai oleh dua rater. (b) Skor hasil penilaiannya bersifat kategori. Biasanya juga hanya dua kategori yang dikode 0 atau 1.

Contoh Kasus Dua orang psikolog menilai 10 orang klien apakah mereka perlu mendapatkan terapi ataukah tidak. Pada kasus ini terlihat jumlah raternya adalah dua dan kategori skor juga dua.

Contoh Ouput Analisis Berikut ini contoh hasil analisis dengan menggunakan program SPSS. Hasil analisis menunjukkan kesepakatan antar rater sebesar 0.72. Tampak bahwa kedua psikolog memiliki indeks kesepakatan yang cukup tinggi.

Count

		Rater_B		
		Tidak Perlu Terapi	Perlu Terapi	Total
Rater_A	Tidak Perlu Terapi	5	1	6
	Perlu Terapi	1	8	9
Total		6	9	15

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.722	.183	2.797	.005
N of Valid Cases		15			

Kategorisasi Ada dua orang ahli yang mengkategorikan nilai koefisien kappa yaitu Landis dan Koch (1977) dan Fleiss (1975). Menurut Landis dan Koch (1977) kategori nilai kappa adalah sebagai berikut :

$\kappa < 0.00$ poor agreement
$0.00 < \kappa < 0.20$ slight
$0.21 < \kappa < 0.40$ fair
$0.41 < \kappa < 0.60$ moderate
$0.61 < \kappa < 0.80$ substantial, and
$0.81 < \kappa < 1.00$ almost perfect agreement.

Menurut Fleiss (1981) kategori nilai adalah sebagai berikut :

$\kappa < 0.40$ poor agreement
$0.40 < \kappa < 0.75$ good, and
$\kappa > 0.75$ excellent agreement.

Prosedur selengkapnya menghitung koefisien Kappa bisa melihat pada tulisan Widhiarso (2005)

2.2.2 Koefisien Korelasi intra Kelas

Koefisien korelasi intra kelas (*intraclass correlation coefficients; ICC*) yang dikembangkan oleh Pearson (1901). Koefisien ini dikembangkan berdasarkan analisis varians namun pada kasus tertentu hasilnya memiliki kemiripan dengan koefisien alpha.

Penggunaan Koefisien ICC tepat digunakan ketika (a) rater yang dipakai banyak dan (b) skor hasil penilaiannya bersifat kontinum.

Contoh Kasus Enam orang psikolog menilai 10 orang anak apakah berapa banyak perilaku bullying yang mereka lakukan pada satu periode amatan. Contoh hasil pengamatan di tampilkan pada Tabel di bawah ini.

Subjek	Rater_1	Rater_2	Rater_k
1	1	2	1
2	4	4	3
3	3	3	3
n	5	5	6

Contoh Ouput Analisis Berikut ini contoh hasil analisis dengan menggunakan program SPSS. Hasil analisis menunjukkan rata-rata kesepakatan antar rater sebesar 0.972 sedangkan untuk satu orang rater konsistensinya adalah 0.811.

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.811	.783	.837	35.287	299.0	2093	.000
Average Measures	.972	.967	.976	35.287	299.0	2093	.000

Prosedur selengkapnya menghitung ICC bisa melihat pada tulisan Widhiarso (2005)

3 Penutup

Kesepakatan antar rater dapat dicapai ketika masing-masing rater memiliki persepsi yang sama terhadap apa yang dinilai dan diobservasi. Oleh karena itu panduan observasi yang jelas dan operasional serta pembekalan yang cukup kepada rater sebelum terjun ke lapangan akan sangat membantu dalam mendapatkan nilai kesepakatan antar rater yang tinggi.

DAFTAR PUSTAKA

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. New Jersey: Prentice Hall.

Fleiss, J. L. (1975). Measuring Agreement between Two Judges on the Presence or Absence of a Trait. *Biometrics*, *31*(3), 651 - 659.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159 - 174.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, (28), 563-575.

Pearson, K. (1901). Mathematical distributions to the theory of evolution. *Philosophical Transactions of the Royal Society of London*, *197*, 385-497.

Widhiarso, W. (2010). Prosedur Pengujian Validitas Isi melalui Indeks Rasio Validitas Isi (CVR). Retrieved February 22, 2011, from http://wahyupsy.blog.ugm.ac.id/2010/06/16/prosedur-pengujian-validitas-isi-melalui-indeks-rasio-validitas-isi-cvr/.

Widhiarso, W. (2005). Mengestimasi Reliabilitas. Yogyakarta: Fakultas Psikologi UGM.