Analisis Butir dalam Pengembangan Pengukuran Psikologi

Oleh : Wahyu Widhiarso
Fakultas Psikologi UGM
Tahun 2010

Analisis aitem adalah suatu proses yang menguji respon subjek terhadap aitem yang dibuat yang bertujuan untuk menilai kualitas dari item-item dan tes secara keseluruhan. Analisis aitem sangat penting dalam meningkatkan kualitas aitem yang akan digunakan kembali dalam pengukuran formal (pengukuran sebenarnya) nantinya. Analisis aitem juga mampu mengeliminasi aitem yang ambigu, menyesatkan dan tidak relevan baik dengan konstrak maupun dengan subjek yang diukur. Analisis aitem penting untuk meningkatkan pengalaman dan keterampilan si pengembang instrumen ukur dalam mengidentifikasi domain-domain konstrak yang ukur, mana domain yang perlu mendapatkan penekanan dan mana yang tidak perlu. Dari paparan diatas dapat kita simpulkan bahwa analisis aitem tidak hanya mengidentifikasi relevansi aitem dengan konstrak ukur akan tetapi juga relevansi aitem dengan sampel yang diukur. Butir “saya menghargai upaya atasa saya” bisa jadi akan gugur dalam analisis jika sampel yang digunakan adalah siswa, sebaliknya aitem tersebut tidak gugur jika sampelnya adalah pegawai.

A. STATISTIK BUTIR

Berikut ini akan dijelaskan statistik yang diidentifikasi dalam analisis aitem.

1. Jumlah Butir.

Jumlah aitem menjadi pertimbangan karena menyangkut kualitas pengukuran misalnya domain ukur. Domain ukur yang sempit hanya membutuhkan aitem yang sedikit dibanding dengan domain ukur yang luas. Skala Self-Efficacy for Self-Regulated Learning yang dikembangkan oleh Bandura (2001) hanya berisi 11 aitem dibanding dengan Skala General Perceived Self-Efficacy Scale Jerusalem dan Schwarzer (1981). Selain itu pengukuran unidimensi akan membutuhkan banyak aitem dibanding pengukuran multidimensi. Alat ukur multidimensi (multifaktor) seperti EPPS, MMPI, Big Five Personality menggunakan aitem yang banyak dibanding dengan skala psikologi yang memusatkan pada satu trait/abilitas. Dalam analisis aitem kita mempertimbangkan apakah jumlah aitem skala yang kita kembangkan sudah cukup kuat untuk merepresentasikan domain ukur. Untuk mengantisipasi akan banyaknya aitem yang gugur, disarankan untuk melipatgandakan dari jumlah aitem target dalam skala kita. Kalau kita mentargetkan skala kita berisi 15 aitem, aitem yang kita ujicobakan bisa berjumlah 30 (2x lipat) atau 15 aitem (3 kali lipat). Proporsi jumlah aitem pada tiap aspek pengukuran diupayakan agar sama jika kita tidak membobot aspek-aspek ukur. Kalaupun ada aspek yang memiliki penekanan lebih dibanding dengan aspek lainnya usakanlah agar jumlah aitem yang sudah jadi sesuai dengan proporsi bobot aspeknya.

2. Rerata dan Deviasi Standar Butir

Rerata dan deviasi standar penting untuk diidentifikasi dalam analisis aitem. Rerata menunjukkan kecenderungan respons maupun skor skala. Devasi standar adalah ukuran penyebaran skor subjek pada aitem yang menunjukkan bagaimana respons atau skor

menyebar. Semakin tinggi nilai deviasi standar menunjukkan respons subjek terhadap aitem bervariasi. Dalam pengembangan instrumen pengukuran tingginya deviasi standar dapat menunjukkan sampel yang kita miliki trait yang bervariasi.

3. Tingkat Kesulitan Butir

Tingkat kesulitan aitem banyak dipakai pada pengukuran abilitas dari pada kepribadian. Tingkat kesulitan aitem adalah persentase siswa yang menjawab item dengan benar yang berkisar dari 0 hingga 100. Semakin tinggi tingkat nilai tingkat kesukaran, semakin mudah aitem tersebut. Dalam skala psikologi, tingkat kesukaran aitem dipakai untuk skala yang menggunakan model Skala Guttman dimana tingkat kesukaran aitem tergradasi dari yang paling mudah hingga paling sulit (scalogram).

Dalam tes prestasi tingkat kesukaran disesuaikan dengan tujuan tes. Tingkat kesulitan aitem dalam tes yang bertujuan untuk seleksi, lebih sulit dibanding dengan tujuan skrining. Tingkat kesulitan aitem pada tes formatif (tes yang menekankan pada konten) lebih bervariasi dibanding dengan tes dengan tujuan penempatan. Namun demikian tingkat kesulitan aitem yang optimal adalah 0.5 dengan rentang yang dapat diterima adalah antara 0.3 hingga 0.7. Lord (1952) menuliskan tingkat kesulitan aitem optimal berdasarkan karakteristik tes.

- Lima-respon pilihan ganda – 0.70
- Empat respons pilihan ganda – 0.74
- Tiga respons pilihan ganda – 0.77
- Dua respons pilihan berganda (Benar-salah) – 0.85

4. Daya Diskriminasi Butir

Daya diskriminasi aitem menunjukkan sebarapa aitem mampu membedakan subjek dengan abilitas/trait tinggi dan rendah. Dalam pendekatan teori skor murni klasik, daya diskriminasi menjadi fokus utama untuk mengevaluasi aitem (dieliminasi atau tidak), sedangkan dalam teori modern daya diskriminasi digunakan untuk mengetahui aitem tersebut bekerja dengan baik pada abilitas/trait level berapa. Dalam tes prestasi daya diskriminasi aitem mengacu pada kemampuan suatu item untuk membedakan antara subjek atas dasar seberapa baik mereka tahu materi yang diuji.

Korelasi aitem total biasa dipakai dalam analisis ini, untuk skala psikologi biasanya menggunakan korelasi product moment sedangkan pada tes prestasi menggunakan korelasi biserial. Disarankan kedua korelasi ini telah terkoreksi oleh spurious overlap. Program SPSS telah menyediakan korelasi product moment aitem-total terkoreksi (corrected aitem-total score) namun tidak menyediakan korelasi biserial dalam menunya. Untuk menghitung korelasi biserial, anda dapat menggunakan program ITEMAN. Namun demikian ada formula yang dapat mentransformasi nilai korelasi product moment menjadi korelasi biserial.

Besarnya daya beda yang direkomendasikan oleh para ahli adalah di atas 0.3. Namun demikian anda masih memungkinkan untuk melibatkan nilai dibawah itu, misalnya 0.275 dengan beberapa pertimbangan. Misalnya untuk mengatasi proporsi aspek ukur yang belum imbang (jika setiap aspek memiliki bobot sama), kita bisa melibatkan nilai korelasi aitem-total dengan nilai tersebut, agar tidak terjadi ketimpangan jumlah aitem dalam aspek ukur. Selain itu kita juga bisa mengeliminasi aitem meski di atas 0.3 jika jumlah aspek-aspek telah mencukupi. Kita

menggugurkan aitem yang memiliki korelasi aitem total paling rendah dibanding teman-temannya di dalam aspek yang sama.

Indeks diskriminasi aitem mencerminkan sejauh mana aitem dan tes secara keseluruhan mengukur atribut ukur, maka nilai koefisien ini akan cenderung lebih rendah untuk tes mengukur atribut yang heterogen (multidimensional). Jika anda mendapati banyak aitem anda berguguran, bisa jadi karena yang anda kembangkan adalah pengukuran multidimensional. Oleh karena itu beberapa ahli menyarankan untuk mendeteksi dimensionalitas data melalui analisis faktor (lihat Wahyu, 2010).

Indeks diskriminasi aitem harus ditafsirkan dalam konteks jenis tes yang sedang dianalisis. Daya diskriminasi yang rendah dapat menunjukkan aitem yang ditulis ambigu atau tidak sesuai dengan karakteristik subjek sehingga item wordingnya harus diperbaiki. Daya diskriminasi tinggi akan tetapi memiliki arah negatif harus diperiksa untuk menentukan mengapa aitem tersebut mampu membedakan subjek kemampuan subjek dengan baik akan tetapi arahnya negatif. Bisa jadi hal ini dikarenakan kesalahan skoring (aitem favorable diskor unfavorable, atau sebaliknya), atau konten aitem tersebut tidak sesuai dengan budaya lokal. Mengkritik atasan mungkin tidak tepat budaya tertentu akan tetapi tepat untuk budaya yang lain. Permasalahan ini sebenarnya dapat diantisipasi ketika kita tidak hanya begutu saja menggunakan aspek-aspek dari teori barat kemudian menerjemahkannya dalam penulisan aitem, namun ada proses adaptasi yang sesuai dengan budaya kita.

B. PERINGATAN

Statistik-statistik harus ditafsirkan dalam konteks jenis tes yang diberikan dan individu yang diukur. Artinya penafsiran hanya dapat diberikan pada alat ukur dan sampel yang kita pakai dalam mengembangkannya. Masih ada peringatan yang lain, berikut ini peringatan-peringatan lainnya.

1. Analisis aitem data tidak sama dengan validitas item. Kriteria eksternal diperlukan untuk secara akurat menilai validitas item tes. Dengan menggunakan kriteria internal total nilai tes, analisis aitem mencerminkan konsistensi internal dari aitem bukan validitas. Banyak sekali yang mengatakan bahwa korelasi aitem-total adalah validitas isi. Jika skor total tersebut adalah alat ukur itu sendiri korelasi-aitem total bukanlah validitas, namun jika skor total tersbut didapatkan dari kriteria (misalnya alat ukur terstandarisasi) korelasi aitem-total baru dapat dimaknai sebagai koefisian validitas.

2. Indeks diskriminasi bukanlah ukuran kualitas item karena ada berbagai alasan yang memungkinkan sebuah aitem memiliki daya diskriminasi rendah, misalnya (a) aitem tersebut terlalu sulit atau terlalu mudah sehingga tidak mampu membedakan kemampuan subjek. Terkadang dalam situasi tes skrining, aitem yang memiliki daya beda rendah pun perlu dilibatkan karena aitem tersebut memiliki konten yang vital sehingga perlu ditanyakan meski itu memiliki indeks diskriminasi rendah. Anekdotnya, aitem “saya mampu memencet tombol” mungkin akan memiliki daya beda rendah, karena hampir semua orang mampu memencet tombol.

Apakah kita akan menggugurkan aitem ini jika aitem ini adalah aitem untuk mengukur kemampuan pilot? Seorang dokter yang memulai mengenali penyakit pasien dari pertanyaan yang umum “apakah ada keluhan?” hingga pertanyaan yang khusus, “apakah dada anda terasa sesak setiap pagi?”. Meski pertanyaan “apakah ada keluhan” memiliki daya beda rendah

(karena semua orang pasti akan memiliki keluhan jika menemui dokter(namun pertanyaan ini penting sehingga tidak boleh digugurkan dalam prosedur identifikasi penyakit. Jadi, aitem yang memiliki daya diskriminasi yang rendah dalam hal ini tidak mendukung model pengukuran yang kita pakai saat ini. Bisa jadi model pengukuran yang lain bisa menerima aitem tersebut. (b) aitem dapat menunjukkan diskriminasi yang rendah bisa disebabkan alat ukur yang kita kembangkan memiliki kandungan berbagai atribut ukur (multidimensi). Untuk menganalisis model ini, kita dapat melakukan analisia aitem secara terpisah berdasarkan dimensi atau faktornya.

3. Analisis adalah masalah tentatif. Data tersebut dipengaruhi oleh jenis dan jumlah sampel yang dipakai dalam pengembangan (e.g try out), prosedur instruksional yang digunakan, dan kesalahan kesempatan.

Jena, 21 November 2010

REFERENSI

Bandura, A. (2000). Self-efficacy. In A. E. Kazdin (Ed.), Encyclopedia of psychology. New York: Oxford University Press.

Lord, F.M. (1952). The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties. Psychometrika, 18, 181-194.

Widhiarso, W. (2008). Reliability Measurement in Personality Multidimensional Measurement. Psikobuana. Vol.1. 39-48 PDF