

# KERANCUAN DALAM PENGGUNAAN ISTILAH “CONSTRUCT RELIABILITY”

**Jahja Umar**

Institut Asesmen Indonesia

umarindo@me.com

## **Abstract**

*In the classic theory where “true-score” for every person that had been tested cannot be obtained, the reliability coefficient is needed so that test result user can see how far the data can be trusted. For example, like making “confident interval” for true score in test score that already obtained. Next, researcher that will be analyzed the result data statistically, with regression analysis for example, can even use reliability coefficient test score to check bias in regression coefficient because of the low score of variable coefficient that used as predictor. So, when the data that analyzed is total-score, researcher have to report the reliability coefficient. Most popular reliability coefficient is Cronbach  $\alpha$ , that can estimate if some of the assumption is required, one of them is that all test item is parallel and assumption that error measurement cannot be correlated. With the advancement of psychometry subject over years, reliability coefficient can be estimate more accurate even when one or many of the assumption cannot be required. One of them is the reliability that measure each item’s representation, that many called it “construct reliability”. The word “construct” is happened to misleded it. This article explain that commonly there are no difference between coefficient and the “new” coefcicient.*

**Keyword:** Reliability, Construct Reliability, Total Score, True-Score

## **Abstrak**

*Dalam teori tes klasik di mana “true-score” untuk setiap orang yang di tes tidak dapat diperoleh, koefisien reliabilitas diperlukan agar pengguna hasil tes dapat melihat sejauh mana data yang dimiliki dapat dijadikan pegangan / dipercaya. Misalnya dengan membuat “confident interval” untuk true-score bagi skor tes (yang dalam hal ini adalah skor total dari skor item) yang diperoleh. Selanjutnya, peneliti yang akan menganalisis data hasil tes secara statistik, misalnya dengan analisis regresi, bahkan dapat menggunakan koefisien reliabilitas skor tes tersebut untuk mengkoreksi bias yang terjadi pada koefisien regresi akibat rendahnya reliabilitas skor variabel yang dijadikan prediktor. Oleh sebab itu, ketika data yang dianalisis adalah skor tes (skor-total) seorang peneliti atau pengguna tes harus melaporkan koefisien reliabilitas itu. Koefisien reliabilitas yang sangat populer adalah Cronbach  $\alpha$ , yang dapat di estimate jika beberapa asumsi terpenuhi, diantaranya asumsi bahwa seluruh item tes adalah paralel dan asumsi bahwa kesalahan pengukuran tidak saling berkorelasi. Seiring dengan kemajuan di bidang psikometri, misalnya dengan dikembangkannya metode Analisis Faktor Konfirmatorik (CFA), koefisien reliabilitas dapat diestimasi dengan lebih akurat meskipun satu atau beberapa dari asumsi tersebut tak dapat terpenuhi. Salah satunya adalah koefisien reliabilitas jika asumsi paralelitas tak terpenuhi, yaitu koefisien dengan memperhitungkan bobot setiap item, yang kemudian ada yang menyebutnya dengan*

istilah “construct reliability”. Penggunaan kata “construct” di sini ternyata telah menimbulkan kesalahan dalam menafsirkan istilah tersebut. Tulisan ini berisi uraian untuk menjelaskan bahwa pada dasarnya tak ada perbedaan yang mendasar antara koefisien seperti Cronbach  $\alpha$  dan sejenisnya dengan koefisien “baru” tersebut.

**Kata Kunci:** Reliabilitas, Realibilitas Konstruk, Skor-Total, True-Score

Diterima: 24 Maret 2014

Direvisi: 29 April 2014

Disetujui: 8 Mei 2014

## PENDAHULUAN

Seperti telah banyak diketahui, persamaan dasar dari teori tes klasik adalah  $X = T + E$  di mana  $X$  adalah skor tes,  $T = \text{true score}$ , dan  $E = \text{measurement error}$ . Dengan persamaan ini, nilai  $T$  tak dapat diperoleh karena persamaannya bersifat tak ada solusi (*underidentified*). Tetapi, jika tes terdiri dari beberapa item, misalnya sebanyak  $k$  item, dengan skor setiap item =  $y_i$ , dan skor tes =

$$X = \sum_{i=1}^k y_i, \text{ yang merupakan skor-total (disebut “composite – score”), yang jika}$$

disertai dengan beberapa asumsi (misalnya: *parallel item* dan *uncorrelated random error*) sehingga  $\text{Var}(X) = \text{Var}(T) + \text{Var}(E)$ , maka persamaan tsb dapat diselesaikan walaupun hanya dalam bentuk solusi terhadap variance dari  $E$ , yaitu  $\text{Var}(E) = S_e^2 = \sum \text{var}(y_i)$ , dan pada gilirannya korelasi antara  $T$  dan  $X$  dapat di estimate dengan rumus (Lord, 1980):

$$r_{XT} = \sqrt{\frac{S_T^2}{S_X^2}} = \sqrt{\left(1 - \frac{\sum \text{var}(y_i)}{S_X^2}\right)}$$

Dimana  $S_T^2$  adalah  $\text{Var}(T)$ , dan  $S_X^2 = \text{Varians} (\sum y_i) = \text{Varians} (y_1 + y_2 + y_3 + \dots + y_k)$ , yang dapat dituliskan:  $S_X^2 = \sum_{i=1}^k S_{y_i}^2 + \sum_{i=1}^g 2S_{y_i y_j}$ , dimana  $S_{y_i}^2$  adalah varians item  $i$ , dan  $S_{y_i y_j}$  adalah kovarians antara item  $i$  dan item  $j$ ,

sedangkan  $g = k(k-1)/2$ , yaitu banyaknya korelasi antara item. Misalkan tes terdiri dari tiga item  $y_1, y_2$ , dan  $y_3$ , maka varians dari skor-total  $X$  adalah:  $S_X^2 =$  Varians  $(y_1 + y_2 + y_3)$ , atau dapat dituliskan dengan  $S_X^2 = (S_{y_1}^2 + S_{y_2}^2 + S_{y_3}^2 + 2S_{y_2y_1} + 2S_{y_3y_1} + 2S_{y_3y_2})$ . Dalam persamaan di atas,  $r_{XT}$  disebut koefisien validitas dari skor-total  $X$  dalam upaya mengukur *true-score*  $T$ , sedangkan  $r_{xx}^2 = (1 - \frac{\sum \text{var}(y_i)}{S_x^2})$  disebut koefisien reliabilitas dari skor-total  $X$ , yaitu besarnya proporsi (persentase) dari varians  $X$  yang merupakan varians  $T$ , dan sering dituliskan dengan simbol  $r_{xx}$ . Salah satu bentuk yang paling populer dari  $r_{xx}$  ini adalah Cronbach  $\alpha$  (Cronbach, 1951). Ada satu hal yang paling penting untuk dicatat di sini: “bahwa yang dapat diperoleh dalam teori tes klasik bukanlah true-scores, melainkan hanya interval dari kemungkinan true-scores untuk sebuah skor-total (composite) pada taraf keyakinan tertentu”. Misalkan ada seorang mendapat skor total = 27 pada sebuah tes terdiri dari 40 item yang memiliki  $r_{xx} = 0,8$ , maka dengan asumsi bahwa kesalahan distribusi adalah mengikuti kurva normal dan dengan tingkat keyakinan 95%, *true-score* orang itu adalah dalam interval  $27 \pm 1.96 \times SEM$ , dimana  $SEM$  (Standard Error of Measurement) =  $\sqrt{(1-0.8)} = 0.04$ . Kelemahan utama dari konsep reliabilitas seperti ini adalah bahwa  $SEM$  itu berlaku bagi semua orang yang menempuh tes walaupun skor yang mereka peroleh berbeda-beda.

## DISKUSI

Jika item dalam suatu tes tidak paralel, maka persamaan  $X = T + E$  tidak dapat berlaku umum dan setiap item harus diperhitungkan bobotnya, sehingga persamaan harus dibuat pada level skor-item yaitu:  $y_i = \lambda_i T + e_i$ , dimana  $\lambda_i$  adalah bobot item  $i$  yang dalam hal ini adalah koefisien regresi linier (yang dalam skala standardized adalah korelasi karena hanya ada satu independent variable yaitu  $T$ ). Jika tersedia tiga atau lebih item dan dihitung matrik korelasi antar item (dalam sampel), maka dengan menggunakan metode analisis faktor konfirmatorik (CFA dengan model satu faktor) akan dapat di estimate koefisien  $\lambda_i$  dan varians ( $e_i$ ). Dalam hal ini, varians ( $e_i$ ) biasanya diberi simbol  $\theta_{ii}$ , karena ia merupakan elemen diagonal yang ke  $i$  dari  $\Theta$ , yaitu matrik kovarians antar residual  $e_i$ .

Jika estimate  $\lambda_i$  dan  $\theta_{ii}$  sudah diperoleh, maka rumus reliabilitas  $r_{xx} =$

$$r_{xt}^2 = \left(1 - \frac{\sum \text{var}(y_i)}{S_x^2}\right), \text{ dapat dituliskan menjadi } r_{XT}^2 = \left(1 - \frac{\sum_{i=1}^k \theta_{ii}}{\left(\sum_{i=1}^k \lambda_i\right)^2 + \sum_{i=1}^k \theta_{ii}}\right)$$

dan akhirnya dapat ditulis dengan  $r_{XT}^2 = \frac{\left(\sum_{i=1}^k \lambda_i\right)^2}{\left(\sum_{i=1}^k \lambda_i\right)^2 + \sum_{i=1}^k \theta_{ii}}$ , di mana:

$$\sum \theta_{ii} = \sum \text{var}(y_i), \text{ dan } S_x^2 = \left(\sum_{i=1}^k \lambda_i\right)^2 + \sum_{i=1}^k \theta_{ii}, \text{ (menggunakan persamaan}$$

$$y_i = \lambda_i T + e_i).$$

Persamaan reliabilitas seperti ini dapat dilihat, misalnya, dalam Lord dan Novick (1968). Untuk pembuktian persamaan tentang  $S_x^2$  tersebut, seperti pada contoh sebelumnya, misalkan ada tes terdiri dari tiga item  $y_1, y_2,$  dan  $y_3$  maka:

$$S_x^2 = \text{Varians} (y_1 + y_2 + y_3) = (S_{y_1}^2 + S_{y_2}^2 + S_{y_3}^2 + 2S_{y_2 y_1} + 2S_{y_3 y_1} + 2S_{y_3 y_2}).$$

Kalau setiap  $y_i$  diganti dengan  $\lambda_i T + e_i$  maka akan diperoleh:

$$S_{y_1}^2 = \lambda_1^2 + \theta_{11}, S_{y_2}^2 = \lambda_2^2 + \theta_{22}, S_{y_3}^2 = \lambda_3^2 + \theta_{33}, \text{ dan}$$

$$S_{y_2 y_1} = \lambda_2 \lambda_1, S_{y_3 y_1} = \lambda_3 \lambda_1, S_{y_3 y_2} = \lambda_3 \lambda_2, \text{ (untuk rinciannya, lihat Umar, 2011).}$$

Dengandemikian,  $S_X^2 = (S_{y_1}^2 + S_{y_2}^2 + S_{y_3}^2 + 2S_{y_2 y_1} + 2S_{y_3 y_1} + 2S_{y_3 y_2})$  dapat ditulis

dengan:

$$S_X^2 = \lambda_1^2 + \theta_{11} + \lambda_2^2 + \theta_{22} + \lambda_3^2 + \theta_{33} + 2\lambda_2 \lambda_1 + 2\lambda_3 \lambda_1 + 2\lambda_3 \lambda_2$$

$$= (\lambda_1^2 + \lambda_2^2 + \lambda_3^2 + 2\lambda_2 \lambda_1 + 2\lambda_3 \lambda_1 + 2\lambda_3 \lambda_2) + (\theta_{11} + \theta_{22} + \theta_{33})$$

$$= (\lambda_1 + \lambda_2 + \lambda_3)^2 + (\theta_{11} + \theta_{22} + \theta_{33}) = \left( \sum_{i=1}^k \lambda_i \right)^2 + \sum_{i=1}^k \theta_{ii}, \text{ yang dalam hal ini}$$

$k=3$ .

Reliabilitas “skor-total” untuk tes yang itemnya tidak memenuhi asumsi paralelitas di mana tak boleh digunakan rumus Cronbach  $\alpha$ , tetapi harus dihitung dengan menggunakan koefisien  $\lambda_i$  sebagai bobot (weight) bagi setiap item  $i$ ,

yang diperoleh dengan rumus :  $r_{xx} = r_{XT}^2 = \frac{\left( \sum_{i=1}^k \lambda_i \right)^2}{\left( \sum_{i=1}^k \lambda_i \right)^2 + \sum_{i=1}^k \theta_{ii}}$  seperti inilah yang

kadang-kadang disebut dengan istilah “*Construct Reliability*” (awalnya oleh Fornell & Larcker, 1981). Istilah ini kemudian ditemukan jugamisalnya dalam Raykov (1997; 2001) dan Hancock dan Mueller (2001). Penamaan “*construct reliability*” sebenarnya kurang tepat karena ternyata telah menimbulkan salah interpretasi dimana sebagian orang justru menafsirkannya sebagai “reliabilitas dari *true-score*”. Padahal koefisien ini adalah tetap saja merupakan “reliabilitas dari skor-total”,(yaitu  $r_{XX}$ , bukan  $r_{TT}$ ) dan sama sekali tak ada bedanya dengan Cronbach  $\alpha$  kecuali adanya aspek pembobotan akibat tidak paralelnya item tes. Konteksnya pun tetap pada teori tes klasik, di mana koefisien reliabilitas adalah diperlukan untuk skor tes (“skor total”) ketika “*true-score*” tak dapat diperoleh.

Oleh sebab itu, dalam konteks di mana “*true-score*” untuk setiap orang bisa diperoleh (misalnya pada *Item Response Theory* dan *Confirmatory Factor Analysis*), jelas sekali bahwa koefisien “reliabilitas skor-total”, baik yang diperoleh melalui rumus seperti Cronbach  $\alpha$  maupun rumus dengan pembobotan item seperti di atas, adalah tidak relevan dan “sama sekali tak diperlukan”. Pada konteks di mana “*true-score*” untuk setiap individu bisa di estimate, biasanya sekaligus diperoleh pula *standard error* (SEM) yang berbeda untuk setiap orang. Kesimpulannya, jika dalam suatu kegiatan penelitian atau asesmen tidak digunakan “skor- total”, maka adalah tidak ada relevansinya (dan tentu saja tidak perlu) bagi peneliti untuk: (1) melaporkan koefisien reliabilitas “skor-total” tersebut, atau (2) mengkaitkan hasil penelitian/asesmen dengan koefisien reliabilitas “skor-total” baik yang diperoleh melalui rumus Cronbach  $\alpha$  maupun rumus dengan pembobotan butir soal seperti di atas. Oleh sebab itu, Raykov (1997) menyebutnya dengan istilah “*Composite Reliability for Congeneric Measures*” (“*composite*” artinya skor-total sebuah tes, sedangkan “*congeneric*” artinya tes yang memenuhi asumsi unidimensionalitas tetapi itemnya tidak paralel).

Kasus “*construct reliability*” seperti diuraikan di atas terjadi ketika para ahli psikometri mencoba menjembatani teori baru yaitu “*structural equation modeling*” dan “*confirmatory factor analysis*” (konteks di mana “*true-scores*” dapat di estimate/ dihitung untuk setiap individu) denganteori tes tradisional (konteks di mana “*true-scores*” itu tidak diperoleh). Tujuannya adalah agar mereka yang secara tradisi telah begitu terfokus kepada konsep reliabilitas  $r_{XX}$  (yang memerlukan banyak asumsi) dapat menemukan semacam analoginya pada teori yang baru (di mana sebagian asumsi tersebut tak diperlukan).

Fenomena yang mirip dengan hal yang dibahas dalam tulisan ini juga terjadi dalam konteks *Item Response Theory* (IRT) di mana “*estimated true-score*” per individu juga diperoleh. Seperti diketahui, pada IRT terdapat konsep tentang “*Item Information Function*” (IIF). Pada setiap item dapat dihitung

koefisien “IIF”, yang jika dijumlahkan menjadi suatu koefisien yang disebut “Test Information Function” (TIF). Koefisien TIF ini sangat berguna untuk memilih himpunan soal yang paling sesuai bagi kelompok peserta tes dengan tingkat kemampuan tertentu (jika tersedia bank soal yang terkalibrasi). Secara logika, konsep tentang TIF ini juga mirip dan sering dianalogikan dengan konsep reliabilitas ( $r_{xx}$ ) seandainya orang menggunakan “skor-total” dari tes tersebut dalam penelitian atau pengetesan. Akibatnya, banyak penulis buku IRT mengilustrasikan bahwa “*Standard Error of Measurement*” yang diperoleh

dalam bentuk:  $SEM = \sqrt{\frac{1}{TIF}}$ , ada kemiripannya dengan konsep reliabilitas

pada teori tes klasik. Tentu saja menggelikan jika kedua hal tersebut lalu dianggap sama, karena dalam IRT justru untuk setiap individu dapat di estimate sebuah “*true-score*”, lengkap dengan standard error nya. Jadi di dalam konteks IRT sama sekali tidak terdapat konsep tentang *standard error* yang berlaku bagi semua orang. Mungkin maksud para penulis buku IRT itu juga sama seperti pada kasus “*construct reliability*”, yaitu untuk memberikan sejenis “jembatan” kepada pembaca yang baru beralih (naik kelas) dari “aplikasi Teori Tes Klasik” ke “aplikasi IRT”. Tetapi akan menjadi salah kaprah jika kemudian sebagian

orang menafsirkan bahwa  $SEM = \sqrt{\frac{1}{TIF}}$  berlaku dan dapat ditafsirkan sebagai

*SEM* yang didapat dari koefisien reliabilitas ( $r_{xx}$ ) bagi skor total.

Semoga tulisan ini dapat memberikan kejelasan kepada peneliti dan pengguna skor tes, terutama mereka yang tidak menggunakan skor-total (misalnya ketika menguji hipotesis dalam “*Structural Equation Models With Latent Variables*”, atau ketika melaporkan hasil tes psikologi dalam bentuk “*true-score*”). Selain itu, semoga tulisan ini dapat pula memberikan kejelasan kepada para pengajar mata kuliah “psikometri” tentang perbedaan antara teori tes klasik di satu sisi (di mana *true-score* tidak diperoleh) dan “*Item Response Theory*” serta “*Confirmatory Factor Analysis*” di sisi yang lain (di mana “*true-score*”

beserta “*standard error*” nya dapat diperoleh untuk setiap individu yang menempuh tes).

## DAFTAR PUSTAKA

- Cronbach, L.J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 3, 297 – 334.
- Fornell, C. and Larcker, D.F. 1981. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, 18, 39 – 50.
- Hancock, G. R. and Mueller, R.O. 2001. Rethinking Construct Reliability Within Latent Variable Systems. In R. Cudeck, S. Du Toit, and D. Sorbom (Eds.): *Structural Equation Modeling: Present and Future*. Lincolnwood, IL: Scientific Software International, Inc.
- Lord, F.M. 1980. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Lord, F.M.; Novick, M.R. 1968. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.
- Raykov, T. 1997. Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement*, 21, 173 – 184.
- Raykov, T. 2001. Studying Change in Scale Reliability for Repeated Multiple Measurements Via Covariance Structure Modeling. In R. Cudeck, S. Du Toit, and D. Sorbom (Eds.): *Structural Equation Modeling: Present and Future*. Lincolnwood, IL: Scientific Software International, Inc.
- Umar, J. 2011. Bahan Pelatihan Analisa Faktor. Tidak diterbitkan. Jakarta: Institut Asesmen Indonesia.

