Tidak sedikit buku yang beredar dipasaran yang membahas dan mengulas tentang analisis, penelitian dan penilaian dan atau kajian statistik. Akan tetapi dari sekian banyak buku yang beredar tersebut, hanya sebagian kecil saja yang dapat dengan mudah dipahami dan dimengerti oleh khalayak pembaca. Hal ini tentu tidak terlepas dari bagaimana si penulis mampu mengemas topik pembahasannya, juga sejauhmana si penulis dapat memberikan contoh dan data-data pendukung yang akurat dan komprehensif.

Buku *Analisis Kuantitatif Instrumen Penelitian* ini merupakan buku panduan bagi peneliti, mahasiswa, dan psikometrian untuk melakukan analisis kuantitatif instrumen yang digunakan, khususnya untuk pengukuran pendidikan. Buku ini dikemas dengan bahasa pembahasan yang simple, mudah dipahami namun tetap komprehensif. Untuk mempermudah pembaca dalam memahami kajiannya, buku ini dilengkapi dengan data bahan praktik yang dapat digunakan pembaca untuk berlatih. Data-data tersebut dapat diunduh di laman http://www.uny.ac.id/ atau http://www.staff.uny.ac.id/ heriretnawati .

Adapun isi pokok pembahasan dalam buku ini sebagaimana tertuang dalam judul bab-bab pembahasan berikut:

Jl. Sadewa No. 1
Sorowajan Baru, Yogyakarta
HP. 0812 2815 3789
email: nuhamedika@gmail.com

parama publishing

ISBN 978-602-1547-98-4
9 786021 547984

Heri Retnawati

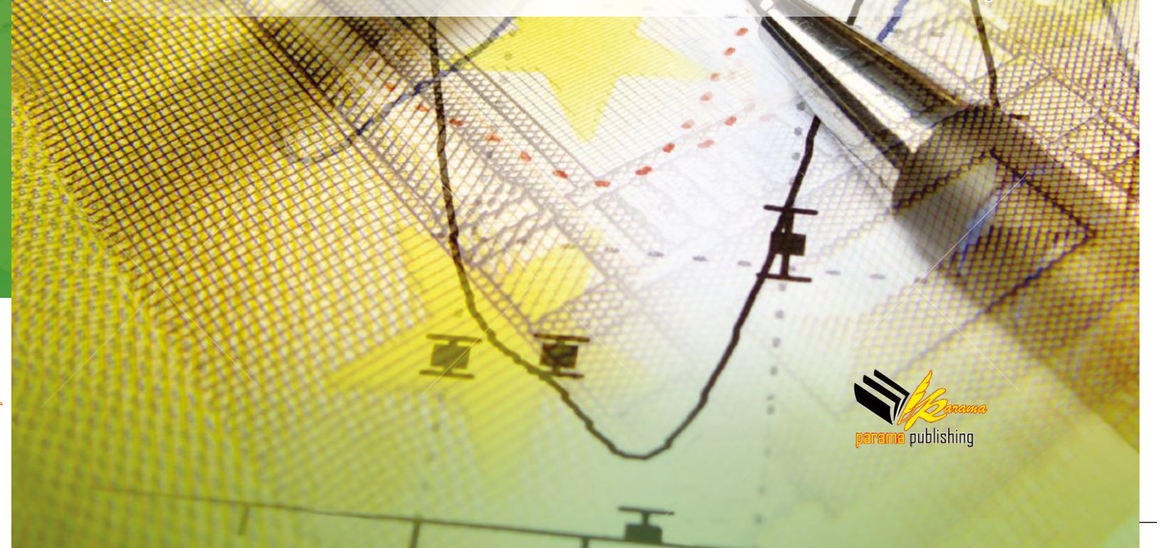
ANALISIS KUANTITATIF INSTRUMEN PENELITIAN

parama publishing

Heri Retnawati

ANALISIS KUANTITATIF INSTRUMEN PENELITIAN

(Panduan Peneliti, Mahasiswa, dan Psikometrian)



parama publishing

Heri Retnawati

ANALISIS KUANTITATIF INSTRUMEN PENELITIAN

(Panduan Peneliti, Mahasiswa, dan Psikometrian)

parama publishing

ANALISIS KUANTITATIF INSTRUMEN PENELITIAN

(Panduan Peneliti, Mahasiswa, dan Psikometrian)

Penulis : Heri Retnawati

Sampul : arteholic numed
Layout : @bay

Cetakan : Pertama, Januari 2016
ISBN : 978-602-1547-98-4

Diterbitkan

Parama Publishing
Jl. Sadewa No. 1 Sorowajan Baru, Yogyakarta
Telp. 0812 2815 3789
email: nuhamedika@gmail.com - nuhamedika@yahoo.com
facebook: www.facebook.com/nuhamedika
homepage: www.nuhamedika.gu.ma

Kata Pengantar

Alhamdulillah, segala puji dan syukur terpanjatkan kepada Allah subhanahu wata'ala atas segala karunia yang diberikan, sehingga buku ini dapat terselesaikan. Buku ini merupakan buku panduan bagi peneliti, mahasiswa, dan psikometrian untuk melakukan analisis kuantitatif instrumen yang digunakan, khususnya untuk pengukuran pendidikan.

Buku ini dilengkapi dengan data bahan praktik yang dapat digunakan pembaca untuk berlatih. Data ini dapat diunduh di laman http://www.uny.ac.id/ atau http://www.staff.uny.ac.id/heriretnawati .

Terimakasih yang tak terhingga kami sampaikan kepada Wakil Rektor I UNY atas bantuan pendanaan untuk penulisan buku ini, Bapak Dr. Hartono selaku Dekan FMIPA UNY atas dukungannya, Bapak Dr. Samsul Hadi atas diskusinya mengenai confirmatory factor analysis, Ibu Dr. Yulia Ayriza dan Ibu Dr. Farida Agus Setyawati sebagai validator instrument self regulated learning yang dijadikan contoh validasi instrument dalam buku ini, dan Bapak Dr. Edi Istiyono atas masukannya baik sebagai validator maupun reviewer buku ini keseluruhan.

Terimakasih atas dukungan dan pengertian dari orangtua kami ayahanda Sugeng Tasijo, Ibunda Warsiyem, dan Ibunda Sujami, Mas Ahmad Madani dan Adik Fatma Fauzia atas motivasinya untuk menjadi "Guru yang Baik", Mas Fauzan Ahmad atas nasehat-nasehatnya untuk bersabar.

Semoga amal kebaikan Bapak/Ibu, saudara/I semua diberikan pahala dan kebaikan yang lebih dari Allah subhanahu wata'ala. Segala masukan dan kritik yang membangun kami harapkan dari pembaca untuk perbaikan buku ini di edisi mendatang.

Yogyakarta, 22 November 2015

Heri Retnawati

Daftar Isi

Bab I
PENGEMBANGAN INSTRUMEN PENELITIAN

Dalam suatu penelitian pendidikan, proses pengumpulan data merupakan suatu hal yang sangat penting. Data yang dikumpulkan sangat terkait dengan fenomena, yang menjadi fokus penelitian. Data ini dimanfaatkan untuk membuat kesimpulan, sesuai dengan tujuan penelitian yang telah ditetapkan oleh peneliti dan menguji hipotesis yang telah dirumuskan (Wiersma, 1986). Sebagai contoh misalnya pada penelitian eksploratif, hasil pengumpulan data ini digunakan untuk penyimpulan dengan membuat deskripsi untuk mengeksplorasi hal-hal terkait dengan permasalahan penelitian. Pada penelitian positivistik, hasil pengumpulan data dianalisis dengan uji statistik tertentu, hasil analisis digunakan untuk membuat kesimpulan.

Dalam pelaksanaan penelitian, peneliti sebaiknya terfokus pada permasalahan penelitian yang akan dipecahkan. Masalah penelitian menentukan jenis data yang diperlukan, dan jenis data ini memandu pemilihan metode atau cara pengumpulan data (Babbie: 2004). Jenis data yang dimaksud yakni data nominal, data ordinal, data interval dan data rasio. Data nominal merupakan ukuran diskrit (terpisah antar data), tidak ada hubungan antara skala yang satu dengan skala yang lain. Contoh data nominal misalnya agama, warna pakaian atau kendaraan, jenis kelamin, hobi, dan lain-lain. Data ordinal merupakan ukuran yang menunjukkan posisi suatu objek, dengan ukuran tersebut dapat diurutkan dari urutan paling rendah sampai yang paling tinggi, namun belum ada jarak atau interval antara posisi ukuran yang satu dengan yang lain. Contoh data ini misalnya skala Likert (Sangat Setuju, Setuju, Ragu-ragu, Tidak Setuju, Sangat Tidak Setuju), dimana belum ada jarak yang jelas antara tidak setuju dengan sangat tidak setuju, dan juga skala lainnya.

Data interval merupakan ukuran yang menunjukkan posisi suatu objek dalam suatu urutan paling rendah sampai yang paling tinggi, dan ada jarak atau interval antara posisi ukuran yang satu dengan yang lain. Contoh data ini adalah nilai/skor dalam pendidikan. Pada data interval, nilai nol juga bukan nilai yang mutlak, yang berarti bahwa seorang peserta didik memeroleh skor nol, belum tentu peserta didik tersebut sama sekali tidak menguasai komptetensi dalam pembelajaran, namun bisa jadi karena alasan lain. Pada data rasio, ukuran menunjukkan posisi suatu objek dalam suatu skala paling rendah sampai skala yang paling tinggi, ada jarak atau interval antara posisi ukuran yang satu dengan yang lain, dan adanya besaran

absolute/mutlak. Sebagai contoh pada data rasio adalah ukuran volume air. Volume air dalam suatu wadah sama dengan nol berarti air dalam wadah tersebut memang telah kosong, atau tidak ada air sedikitpun dalam wadah tersebut.

Jenis data tersebut berdampak pada pelaksanaan pengukuran dalam penelitian. Sebagai contoh seorang peneliti ingin mengetahui kemampuan berpikir tingkat tinggi (*higher order thinking*) siswa SMP di kabupaten Subur Makmur. Fokus permasalahan yang menjadi kata kunci penelitian ini adalah kemampuan berpikir tingkat tinggi. Ini berarti data yang harus dikumpulkan peneliti tersebut adalah data kemampuan berpikir tingkat tinggi siswa SMP. Karena kemampuan berfikir tingkat tinggi merupakan kemampuan yang abstrak, diperlukan suatu tes untuk mengukurnya. Kemampuan ini dapat diukur dengan teknik tes, dan data yang kita peroleh berupa data interval. Pada kasus lain, seorang peneliti ingin mengetahui motivasi kerja karyawan. Permasalahan yang menjadi fokus penelitian adalah motivasi, yang dapat diukur dengan angket/kuisioner motivasi. Untuk pengumpulan data ini, perlu digunakan teknik nontes.

Pengumpulan data sangat terkait dengan kegiatan pengukuran (*measurement*). Pengukuran dilaksanakan untuk mengetahui kemampuan atau performa dari sesuatu atau seseorang, baik berupa kemampuan, sikap, keterampilan, persepsi, dan lain-lain. Pengumpulan data pada dasarnya dikategorikan menjadi 2 teknik, yakni teknik tes dan nontes. Teknik tes dengan menggunakan instrumen tes, baik tes lisan, tulisan, atau tes berbasis komputer (*computer-based testing, CBT*) dan ada pula tes adaptif berbasis computer (*computer adaptive test, CAT*). Untuk instrumen non tes, dapat dikategorikan menjadi angket, wawancara, observasi, dan dokumentasi. Instrumen pengumpulan data ini masing-masing disajikan berikut ini.

Pengumpulan data dengan teknik tes ini dilakukan dengan melakukan pengujian pada responden penelitian. Tes ini biasanya dilakukan untuk melihat kemampuan responden penelitian. Sebagai contoh kemampuan kognitif, menggunakan berbagai tes seperti tes kemampuan bahasa Inggris, tes kemampuan matematika, tes kemampuan membaca, tes bakat akademis, dan lain-lain. Tes-tes ini merupakan salah satu bentuk instrumen, terdiri dari sejumlah pertanyaan, atau butir-butir soal digunakan untuk memperoleh data atau informasi melalui jawaban peserta tes. Melalui hasil jawaban tersebut, diperoleh suatu ukuran mengenai karakteristik peserta tes.

Ada dua tipe tes, yakni tes objektif dan tes uraian (essay, disebut pula dengan *constructed response*). Tes objektif merupakan tes yang telah disediakan pilihan

jawabannya. Tes objektif dapat berbentuk tes benar salah, tes pilihan ganda, tes menjodohkan, dan tes isian singkat atau jawaban pendek. Tes uraian berupa tes yang masing-masing mengandung permasalahan dan menuntut peserta tes mengkonstruk sendiri jawabannya.

Untuk instrumen non tes, dapat dikategorikan menjadi angket, wawancara, observasi, dan dokumentasi. Angket berupa sekumpulan pertanyaan yang biasanya dalam bentuk tertulis kemudian diberikan kepada responden. Jika peneliti menanyakan sekumpulan pertanyaan kepada responden secara langsung, teknik ini disebut dengan wawancara. Observasi terjadi jika peneliti megamati langsung fenomena-fenomena yang terkait dengan penelitian. Adapun dokumentasi merupakan teknik mengumpulkan data dengan menggunakan dokumen-dokumen, baik yang disimpan peneliti sendiri maupun orang lain terkait dengan fokus penelitian.

Pertanyaan-pertanyaan dalam angket atau disebut pula dengan kuisioner bermacam-macam, diantaranya pertanyaan dikotomi, pertanyaan pilihan ganda, urutan bertingkat (*rank ordering*), *rating scale*, dan pertanyaan terbuka (Cohen, Manio, Morrison, 2011). Masing-masing bentuk memiliki ciri khas yang tersendiri, yang disajikan sebagai berikut.

Pertanyaan dikotomi dalam angket hanya memuat 2 pilihan jawaban jawaban saja. Pertanyaan ini digunakan jika peneliti ingin menanyakan kepada responden terkait dengan variabel yang hanya memuat dua jawaban saja. Sebagai contoh jenis kelamin (laki-laki atau perempuan, ya atau tidak, benar atau salah, dan lain-lainnya.

Pertanyaan kuisioner pilihan ganda pada dasarnya seperti pilihan ganda pada soal uraian. Pada pilihan ganda ini, responden biasanya diperkenankan memilih salah satu jawaban saja. Penskoran dapat dilakukan dengan benar-salah saja, atau bertingkat. Jika penskoran dilakukan bertingkat, kondisi ideal yang dihadapi responden dan berbagai kemungkinan kondisi yang dialami responden perlu menjadi pertimbangan penyusun kuisioner.

Untuk mengembangkan instrumen yang baik, ada langkah-langkah yang perlu diperhatikan. Langkah-langkah mengembangkan instrumen baik tes maupun nontes sebagai berikut.

1. Menentukan tujuan penyusunan instrumen

   Pada awal menyusun instrumen, perlu ditetapkan tujuan penyusunan instrumen. Tujuan penyusunan ini memandu teori untuk mengonstruk instrumen, bentuk instrumen,

penyekoran sekaligus pemaknaan hasil penyekoran pada intrumen yang akan dikembangkan. Tujuan penyusunan instrumen ini perlu disesuaikan dengan tujuan penelitian.

Sebagai contoh, ketika peneliti akan mengetahui pengaruh pembelajaran berbasis masalah terhadap motivasi dan kemampuan berfikir tingkat tinggi. Tentunya ada dua intrumen yang perlu dikembangkan, instrumen pengukur motivasi dan instrumen pengukur kemampuan berfikir tingkat tinggi.

2. Mencari teori yang relevan atau cakupan materi

   Setelah tujuan penyusunan instrument ditetapkan, selanjutnya perlu dicari teori atau cakupan materi yang relevan. Teori yang relevan digunakan untuk membuat konstruk, apa saja indicator suatu variabel yang akan diukur. Kaitannya dengan tes, perlu dibatasi juga cakupan materi apa saja yang menjadi bahan menyusun tes. Sebagai contoh pada kemampuan berfikir tingkat tinggi, yang akan diukur harus memiliki indikator pemecahan masalah (*problem solving*), kebaruan, kreativitas, kontekstual dan lain-lain. Jika yang akan diukur adalah siswa SMP, cakupan materi apa saja yang akan diukur perlu menjadi bahan pertimbangan.

3. Menyusun indikator butir instrumen/soal

   Indikator soal ini ditentukan berdasarkan kajian teori yang relevan pada instrumen nontes. Adapun pada instrumen tes, selain mempertimbangkan kajian teori, perlu dipertimbagkan cakupan dan kedalaman materi. Indikator ini telah bersifat khusus, sehingga dengan menggunakan indicator dapat disusun menjadi butir instrumen. Biasanya aspek yang akan diukur dengan indikatornya disusun menjadi suatu tabel. Tabel tersebut kemudian disebut dengan kisi-kisi (*blue print*). Penyusunan kisi-kisi ini mempermudah peneliti menyusun butir soal.

4. Menyusun butir instrumen

   Langkah selanjutnya adalah menyusun butir-butir instrumen. Penyusunan butir ini dilakukan dengan melihat indikator yang sudah disusun pada kisi-kisi. Pada penyusunan butir ini, peneliti perlu mempertimbangkan bentuknya. Misal untuk nontes akan

menggunakan angket, angket jenis yang mana, menggunakan berapa skala, penskorannya dan analisisnya.

Jika peneliti akan menggunakan instrumen berupa tes, perlu dipikirkan apakah akan menggunakan bentuk objektif atau menggunakan bentuk uraian (*construted response)*. Pada penyusunan butir ini, peneliti telah mempertimbangkan penskoran untuk tiap butir, sehingga memudahkan analisis. Jika perlu, pedoman penskoran disusun setelah peneliti menyelesaikan penyusunan butir instrumen.

5. Validasi isi

   Setelah butir-butir soal tersusun, langkah selanjutnya adalah validasi. Validasi ini dilakukan dengan menyampaikan kisi-kisi, butir instrumen, dan lembar diberikan kepada ahli untuk ditelaah secara kuantitatif dan kualitatif. Tugas ahli adalah melihat kesuaian indicator dengan tujuan pengembangan instrumen, kesesuaian indicator dengan cakupan materi atau kesesuaian teori, melihat kesuaian instrumen dengan indicator butir, melihat kebenaran konsep butir soal, melihat kebenaran isi, kebenaran kunci (pada tes), bahasa dan budaya. Proses ini disebut dengan validasi isi dengan mempertimbangkan penilaian ahli (*expert judgement*).

   Jika validasi isi akan dikuantifikasi, peneliti dapat meminta ahli mengisi lembar penilaian validasi. Paling tidak, ada 3 ahli yang dilibatkan untuk proses validasi instrumen penelitian. Berdasarkan isian 3 ahli, selanjutnya penelitian menghitung indeks kesepakatan ahli atau kesepakatan validator dengan menggunakan indeks Aiken atau indeks Gregory.

6. Revisi berdasarkan masukan validator

   Biasanya validator memberikan masukan. Masukan-masukan ini kemudian digunakan peneliti untuk merevisinya. Jika perlu, peneliti perlu mengkonsultasikan lagi hasil perbaikan tersebut, sehingga diperoleh instrumen yang benar-benar valid.

7. Melakukan ujicoba kepada responden yang bersesuaian untuk memeroleh data respons peserta

Setelah revisi, butir-butir instrumen kemudian disusun lengkap (dirakit) dan siap diujicobakan. Ujicoba ini dilakukan dalam rangka memeroleh bukti empiris. Ujicoba ini dilakukan kepada responden yang bersesuaian dengan subjek penelitian. Peneliti dapat pula menggunakan anggota populasi yang tidak menjadi anggota sampel.

8. Melakukan analisis (reliabilitas, tingkat kesulitan, dan daya pembeda)
   Setelah melakukan ujicoba, peneliti memeroleh data respons peserta ujicoba. Dengan menggunakan respons peserta, peneliti kemudian melakukan penskoran tiap butir. Selanjutnya hasil penskoran ini digunakan untuk melakukan analisis reliabilitas skor perangkat tes dan juga analisis karakteristik butir. Analisis karakteristik butir dapat dilakukan dengan pendektatan teori tes klasik maupun teori respons butir. Analisis pada kedua pendekatan ini akan dibahas pada bab-bab selanjutnya.

9. Merakit instrumen
   Setelah karakteristik butir diketahui, peneliti dapat merakit ulang perangkat instrumen. Pemilihan butir-butir dalam merakit perangkat ini mempertimbangkan karakteristik tertentu yang dikehendaki peneliti, misalnya tingkat kesulitan butir. Setelah diberi instruksi pengerjaan, peneliti kemudian dapat mempergunakan instrumen tersebut untuk mengumpulkan data penelitian.

Sebagai contoh, berikut disajikan contoh penyusunan instrumen tes dan nontes. Penyusunan contoh ini hanya sampai menyusun butir saja, karena langkah berikutnya disajikan pada bab-bab selanjutnya dalam buku ini.

Contoh Instrumen Tes (Pengembangan Soal Try-out Ujian Nasional Matematika Sekolah Menengah Atas (SMA)-IPS).

Langkah-langkah pengembangan instrument sebagai berikut.

1. Menentukan Tujuan Penyusunan Instrumen

Tujuan penyusunan instrumen ini adalah mengembangkan soal try-out ujian nasional matematika SMA-IPS, jadi instrument berbentuk tes. Karena tes yang digunakan pada ujian

nasional matematika berbentuk pilihan ganda, maka pada instrumen yang dikembangkan ini berbentuk tes pilihan ganda.

2. Menentukan cakupan materi

Karena yang dikembangkan adalah soal try-out ujian nasional matematika SMA-IPS, maka cakupan materi meliputi standar kompetensi lulusan (SKL) matematika SMA-IPS. Untuk keperluan ini, SKL dapat diambil dari Peraturan Menteri Pendidikan Nasional (Permendiknas) mengenai SKL. SKL untuk SMA-IPS disajikan di laman bsnp-indonesia.org.

3. Menyusun Indikator Soal

Dengan menggunakan cakupan materi dalam SKL, pengembang dapat menyusun indikator. Indikator untuk butir soal, dapat disusun menggunakan kata kerja yang sesuai dengan kedalaman soal yang diinginkan, diantaranya dapat menggunakan kata kerja operasional pada taksonomi Bloom yang telah direvisi, misalnya disajikan pada Tabel 1.1 berikut.

Tabel 1.1. Kata Kerja Operasional pada Indikator

Kemampuan yang Diukur	Kata Kerja yang Biasa Digunakan	
Kemampuan *mengingat*	Mengutip	Menyadari
	Menyebutkan	Menghafal
	Menjelaskan	Meniru
	Menggambar	Mencatat
	Membilang	Mengulang
	Mengidentifikasi	Mereproduksi
	Mendaftar	Meninjau
	Menunjukkan	Memilih
	Memberi label	Menyatakan
	Memberi indeks	Mempelajari
	Memasangkan	Mentabulasi
	Menamai	Memberi kode
	Manandai	Menelusuri
	Membaca	Menulis
Kemampuan *memahami*	Memperkirakan	Menjalin
	Menjelaskan	Membedakan
	Mengkategorikan	Mendiskusikan
	Mencirikan	Menggali

Kemampuan yang Diukur	Kata Kerja yang Biasa Digunakan	
	Merinci	Mencontohkan
	Mengasosiasikan	Menerangkan
	Membandingkan	Mengemukakan
	Menghitung	Mempolakan
	Mengkontraskan	Memperluas
	Mengubah	Menyimpulkan
	Mempertahankan	Meramalkan
	Menguraikan	Merangkum
Kemampuan *menerapkan* pengetahuan (aplikasi)	Menugaskan	Menggali
	Mengurutkan	Mengemukakan
	Menentukan	Mengadaptasi
	Menerapkan	Menyelidiki
	Menyesuaikan	Mengoperasikan
	Mengkalkulasi	Mempersoalkan
	Memodifikasi	Mengkonsepkan
	Mengklasifiksi	Melaksanakan
	Menghitung	Meramalkan
	Membangun	Memproduksi
	Mengurutkan	Memproses
	Membiasakan	Mengaitkan
	Mencegah	Menyusun
	Menggambarkan	Mensimulasikan
	Menggunakan	Memecahkan
	Menilai	Melakukan
	Melatih	Mentabulasi
Kemampuan *menganalisis*	Menganalisis	Menjelajah
	Mengaudit	Membagankan
	Memecahkan	Menyimpulkan
	Menegaskan	Menemukan
	Mendeteksi	Menelaah
	Mendiagnosis	Memaksimalkan
	Menyeleksi	Memerintahkan
	Memerinci	Mengedit
	Menominasikan	Mengaitkan
	Mendiagramkan	Memilih
	Mengkorelasikan	Mengukur
	Merasionalkan	Melatih
	Menguji	Mentransfer
	Mencerahkan	
Kemampuan *mengevaluasi*	Membandingkan	Menafsirkan
	Menyimpulkan	Mempertahankan
	Menilai	Memerinci
	Mengarahkan	Mengukur
	Mengkritik	Merangkum
	Menimbang	Membuktikan

Kemampuan yang Diukur	Kata Kerja yang Biasa Digunakan	
	Memutuskan	Memvalidasi
	Memisahkan	Mengetes
	Memprediksi	Mendukung
	Memperjelas	Memilih
	Menugaskan	Memproyeksikan
Kemampuan *Mencipta*	Mengabstraksi	Mendikte
	Mengatur	Meningkatkan
	Menganimasi	Memperjelas
	Mengumpulkan	Memfasilitasi
	Mengkategorikan	Membentuk
	Mengkode	Merumuskan
	Mengkombinasikan	Menggeneralisasi
	Menyusun	Menggabungkan
	Mengarang	Memadukan
	Membangun	Membatas
	Menanggulangi	Mereparasi
	Menghubungkan	Menampilkan
	Menciptakan	Menyiapkan
	Mengkreasikan	Memproduksi
	Mengoreksi	Merangkum
	Merancang	Merekonstruksi
	Merencanakan	Membuat

Sumber: Panduan Pelaksanaan Penilaian Menggunakan Kurikulum 2013 Direktorat PSMP

4. Menyusun Butir Instrumen

Standar kompetensi dan indicator biasanya disajikan dalam tabel. Tabel tersebut kemudian dilengkapi dengan butir soal yang disusun berdasarkan indicator yang telah dirumuskan. Contoh kompetensi dasar, indikator, dan butir soal disajikan dalam Tabel 1.2.

Dalam menyusun butir, perlu dipertimbangkan bentuk butir yang akan digunakan. Pada kasus ini, karena yang dikembangkan adalah soal tryout ujian nasional, maka soal-soal yang disusun adalah soal-soal yang mirip ujian nasional sehigga berbentuk pilihan ganda. Untuk bentuk ini, soal-soal yang disusun perlu memerhatikan aturan penyusunan soal pilihan ganda, misalnya aturan pernyataan (*stem*) jelas dan tidak membingngkan, ketersediaan kunci, distraktor berfungsi baik, dan lain-lainnya.

Tabel 1.2

KISI-KISI SOAL TRYOUT UN MATEMATIKA SMA IPS

Satuan Pendidikan : SMA
Jumlah Butir : 120 menit

Mata Pelajaran : Matematika-IPS
Jumlah Soal : 40 butir

No. Urut	Standar Kompetensi Lulusan	Indikator	Indikator Soal	Materi	Bahan Kelas	No. Soal
1	Memahami pernyataan dan Ingkarannya, menentukan nilai kebenaran pernyataan majemuk, serta mampu menggunakan prinsip logika matematika dalam pemecahan masalah yang berkaitan dengan penarikan kesimpulan	Menentukan kesimpulan dari beberapa premis	Disajikan dua premis Premis 1 berbentuk implikasi Premis 2 berbentuk implikasi Siswa dapat menentukan kesimpulan yang sah dari kedua premis	Penarikan kesimpulan	X/2	1
		Menentukan ingkaran dari suatu pernyataan	Dapat menentukan ingkaran suatu implikasi	Ingkaran pernyataan	X/2	2
		Menentukan pernyataan yang setara	Disajikan pernyataan berbentuk implikasi, siswa dapat menentukan pernyataan yang setara dengan implikasi	Pernyataan yang setara	X/2	3
2	Memahami konsep yang berkaitan dengan aturan pangkat, akar dan logaritma,fungsi aljabar sederhana, persamaan dan pertidaksamaan kuadrat, sistem persamaan linear, program linear, matriks, barisan dan deret, serta mampu menggunakannya dalam pemecahan masalah.	Menyederhanakan hasil operasi bentuk pangkat, akar dan logaritma	Disajikan penjumlahan dan pengurangan berbentuk logaritma dengan bilangan pokok yang sama, maka siswa dapat menentukan nilainya.	Operasi logaritma	X/1	4
	Dst.					

Sumber: SKL dan Indikator dari bsnp-indonesia.org, indikator soal disusun sendiri.

No	Indikator Soal	Soal	Kunci Jawaban	Keterangan Distraktor
1.	Dapat menentukan nilai kebenaran pernyataan majemuk	Terdapat dua pernyataan *s* dan *t*. Apabila pernyataan s bernilai benar dan pernyataan t bernilai salah, maka pernyataan berikut ini yang bernilai salah adalah.... A. $s \wedge \sim t$ B. $\sim s \vee \sim t$ C. $t \Rightarrow s$ D. $t \Rightarrow \sim s$ E. $\sim t \Rightarrow \sim s$	E	A. Salah konsep B. Salah konsep C. Salah konsep D. Salah konsep E. Kunci Jawaban
2.	Dapat menentukan ingkaran suatu implikasi	Negasi dari pernyataan "Jika pentas musik daerah jadi dilaksanakan maka semua siswa bergembira" adalah…. A. Pentas musik daerah jadi dilaksanakan dan semua siswa bergembira B. Pentas musik daerah tidak jadi dilaksanakan dan ada siswa yang bergembira C. Pentas musik daerah jadi dilaksanakan dan ada siswa yang tidak bergembira D. Pentas musik daerah tidak jadi dilaksanakan atau tidak ada siswa yang bergembira E. Pentas seni jadi dilaksanakan atau ada siswa tidak yang bergembira	C	A. Salah konsep B. Siswa salah konsep C. Kunci Jawaban D. Salah konsep E. Salah konsep
3.	Disajikan tiga premis Premis 1 berbentuk implikasi Premis 2 berbentuk implikasi Premis 3 berbentuk implikasi Siswa dapat menentukan kesimpulan yang sah dari ketiga premis	Diketahui premis-premis berikut ini. Premis 1 : Jika saya tidak rajin belajar, maka saya tidak akan berhasil lulus ujian. Premis 2 : Jika saya tidak berhasil lulus ujian, maka orang tua tidak akan bahagia. Premis 3 : Orang tua saya bahagia. Kesimpulan yang sah dari premis-premis tersebut adalah... . A. Jika saya rajin belajar, maka orang tua saya bahagia B. Jika saya tidak rajin belajar , maka orang tua tidak bahadia C. Saya tidak berhasil lulus ujian dan orang tua saya bahagia D. Saya tidak berhasil lulus ujian E. Saya rajin belajar	E	A. Salah konsep B. Salah konsep C. Salah konsep D. Salah konsep E. Kunci Jawaban
4.	Dst.			

Contoh Instrumen Nontes (Pengembangan Instrumen Self Regulated Learning)

1. Menentukan Tujuan Penyusunan Instrumen

Tujuan penyusunan instrumen ini adalah mengembangkan instrumen *self regulated learning* (SRL).

2. Menentukan teori yang relevan

Untuk keperluan ini, peneliti dapat mencari teori yang relevan dari buku referensi, jurnal-jurnal ataupun penelitian yang telah terdahulu. Contohnya adalah ketika mengembangkan instrumen SRL, dicari teori-teori yang relevan, sebagai berikut (diambilkan dari Heri Retnawati, 2015).

Berbagai pendapat disampaikan ahli terkait dengan *self regulated learning*. Pintrich menyatakan bahwa

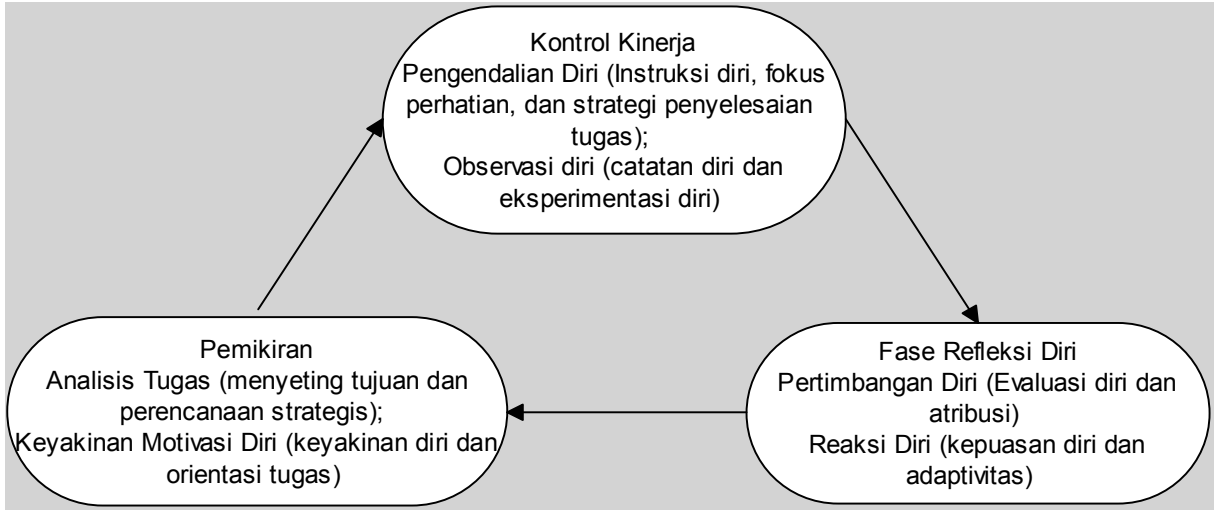
"*Self-regulated learning, or self-regulation, is an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment*" (Schunk, 2005).

Zimmerman menyatakan bahwa "*Self-regulated learning strategies* are actions and processes directed at acquiring information or skill that involve agency, purpose, and instrumenality perceptions by learners" (1989, 1990). Hal ini berarti bahwa seseorang melaksanakan self regulated learning dalam belajar bila yang bersangkutan mengatur perilaku dan kogisinya secara sistematis dengan memperhatikan aturan yang dibuatnya sendiri, mengontrol jalannya pembelajaran yang dilakukannya, mengintegrasikan pengetahuan, melatih untuk mengingat informasi yang diperoleh, serta mengembangkan dan mempertahankan nilai-nilai positif pembelajarannya.

Teori kognitif social dari Bandura (Kivinen, 2013) menyajikan dasar teori dari pengembangan model self regulated learning dalam diri seseorang, dimana faktor-faktor kontekstual dan perilaku berinteraksi dalam suatu ara yang memberikan keuntungan kepada siswa untuk mengatur belajarnya dimana pada waktu yang sama siswa mengatur dirinya sendiri. Suatu perspektif kognitif soasial berbeda dari sudut pandang interaksi personal, perilaku dan linkungannya yang sering disebut proses triadik dari Bandura.

Regulasi diri merupakan proses siklis, karena masukan dari kemampuan awal digunakan untuk membuat keputusan untuk mengulangi usaha-usaha yang telah dilakukan. Upaya pengulangan-pengulangan ini diperlukan karena orang, lingkungan, dan perilaku selaluu berubah selama pembelajaran yang selalu diobservasi dan dipantau.

Pembicaraan *self regulated learning* mencakup 3 fase, meliputi *forethought and planning phase, performance monitoring phase*, dan *reflection on performance phase* (Zumbrunn, S., Tadlock, J., Danielle, E. 2011). Pada fase pemikiran, ada dua hal yang sangat terkait yakni analisis tugas dan keyakinan dan motivasi diri. Fase control kehendak atau kinerja meliputi pengendalian diri dan pengamatan yang khusus. Fase Refleksi diri terdiri dari perkembangan diri, dan reaksi diri. Ketiga fase ini saling terkait dan saling mempengaruhi yang membentuk siklus. Siklus tersebut digambarkan sebagai berikut.



Gambar 2. Fase SRL (Zumbrunn, S., Tadlock, J., Danielle, E. 2011)

Pada fase pemikiran, dapat diklasifikasikan menjadi dua hal yakni analisis tugas (meliputi tujuan pengaturan diri, perencanaan strategis) dan keyakinan motivasi diri (keyakinan diri dan orietasi tugas). Fase kontrol kinerja meliputi pengendalian diri (instruksi diri, fokus perhatian, strategi penyelesaian tugas). Refleksi diri terdiri dari pertimbangan diri (evaluasi diri dan atribusi) dan juga reaksi diri (kepuasan diri dan adaptivitas). Untuk mengetahui skala SRL, Wolkers, Pintrich, Karanenink (2009) mengatakan bahwa perlunya dikembangkan butir terlebih dahulu untuk mengukur

pengaturan kognisi, diikuti dengan regulasi, motivasi, dan perilaku. Ketiga hal ini perlu diukur dalam konteks akademik.

3. Menyusun indikator butir instrumen

   Dari teori-teori yang relevan, dikonstruk indikator-indikator untuk SRL. Untuk memperjelas tiap indicator, peneliti dapat mengembangkan subindikator. Subindikator ini digunakan untuk menyusun butir instrumen. Contohnya sebagai berikut.

Komponen dan Indikator SRL (dikembangkan dari Zimmerman (2000))

Komponen	Indikator	Sub Indikator	No. Butir
Pemikiran	Analisis Tugas	Pengaturan tujuan	1
		Perencanaan Strategis	2
	Keyakinan Diri	Kemampuan diri	3
		Orientasi tugas	4
Kontrol Kinerja	Pengendalian Diri	Instruksi diri	5
		Usaha untuk Fokus belajar	6
		Strategi penyelesaian tugas	7
	Pengamatan yang Cukup	Pemantauan metakognitif	8
		Catatan diri	9
		Eksperimentasi diri	10
Refleksi Diri	Pertimbangan Diri	Evaluasi diri	11
		Atribusi kausal	12
	Reaksi diri	Kepuasan diri (Hadiah)	13
		Kepuasan diri (Hukuman)	14
		Adaptif/defensif	15

Butir-butir instrumen kemudian disusun berdasarkan subindikator tersebut. Contohnya sebagai berikut. Bentuk yang sesuai dengan kasus ini adalah angket dengan skala Likert, dengan 4 pilihan TP (tidak pernah), J (jarang), S (Sering), SL (Selalu).

Butir untuk mengukur SRL dengan Likert

No	Pernyataan	TP	J	S	SL
1	Saya merumuskan tujuan-tujuan kuliah/belajar saya, sebelum kegiatan dimulai				
2	Saya merencanakan strategi untuk mencapai tujuan kuliah/belajar saya				
3	Saya mempercayai kemampuan diri saya untuk berhasil dalam kuliah/belajar				
4	Saya menitikberatkan usaha mencapai tujuan kuliah/belajar saya dibandingkan dengan kegiatan lain.				
5	Saya membuat jadwal untuk diri sendiri terkait dengan pencapaian tujuan kuliah/belajar saya				
6	Saya mengupayakan diri untuk focus belajar				
7	Saya menyusun strategi paling tepat untuk penyelesaian tugas kuliah/belajar				
8	Saya membuat peta kegiatan/aktivitas telah saya lakukan				
9	Saya membuat catatan apa yang telah saya lakukan baik yang berhasil maupun yang belum				
10	Jika ada hal yang membuat saya gagal, saya akan berusaha lagi dengan strategi lain.				
11	Setelah selesai melakukan kegiatan dan melihat hasilya (misal akhir semester) saya melakukan evaluasi.				
12	Saya mencermati penyebab keberhasilan atau kegagalan usaha saya.				
13	Setelah mencapai hal sesuai target kuliah/belajar, saya memberi hadiah untuk diri sendiri.				
14	Saya menghukum diri sendiri jika ada hal dari diri sendiri yang menyebabkan saya gagal mencapai target kuliah/belajar.				
15	Jika suatu strategi kuliah/belajar yang saya gunakan berhasil, saya akan menggunakannya lagi.				

Bab II
VALIDITAS INSTRUMEN

Ada berbagai pendapat mengenai validitas untuk instrumen yang digunakan pengukuran, baik di bidang pendidikan maupun psikologi. Menurut *American Educational Research Association, American Psychological Association, and National Council on Measurement* in Education (AERA, APA, and NCME) dalam *Standards for Educational and Psychological Testing*, validitas merujuk pada derajat dari fakta dan teori yang mendukung interpretasi skor tes, dan merupakan pertimbangan paling penting dalam pengembangan tes (1999). Ahli lain mengemukakan bahwa validitas suatu alat ukur adalah sejauhmana alat ukur itu mampu mengukur apa yang seharusnya diukur (Nunnally, 1978, Allen & Yen, 1979: 97; Kerlinger, 1986; Syaifudin Azwar, 2000: 45).

Sementara itu, Linn & Gronlund (1995) menjelaskan validitas mengacu pada kecukupan dan kelayakan interpretasi yang dibuat dari penilaian, berkenaan dengan penggunaan khusus. Pendapat ini diperkuat oleh Messick (1989) bahwa validitas merupakan kebijakan evaluatif yang terintegrasi tentang sejauhmana fakta empiris dan alasan teoretis mendukung kecukupan dan kesesuaian inferensi dan tindakan berdasarkan skor tes atau skor suatu instrumen. Berdasarkan beberapa pendapat tersebut, dapat disimpulkan bahwa validitas akan menunjukkan dukungan fakta empiris dan alasan teoretis terhadap terhadap interpretasi skor tes atau skor suatu instrumen, dan terkait dengan kecermatan pengukuran.

Validitas itu dapat dikelompokkan menjadi tiga tipe, yaitu: (1) validitas kriteria (*criterion-related*), (2) validitas isi, dan (3) validitas konstruk (Nunnally, 1978, Allen & Yen, 1979, Fernandes, 1984, Woolfolk & McCane, 1984, Kerlinger, 1986, dan Lawrence, 1994). Validitas ini dapat diketahui melalui fakta keberadaaan validitas. Sumber fakta validitas dapat dikelompokkan menjadi isi tes, proses respons,

struktur internal, hubungan dengan variabel lain, dan konsekuensi dari pelaksanaan tes (AERA, APA, and NCME, 1999; Cizek, et al., 2008). Keberadaan validitas dari suatu perangkat tes ini dapat diketahui melalui analisis isi tes dan analisis empiris dari skor tes data respons butir (Lissitz & Samuelsen, 2007).

Validitas berdasarkan kriteria dibedakan menjadi dua, yaitu validitas prediktif dan validitas konkuren. Fernandes (1984) mengatakan validitas berdasarkan kriteria dimaksudkan untuk menjawab pertanyaan sejauh mana tes memprediksi kemampuan peserta di masa mendatang *(predictive validity)* atau mengestimasi kemampuan dengan alat ukur lain dengan tenggang waktu yang hampir bersamaan *(concurrent validity)*. Hal senada juga disampaikan oleh Lawrence (1994) yang mengatakan bahwa tes dikatakan memiliki validitas prediktif bila tes itu mampu memprediksikan kemampuan yang akan datang. Dalam analisis validitas prediktif, performansi yang hendak diprediksikan disebut dengan kriteria. Besar kecilnya harga estimasi validitas prediktif suatu instrumen digambarkan dengan koefisien korelasi antara prediktor dengan kriteria tersebut.

Validitas isi suatu instrumen adalah sejauhmana butir-butir dalam instrumen itu mewakili komponen-komponen dalam keseluruhan kawasan isi objek yang hendak diukur dan sejauh mana butir-butir itu mencerminkan ciri perilaku yang hendak diukur (Nunnally, 1978; Fernandes, 1984). Sementara itu Lawrence (1994) menjelaskan bahwa validitas isi itu keterwakilan pertanyaan terhadap kemampuan khusus yang harus diukur. Berdasarkan hal ini, dapat disimpulkan bahwa validitas isi terkait dengan analisis rasional terhadap domain yang hendak diukur untuk mengetahui keterwakilan instrumen dengan kemampuan yang hendak diukur.

Validitas konstruk adalah validitas yang menunjukkan sejauhmana instrumen mengungkap suatu kemampuan atau konstruk teoretis tertentu yang hendak diukurnya (Nunnally, 1978, Fernandes, 1984). Prosedur validasi konstruk diawali dari suatu identifikasi dan batasan mengenai variabel yang hendak diukur dan dinyatakan dalam bentuk konstruk logis berdasarkan teori mengenai variabel tersebut. Dari teori ini ditarik suatu konskuensi praktis mengenai hasil pengukuran pada

kondisi tertentu, dan konskuensi inilah yang akan diuji. Apabila hasilnya sesuai dengan harapan maka instrumen itu dianggap memiliki validitas konstruk yang baik.

Pada tes prestasi belajar, validitas merupakan syarat yang sangat diperlukan dalam pengembangan tes. Menurut pendapat Sireci yang didukung Lissitz & Samuelsen (2007), validasi tes yang dipergunakan dalam dunia pendidikan sebaiknya melibatkan analisis isi tes dan analisis empiris dari skor tes dan data respons terhadap butir oleh peserta tes. Analisis isi tes terkait dengan validitas isi yang selanjutnya diperlukan juga analisis empiris untuk mengetahui validitas konstruk. Kedua analisis ini dimaksudkan agar tes di dunia pendidikan memenuhi syarat tes yang standar.

A. Membuktikan Validitas Isi

Validitas isi ditentukan menggunakan kesepakatan ahli. Kesepakatan ahli bidang studi atau sering disebut dengan *domain* yang diukur menentukan tingkatan validitas isi (*content related*). Hal ini dikarenakan instrumen pengukuran, misalnya berupa tes atau angket dibuktikan valid jika ahli (*expert*) meyakini bahwa bahwa instrumen tersebut mengukur penguasaan kemampuan yang didefinisikan dalam domain ataupun juga konstruk psikologi yang diukur. Untuk mengetahui kesepakatan ini, dapat digunakan indeks validitas, diantaranya dengan indeks yang diusulkan oleh Aiken (1980; 1985; Kumaidi, 2014). Indeks validitas butir yang diusulkan Aiken ini dirumuskan sebagai berikut:

$$V = \frac{\sum s}{n(c-1)} \quad \text{.....(2.1)}$$

dengan V adalah indeks kesepakatan rater mengenai validitas butir; s skor yang ditetapkan setiap rater dikurangi skor terendah dalam kategori yang dipakai ($s = r - l_o$, dengan r = skor kategori pilihan rater dan l_o skor terendah dalam kategori penyekoran); n banyaknya rater; dan c banyaknya kategori yang dapat dipilih rater.

Berdasarkan pendapat tersebut, indeks Aiken V merupakan indeks kesepakatan rater terhadap kesesuaian butir (atau sesuai tidaknya butir) dengan indikator yang ingin diukur menggunakan butir tersebut. Jika diterapkan untuk

instrument pengukuran, menurut seorang rater maka n dapat diganti dengan m (banyaknya butir dalam satu instrumen). Indeks V ini nilainya berkisar diantara 0-1. Contoh membuktikan validitas isi dari instrumen yang disajikan selengkapnya pada Bab 3. Dari hasil perhitungan indeks V, suatu butir atau perangkat dapat dikategorikan berdasarkan indeknya. Jika indeksnya kurang atau sama dengan 0,4 dikatakan validitasnya kurang, 0,4-0,8 dikatakan validitasnya sedang, dan jika lebih besar dari 0,8 dikatakan sangat valid.

Cara lain membuktikan validitas isi dengan kesepakatan ahli adalah dengan menggunakan indeks kesepakatan ahli yang disarankan oleh Gregory (2007). Indeks ini juga berkisar diantara 0-1. Dengan membuat tabel kontingensi pada dua ahli, dengan kategori pertama tidak relevan dan kurang relevan menjadi kategori relevansi lemah, dan kategori kedua untuk yang cukup relevan dan sangat relevan yang dibuat kategori baru relevansi kuat. Indeks kesepakatan ahli untuk validitas isi merupakan perbandingan banyaknya butir dari kedua ahli dengan kategori relevansi kuat dengan keseluruhan butir. Contoh selengkapnya disajikan pada Bab 3.

Ada hal lain yang perlu diperhatikan terkait dengan validitas isi. Keterwakilan indikator dari domain yang akan diukur benar-benar perlu menjadi perhatian. Beberapa ahli menggolongkan hal ini sebagai validitas logis. Kebenaran konsep yang dinyatakan dalam instrumen merupakan hal yang dapat dijadikan kriteria dan bahan pertimbangan untuk mengisi skor dalam format penilaian. Jika instrumen berbentuk pilihan ganda, maka keberadaan kunci jawaban, keberfungsian distraktor, format penulisan, keterbacaan butir,dan juga berfungsinya gambar atau tabel juga dapat dijadikan pertimbangan. Beberapa ahli mengategorikan ini sebagai validitas kenampakan (*face validity*).

B. Membuktikan Validitas Konstruk

Cara kedua pembuktian validitas interpretasi skor hasil pengukuran adalah dengan membuktikan kebermaknaan skor hasil pengukuran (*meaningfulness)*. Cara ini oleh Popham (1995) disamakan dengan pembuktian *construct related validity*. Proses pembuktiannya dapat dilakukan dengan membuktikan bahwa konstruk instrumen memang ada (*exists*) dan kemudian dibuktikan hasil pengukurannya secara empiris. Pendapat tersebut juga didukung Kumaidi (2014). Pendekatan yang dipilih berupa pembuktian bahwa konstruk yang dihipotesiskan dapat dikonfirmasi keberadaannya. Analisis yang banyak digunakan antara lain dengan analisis faktor eksploratori (*exploratory factor analysis, EFA*) maupun konfirmatori (*confirmatory factor analysis, CFA)*.

Dalam suatu penelitian, biasanya digunakan instrument yang melibatkan butir-butir yang banyak. Untuk memahami data seperti ini, biasanya digunakan analisis faktor. Analisis faktor digunakan untuk mereduksi data, dengan menemukan hubungan antar variabel yang saling bebas (Stapleton, 1997), yang kemudian terkumpul dalam variable yang jumlahnya lebih sedikit untuk mengetahui struktur dimensi laten (Anonim, 2001; Garson,2006) , yang disebut dengan faktor. Faktor ini merupakan variable yang baru, yang disebut juga dengan variable latent, variable konstruk dan memiliki sifat tidak dapat diketahui langsung (unobservable). Analisis faktor dapat dilakukan dengan dua cara, yakni analisis faktor eksploratori (*eksploratory factor analysis*) dan analisis faktor confirmatory (*confirmatory faktor analysis*).

Ide dasar analisis faktor baik eksploratori maupun konfirmatori adalah mereduksi banyaknya variable. Misalkan variabel awalnya adalah x_1, …, x_q, yang selanjutnya akan ditemukan himpunan faktor laten ξ_1, …, ξ_n (dengan $q > n$). Variabel observable tergantung pada kombinasi linear faktor laten ξ_1 yang dinyatakan dengan

$$X_i = \lambda_{i1}\,\xi_1 + \lambda_{i2}\,\xi_2 + \ldots + \lambda_{in}\,\xi_n + \delta_i$$

Dengan δ_i (kesalahan pengukuran) merupakan bagian unik dari x_i yang diasumsikan tidak berkorelasi dengan $\xi_1, \xi_2, ..., \xi_n$. Untuk $i \neq j$, maka $\delta_i \neq \delta_j$. Bagian unik terdiri dari faktor khusus s_i dan suatu kesalahan pengukuran acak e_i.

Analisis faktor eksploratori merupakan suatu teknik untuk mendeteksi dan mengases sumber laten dari variasi atau kovariasi dalam suatu pengukuran (Joreskog & Sorbom, 1993). Analisis faktor eksploratori bersifat mengeksplor data empiris untuk menemukan dan mendeteksi karakteristik dan hubungan antar variable tanpa menentukan model pada data. Pada analisis ini, peneliti tidak memiliki teori *a priori* untuk menyusun hipotesis (Stapleton, 1997). Mengingat sifatnya yang ekplorasi inilah, hasil analisis faktor eksploratori ini lemah. Hasil analisis, yang menjelaskan hubungan antar variable semata, juga tidak didasarkan pada teori yang ada. Hasil analisis juga hanya tergantung data empiris, dan jika variable terobservasinya banyak, hasil analisis akan sulit dimaknai (Stapleton, 1997). Biasanya analisis faktor terkait erat dengan pertanyaan tentang validitas (Nunally, 1978). Ketika faktor-faktor teridentifikasi dihubungkan, analisis faktor eksploratori menjawab pertanyaan tenang validitas konstruk, apakah suatu skor mengukur apa yang seharusnya diukur.

Sebagai contoh data NPV.RAW pada *TUTORIAL LISREL 8.51*. Data ini merupakan data yang dikumpulkan Holzinger dan Swinwford pada tahun 1939 dengan menggunakan 21 tes psikologi yang diikuti 145 siswa di Chicago. Ada 9 jenis tes, yang dianggap sebagai variabel observable, yakni *VISPERC, CUBES, LOZENGES, PARCOM, SENCOM, WORDMEAN, ADDITION, COUNTDOT* dan *SCCAPS*. Secara eksploratori, yang disajikan dalam Tabel 2.1 setelah melalui proses rotasi Promax, 9 variabel tersebut dapat disederhanakan menjadi 3 faktor baru yakni *component* 1 *(VISPERC, CUBES, LOZENGES*), *component* 2 *(PARCOM, SENCOM, WORDMEAN)* dan komponen 3 *(ADDITION, COUNTDOT dan SCCAPS)*(Disarikan dari Heri Retnawati, 2009, evaluation-edu.com).

Pada komponen 1, mungkin akan dapat mudah diintrepretasikan, bahwa variable *VISPERC* (pandangan visual), *CUBES* (kubus), *LOZENGES* (belah ketupat)

terkait dengan konsep geometri. Demikian pula komponen 2, *PARCOM* (parable, perumpamaan/ parafrase), *SENCOM* (sentence-kalimat), *WORDMEAN* (arti kata) dapat dimaknai bahwa faktor ini terkait dengan kemampuan verbal. Namun pada komponen ketiga yang merupakan kumpulan variable *ADDITION* (penjumlahan), *COUNTDOT* (menghitung titik) dan *SCCAPS (straight-curved capital*, huruf lurus-lengkung) akan sulit dimaknai. variable *ADDITION* (penjumlahan) terkait dengan ketepatan, *COUNTDOT* (menghitung titik) terkait dengan ketelitian dan *SCCAPS* masih berbau konsep geometri. Namun hasil ini tetap dapat digunakan untuk membangun model hubungan antar variable, yang dapat digunakan untuk membuat/menyusun hipotesis penelitian yang lain. Hasil selengkapnya disajikan pada Tabel 2.1.

Ada beberapa kritik ang terkait dengan analisis faktor ekploratori. Menurut Mulaik (Stapleton, 1997), untuk memperoleh pengetahuan, yang perlu dilakukan terlebih dahulu adalah membuat asumsi prior. Pada analisis faktor ekploratori, hubungan kausal diasumsikan linear. Kenyataannya, tidak semua variabel bersifat linear. Proses penemuan struktur faktor, semata-mata dilakukan secara mekanik dengan metode tertentu dan dengan rotasi.

Tabel 2.1

Contoh Hasil Analisis Faktor Eksploratori

Component Score Coefficient Matrix

Variabel	Component		
	1	2	3
VISPERC	.028	.059	.349
CUBES	-.046	-.049	.415
LOZENGES	.027	-.022	.400
PARCOM	.365	-.020	.017
SENCOM	.362	.047	-.036
WORDMEAN	.363	-.029	.013
ADDITION	.048	.450	-.142
COUNTDOT	-.060	.441	.057
SCCAPS	.024	.303	.177

Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization.

Ada beberapa kritik yang terkait dengan analisis faktor ekploratori. Menurut Mulaik (Stapleton, 1997), untuk memperoleh pengetahuan, yang perlu dilakukan terlebih dahulu adalah membuat asumsi prior. Pada analisis faktor ekploratori, hubungan kausal diasumsikan linear. Kenyataannya, tidak semua variabel bersifat lnear. Alam penemuan struktur faktor, semata-mata dilakukan secara mekanik dengan metode tertentu dan dengan rotasi.

Berbeda dengan analisis faktor eksploratori, analisis faktor konfirmatori digunakan untuk menguji model yang telah diasumsikan untuk dideskripsikan, dijelaskan untuk model data empiris dengan menggunakan parameter yang lebih sedikit dibandingkan dengan variable terobservasi (Joreskog dan Sorbom, 1993; Steward, dalam Anonim, 2001). Model yang dibangun didasarkan pada informasi a priori tentang struktur data dalam bentuk teori khusus atau hipotesis (Garson, 2006). Teori khusus atau hipotesis yang dibangun didasarkan pada teori yang telah ada atau hasil penelitian sebelumnya.

EFA digunakan ketika model pengukuran dari konstruk instrumen masih dicari ataupun dilakukan eksplorasi. Namun pada CFA, ketika model pengukuran telah ada teorinya, konstruk instumen tersebut tinggal dibuktikan atau dikonfirmasi. Pada CFA, membuktikan validitas konstruk ini khususnya menggunakan model pengukuran (*measurement model*). Menurut Khumaidi, (2014) analisis dapat dilakukan dengan *first order CFA*, dan jika belum konklusif perlu dilakukan *second order analysis*. Analisis dengan EFA dan CFA disajikan pada Bab 4.

C. Membuktikan Validitas Kriteria

Membuktikan validitas kriteria merupakan cara ketiga dalam membuktikan validitas. Validitas ini dibuktikan dengan melihat kebermanfaatan dari interpretasi skor hasil pengukuran (*usefulness*). Pendekatan yang dipakai dapat dalam bentuk *criterion-related validation* (Popham, 1995). Pada pembuktian validitas dengan cara ini, diperlukan skor hasil pengukuran menggunakan instrumen lain yang lebih

terstandar. Misalnya ketika membuktikan validitas tes bahasa Inggris, digunakan tes bahasa Inggris yang lebih terstandar sebagai kriterianya, misalnya TOEFL atau IELTS yang telah diakui di seluruh dunia. Pendekatan analisisnya sering menggunakan yakni analisis dengan korelasi, misalnya korelasi *product-moment*. Jika kriteria yang telah ada saat skor penilaian diperoleh atau rentang waktu perolehan kedua data tidak terlalu lama, maka validasinya bersifat konkuren sehingga sering disebut dengan *concurrent validity*. Jika kriteria keberhasilan ditunggu beberapa lama, misalnya kurun waktu tertentu, maka validasinya bersifat prediktif, sehingga sering disebut dengan *predictive validity*. Pendekatan korelasi ini perlu dikoreksi terlebih dahulu, yang dalam psikometri disebut rumus "*correction for attenuation*" (Allen & Yenn, 1979). Koreksi atenuasi merupakan koreksi terhadap ketidakreliabelan pengukuran konstruk dan kriterianya.

Validitas kriteria diketahui dengan mengestimasi korelasi skor tes peserta dengan skor kriteria. Korelasi ini disebut dengan koefisien validitas (Linn & Gronlund, 1995), yang menyatakan derajat hubungan antara prediktor dengan kriteria. Salah satu manfaat dengan adanya validitas kriteria yakni dapat memprediksikan suatu skor kemampuan ke skor kriteria dalam rangka memprediksikan kemampuan atau performen peserta tes. Prediksi ini dilakukan melalui persamaan regresi.

Ada dua macam regresi yang dapat digunakan. Model yang pertama yakni regresi sederhana atau regresi tunggal, dengan prediktor hanya satu variabel saja (Pedhazur, 1973, Kleinbaum, dkk.,1988; Walpole, dkk., 2002). Model ini dituliskan dengan

$$\hat{Y} = b_0 + b_1 X \quad \text{..... (2.2)}$$

dengan $\hat{Y}$ merupakan hasil prediksi, b_0 konstanta, b_1 koefisien prediktor, dan X merupakan prediktor.

Model yang kedua yakni regresi ganda, dengan prediktor lebih dari satu variabel. Pada kasus kedua ini, digunakan jika tes terdiri dari beberapa subtes, dan prediktor

merupakan jumlahan skor dari subtes-subtes yang berada dalam seperangkat tes. Model regresi ganda dengan dua prediktor disajikan pada persamaan 2.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad \text{.....(2.3)}$$

dengan $\hat{Y}$ merupakan hasil prediksi, b_0 konstanta, b_1 koefisien prediktor pertama, X_1 prediktor pertama, b_2 koefisien prediktor kedua, dan X_2 merupakan prediktor kedua. Kedua model ini belum dibandingkan yang paling akurat, untuk memprediksikan skor kriteria kemampuan peserta tes.

D. Praktek Pembuktian Validitas yang Keliru

Di masyarakat ilmiah, seperti pada laporan penelitian, skripsi, maupun tesis, ada beberapa praktek pembuktian validitas yang belum memenuhi definisi validitas yang telah dipaparkan sebelumnya. Kumaidi (2014) menyatakan bahwa

> ".....banyak praktik pembuktian validitas yang dilakukan oleh mahasiswa atau peneliti yang sebenarnya belum memenuhi definisi validitas. Pendekatan ini tentu perlu dipertanyakan dan sebaiknya ditinggalkan dan dihindari. Pendekatan yang harus dihindari yang dimaksudkan adalah pembuktian validitas yang didasarkan pada analisis butir (*item analysis*), terutama pemakaian koefisien korelasi skor butir dan skor total tes (r_{ix})."

Berdasarkan pernyataan tersebut, membuktikan validita butir dengan menghitung korelasi butir dengan total perlu dihindari. Lebih lanjut Kumaidi (2004) menguraikan ketidaktepatan pemakaian korelasi *product-moment* (r_{ix}) ini sebagai indeks validitas butir sebagai sebuah kekeliruan dan perlu dihindari. Alasan yang dikemukakan r_{ix} hanya merupakan (1) indeks daya beda butir; (2) bagian dari indeks reliabilitas butir; dan (3) homogenitas item (dalam satu set tes atau instrumen). Alasan r_{ix} dipakai sebagai indeks validitas dengan menggunakan *internal criterion* dikarenakan kesulitan menemukan *external criterion* tidak dapat diterima, karena dengan menggunakan *internal criterion* r_{ix} lebih dekat ke analisis reliabilitas dibanding kepada analisis validitas.

Terkait dengan pernyataan-pernyataan tersebut, **kesalahan** pembuktian validitas yang sering terjadi di dunia akademis maupun penelitian pada umumnya adalah membuktikan validitas dengan menghitung korelasi butir dengan total. Demikian pula halnya dengan istilah, yang sering digunakan yakni menguji validitas. Seharusnya, berdasarkan pendapat ahli, terminologi **yang betul adalah membuktikan validitas, bukan menguji validitas**. Adapun caranya, pembuktian validitas yang sesuai dengan definisi validitas, perlu digarisbawahi terbuktinya validitas isi, konstruk, dan kriteria. Dalam suatu penelitian, validitas isi dapat dibuktikan melalui ahli yang menilai relevansi tiap butir instrumen kemudian hasil penilaian ini digunakan untuk menghitung indeks kesepakatan ahli dengan indeks Aiken atau indeks Gregory. Validitas konstruk dapat dibuktikan dengan analisis faktor, baik eksploratori maupun konfirmatori. Validitas kriteria dapat dibuktikan dengan mengetahui besarnya korelasi antara skor responden yang diperoleh dengan instrumen tersebut terhadap skor yang dianggap sebagai kriteria. Prosedur membuktikan validitas isi dibahas secara detail pada Bab 3.

Bab III
MEMBUKTIKAN VALIDITAS ISI

Seperti telah dijelaskan pada Bab 2, membuktikan validitas isi dilakukan melalui kesepakatan ahli (*expert judgement*). *Expert* atau ahli yang dimaksudkan adalah orang yang memiliki kepakaran pada pada bidangnya, tentu saja dengan bidang yang sesuai dengan instrumen untuk penelitian. Langkah-langkah untuk membuktikan validitas isi yaitu:

1. Memberikan kisi-kisi dan butir instrumen, berikut rubrik penskorannya jika ada kepada beberapa ahli yang sesuai dengan bidang yang diteliti untuk mohon masukan. Banyaknya ahli yang dimohon untuk memberi masukan paling tidak 3 orang ahli dengan kepakaran yang relevan dengan bidang yang diteliti.
2. Masukan yang diharapakan dari ahli berupa kesesuaian komponen instrumen dengan indicator, indikator dengan butir, benarnya substansi butir, kejelasan kalimat dalam butir, jika merupakan tes, maka pertanyaan harus ada jawabannya/kuncinya, kalimat-kalimat tidak membingungkan, format tulisan, simbol, dan gambar yang cukup jelas. Proses ini sering disebut telaah kualitatif yang meliputi aspek substansi, bahasa, dan budaya.
3. Berdasarkan masukan ahli tersebut, kisi-kisi dan atau instrumen kemudian diperbaiki.
4. Meminta ahli untuk menilai validitas butir, berupa kesesuaian antara butir dengan indikator. Penilaian ini dapat dilakukan misalnya dengan skala Likert (Skor1: Tidak Valid, Skor 2= kurang valid, Skor 3= cukup valid, skor 4= valid, skor 5 = sangat valid). Dapat pula penskoran dengan melihat relevansi butir dengan indicator (Skor1: Tidak Relevan, Skor 2= kurang relevan, Skor 3= cukup relevan, skor 4= relevan, skor 5 = sangat relevan).

5. Menghitung indeks kesepakatan ahli (*rater agreement*) dengan indeks Aiken V atau indeks Gregory, yang merupakan indeks untuk menunjukkan kesepakatan hasil penilaian para ahli tentang validitas, baik untuk butir maupun untuk perangkatnya.

Pada bab ini disajikan dua contoh kasus, yaitu membuktikan validitas isi instrumen penelitian yang berupa tes dan yang berupa nontes.

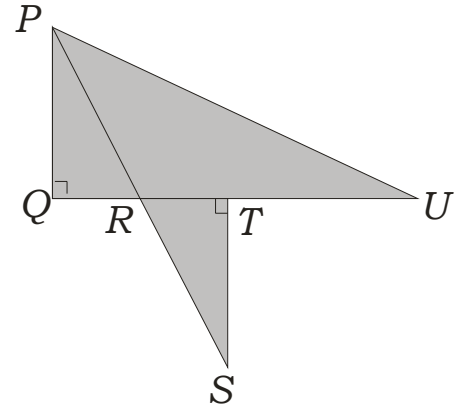
A. Membuktikan Validitas Isi Instrumen Tes

Pada pengembangan instrumen pengukuran, misalnya tes, dikembangkan kisi-kisi dahulu dan butir perangkatnya, minimal memuat indikator, bentuk instrumen/tes, kemudian butir soal, dan penskorannya (pada kasus ini penskoran tidak dituliskan karena perlu tempat yang cukup banyak). Kemudian kisi-kisi berikut (diajikan pada Tabel 3.1) dan butir-butir instrumennya diberikan kepada 3 orang ahli untuk divalidasi, dengan memberikan masukan terkait butir-butir instrumen sekaligus mengisi kesesuaian butir dengan indikator. Contoh format penilaian oleh ahli untuk mengetahui kesesuaian butir dengan indikator disajikan pada Tabel 3.2.

Tabel 3.1. Contoh Instrumen (Tes Prestasi Matematika)

(Instrumen ini diberikan kepada pendidik matematia SMP kelas IX, dalam rangka memotret kesulitan pendidik dalam memecahkan masalah sulit dalam ujian nasional SMP, Heri Retnawati, dkk. 2013)

No.	Indikator	Bentuk instrumen	Butir
1	Menentukan luas permukaan kerucut.	Tes Uraian (Rubrik disajikan di form lain)	Diketahui sebuah kerucut tanpa alas. Jari-jari kerucut tersebut 5 cm dan tingginya 12 cm. Tentukan luas permukaan kerucut tersebut!
2	Menentukan persamaan garis yang tegal lurus degan garis lain kemudian	Tes Uraian (Rubrik disajikan di form lain)	Suatu garis memiliki persamaan 2y=3x +5. Tentukanlah sebuat persamaan garis yang tegak lurus dengan garis tersebut dan

	menggambarkannya dalam bidang koordinat yang sama.		melalui (-2,3). Gambarlah kedua garis tersebut pada satu bidang koordinat.
3	Menentukan panjang sisi pada segitiga-segitiga yang sebangun.	Tes Uraian (Rubrik disajikan di form lain)	Perhatikanlah gambar berikut. Perhatikanlah gambar berikut. Jika $PQ = ST = 12$ cm, $SR = 13$ cm dan $UP = 20$ cm, tentukanlah panjang TU. 
4	Menentukan volum kerucut	Tes Uraian (Rubrik disajikan di form lain)	Sebuah kerucut yang tingginya 24 cm dan panjang garis pelukisnya 25 cm. Tentukanlah volume kerucut tersebut.
5	Menyelesaikan permasalahan yang terkait dengan rerata dari suatu data.	Tes Uraian (Rubrik disajikan di form lain)	Rataan nilai ujian Matematika dari 30 siswa adalah 55. Ada 3 orang yang mengikuti ujian susulan, rerata ketiga orang yang mengikuti ujian susulan tersebut adalah 61. Berapakah nilai rerata 3 orang tersebut?

Setelah masing-masing ahli melakukan penilaian, selanjutnya direkap hasil untuk semua ahli, yang juga disebut dengan rater atau validator dalam satu tabel. Contoh hasil tabulasi dari 3 ahli misalnya disajikan pada Tabel 3.3.

Tabel 3.2. Format penilaian ahli untuk mengetahui kesesuaian butir dengan indikator

No.	Skor Relevansi Butir dengan Indikator					Keterangan
	1	2	3	4	5	
	Tidak Relevan	Kurang Relevan	Cukup Relevan	Relevan	Sangat Relevan	
1						
2						
3						
4						
5						

Tabel 3.3. Contoh hasil tabulasi dari 3 rater

No. Butir	Ahli 1	Ahli 2	Ahli 3
1	5	4	4
2	3	5	5
3	5	3	4
4	4	3	3
5	3	4	5

Dengan menggunakan formula (2.1), selanjutnya dihitung V untuk masing-masing butir. Hasil Selengkapnya disajikan pada Tabel 3.4., yang selanjutnya diperoleh indeks indeks kesepakatan ahli mengenai validitas butir.

Tabel 3.4. Contoh hasil menghitung indeks kesepakatan ahli mengenai validitas

No. Butir	Rater 1	Rater 2	Rater 3	s_1	s_2	s_3	$\sum s$	V
1	5	4	4	4	3	3	10	0,83
2	3	5	5	2	4	4	10	0,83
3	5	3	4	4	2	3	9	0,75
4	4	3	3	3	2	2	7	0,58
5	3	4	5	2	3	4	9	0,75

Selanjutnya hasil tersebut diinterpretasikan, Jika indeks kesepakatan tersebut kurang dari 0,4 maka dikatakan validitasnya rendah, diantara 0,4-0,8 dikatakan validitasnya sedang (*mediocare*) dan jika lebih dari 0,8 dikatakan tinggi,

Cara lain membuktikan validitas isi dengan kesepakatan ahli adalah dengan menggunakan indeks kesepakatan ahli yang disarankan oleh Lawshe dan Martuza (Gregory, 2007).

Pada contoh instrumen tes di atas, sebelum ke ahli dibuat instrumen terlebih dahulu. Instrumen ini untuk mengetahui skor relevansi butir. Selanjutnya ahli memberikan penilaian, apakah butir relevan dengan indicator. Contoh instrumen disajikan pada Tabel 3.5.

Tabel 3.5. Penilaian Relevansi Butir dengan Indikator

No. Butir	Skor Relevansi Butir dengan Indikator			
	1	2	3	4
	Tidak Relevan	Kurang Relevan	Cukup Relevan	Sangat Relevan
1.				
2.				
3.				
4.				
5.				

Setelah ahli menilai kesesuaian butir dengan indikator, dibuat tabulasi dari kedua ahli atau rater tersebut. Contoh hasil penilaian disajikan pada Tabel 3.6.

Tabel 3.6. Contoh hasil tabulasi dari 2 rater

No. Butir	Ahli 1	Ahli 2
1	4	4
2	3	4
3	2	3
4	2	1
5	3	4

Selanjutnya kategori pertama tidak relevan (skor 1) dan kurang relevan (skor 2) dikategorikan ulang menjadi kategori relevansi lemah, dan kategori kedua untuk yang cukup relevan (skor 3) dan sangat relevan (skor 4) yang dibuat kategori baru relevansi kuat. Contoh hasil membuat kategori ulang disajikan pada Tabel 3.7.

Tabel 3.7. Contoh hasil kategori ulang dari dua ahli

No. Butir	Ahli 1	Ahli 2
1	Kuat	Kuat
2	Kuat	Kuat
3	Lemah	Kuat
4	Lemah	Lemah
5	Kuat	Kuat

Dengan membuat tabel kontingensi pada dua ahli, untuk ahli 1 dan ahli 2, pada relevansi lemah dan kuat.

Tabel 3.7. Tabel Kontingensi kategori ulang dari dua ahli

		Rater 1	
		Lemah	Kuat
Rater 2	Lemah	1	0
	Kuat	1	3

Indeks kesepakatan ahli untuk validitas isi merupakan perbandingan banyaknya butir dari kedua ahli dengan kategori relevansi kuat dengan keseluruhan butir. Contoh pertolongan untuk menghitung indeks Gregory disajikan pada Tabel 3.8.

Tabel 3.8. Tabel Kontingensi untuk Menghitung Indeks Gregory

		Rater 1	
		Lemah	Kuat
Rater 2	Lemah	A	B
	Kuat	C	D

Koefisien validitas isi = $\frac{D}{(A+B+C+D)}$ (Gregory, 2007).

Pada contoh kasus tersebut, koefisien validitas isi =3/(1+0+1+3) =0,6.

Selanjutnya hasil tersebut diinterpretasikan, Jika indeks kesepakatan tersebut kurang dari 0,4 maka dikatakan validitasnya rendah, diantara 0,4-0,8 dikatakan validitasnya sedang (*mediocare*) dan jika lebih dari 0,8 dikatakan tinggi. Pada kasus ini karena koefisien validitas isinya 0,6, maka dikatakan validitasnya sedang.

B. Membuktikan Validitas Isi Instrumen Nontes

Pada contoh kedua, akan disajikan membuktikan validitas isi instrumen non tes yang berupa angket untuk mengukur *self-regulated learning* mahasiswa pendidikan Matematika (Heri Retnawati, 2015). Pada awalnya peneliti mendefinisikan dahulu *self-regulated learning* berikut konstruk teorinya berdasarkan pendapat ahli. Selanjutnya disusun kisi-kisi instrumen untuk *self-regulated learning* yang terdiri dari 16 butir. Komponen dan indicator SRL disajikan pada Tabel 3.9.

Tabel 3.9. Komponen dan Indikator SRL

Komponen	Indikator	Sub Indikator	Butir
Pemikiran	Analisis Tugas	Pengaturan tujuan	1
		Perencanaan Strategis	2
	Keyakinan Diri	Kemampuan diri	3
		Kejelasan tujuan yang akan dicapai	4
Kontrol Kinerja	Pengendalian Diri	Instruksi diri, citra diri	5
		Usaha untuk Fokus belajar	7
		Strategi penyelesaian tugas	6,8
	Pengamatan yang Cukup	Pemantauan metakognitif	9
		Catatan diri	10
Refleksi Diri	Pertimbangan Diri	Evaluasi diri	11
		Atribusi kausal	12
	Reaksi diri	Kepuasan diri	13,14
		Adaptif/defensif	15,16

Berdasarkan indikator tersebut, kemudian disusun menjadi butir-butir angket. Hasil penyusunan tersebut disajikan pada Tabel 3.10. Indikator dan butir ini kemudian diberikan kepada 3 orang ahli, 2 ahli psikologi dan 1 orang ahli pengukuran pendidikan MIPA. Ketiga ahli memberikan masukan-masukan mengenai konstruk instrumen dan butir-butir pada angket. Masukan-masukan dari ahli disajikan pada Tabel 3.11.

Tabel 3.10. Butir untuk mengukur SRL dengan Skala Likert

No	Pernyataan	TP	J	S	SL
1	Saya merumuskan tujuan-tujuan kuliah/belajar saya				
2	Saya merencanakan strategi untuk mencapai tujuan kuliah/belajar saya				
3	Saya mempercayai kemampuan diri saya untuk berhasil dalam kuliah/belajar				
4	Saya mengetahui cara mencapai tujuan kuliah/belajar saya dengan jelas				
5	Saya membuat jadwal untuk diri sendiri terkait dengan pencapaian tujuan kuliah/belajar saya				
6	Saya meniru strategi orang yang berhasil dalam kuliah/belajar				
7	Saya mengupayakan diri untuk focus belajar				
8	Saya menyusun strategi yang kira-kira paling tepat untuk penyelesaian tugas kuliah/belajar				
9	Saya membuat peta dari apa yang telah saya lakukan				
10	Saya membuat catatan apa yang telah saya lakukan baik yang berhasil maupun yang belum				
11	Setelah selesai melakukan kegiatan dan melihat hasilya (misal akhir semester) saya melakukan evaluasi,				
12	Saya mencermati penyebab keberhasilan atau kegagalan usaha saya,				
13	Setelah mencapai hal sesuai target kuliah/belajar, saya memberi hadiah untuk diri sendiri,				
14	Saya menghukum diri sendiri jika ada hal yang membuat saya gagal mencapai target kuliah/belajar,				
15	Jika ada hal yang membuat saya gagal, saya akan berusaha lagi dengan strategi lain,				
16	Jika suatu strategi kuliah/belajar yang saya gunakan berhasil, saya akan menggunakannya lagi,				

TP: tidak pernah, J: jarang, S: sering. SL: Selalu

Tabel 3.11 Masukan Perbaikan Butir dari Ahli

Ahli ke-	Masukan	Tindak lanjut
1,2	Citra diri tidak masuk pengendalian diri	Butir 6 tidak dipakai, kisi-kisi diperbaiki
1,3	Perbaikan redaksi butir 1	Perbaikan butir 1
1,2	Perbaikan redaksi butir 4	Perbaikan butir 4
2,3	Perbaikan redaksi butir 14	Perbaikan butir 4

Selanjutnya kisi-kisi dan butir-butir diperbaiki berdasarkan masukan ahli tersebut. Hasilnya disajikan pada Tabel 3.12 untuk kisi-kisi dan Tabel 3.13 untuk butir.

Tabel 3.12. Komponen dan Indikator (Hasil Revisi)

Komponen	Indikator	Sub Indikator	Butir
Pemikiran	Analisis Tugas	Pengaturan tujuan	1
		Perencanaan Strategis	2
	Keyakinan Diri	Kemampuan diri	3
		Orientasi tugas	4
Kontrol Kinerja	Pengendalian Diri	Instruksi diri	5
		Usaha untuk Fokus belajar	6
		Strategi penyelesaian tugas	7
	Pengamatan yang Cukup	Pemantauan metakognitif	8
		Catatan diri	9
		Eksperimentasi diri	10
Refleksi Diri	Pertimbangan Diri	Evaluasi diri	11
		Atribusi kausal	12
	Reaksi diri	Kepuasan diri (Hadiah)	13
		Kepuasan diri (Hukuman)	14
		Adaptif/defensive	15

Setelah itu kepada ketiga ahli dimohon untuk menilai butir-butir dengan mengisi skor (Skor1: Tidak relevan, Skor 2= kurang relevan, Skor 3= cukup relevan, skor 4= relevan, skor 5 = sangat relevan) pada format penilaian pada Tabel 3.14. Hasil dari ketiga ahli sebagai validator kemudian ditabulasikan kemudian disajikan pada tabel 3.15.

Tabel 3.13. Butir untuk mengukur SRL dengan Likert (Hasil Revisi)

No	Pernyataan	STS	TS	S	SS
1	Saya merumuskan tujuan-tujuan kuliah/belajar saya, sebelum kegiatan dimulai				
2	Saya merencanakan strategi untuk mencapai tujuan kuliah/belajar saya				
3	Saya mempercayai kemampuan diri saya untuk berhasil dalam kuliah/belajar				
4	Saya menitikberatkan usaha mencapai tujuan kuliah/belajar saya dibandingkan dengan kegiatan lain,				
5	Saya membuat jadwal untuk diri sendiri terkait dengan pencapaian tujuan kuliah/belajar saya				
6	Saya mengupayakan diri untuk focus belajar				
7	Saya menyusun strategi paling tepat untuk penyelesaian tugas kuliah/belajar				
8	Saya membuat peta kegiatan/aktivitas telah saya lakukan				
9	Saya membuat catatan apa yang telah saya lakukan baik yang berhasil maupun yang belum				
10	Jika ada hal yang membuat saya gagal, saya akan berusaha lagi dengan strategi lain,				
11	Setelah selesai melakukan kegiatan dan melihat hasilya (misal akhir semester) saya melakukan evaluasi,				
12	Saya mencermati penyebab keberhasilan atau kegagalan usaha saya,				
13	Setelah mencapai hal sesuai target kuliah/belajar, saya memberi hadiah untuk diri sendiri,				
14	Saya menghukum diri sendiri jika ada hal dari diri sendiri yang menyebabkan saya gagal mencapai target kuliah/belajar,				
15	Jika suatu strategi kuliah/belajar yang saya gunakan berhasil, saya akan menggunakannya lagi,				

3.14. Lembar Penilaian Validator

No. Butir	Hasil Penilaian ahli dengan memberi tanda cek (√)				
	Tidak Relevan	Kurang Relevan	Cukup Relevan	Relevan	Sangat Relevan
1.					
2.					
3.	dst				
15.					

Tabel 3.15. Hasil Penilaian dari 3 ahli sebagai Validator

Butir	Validator 1	Validator 2	Validator 3
1	5	4	2
2	5	4	5
3	5	5	5
4	5	4	4
5	5	2	5
6	5	4	4
7	5	2	3
8	5	2	4
9	5	2	5
10	5	4	4
11	5	5	4
12	5	4	4
13	5	2	4
14	4	4	3
15	5	4	5

Dengan menggunakan rumus 2.1 dari bab 2, indeks Aiken masing-masing butir dihitung, kemudian hasilnya disajikan pada Tabel 3.16.

Tabel 3.16. Hasil Penghitungan Indeks Aiken

Butir	Rater1	Rater2	Rater3	s1	s2	s3	Σs	V
1	5	4	2	4	3	1	8	0,67
2	5	4	5	4	3	4	11	0,92
3	5	5	5	4	4	4	12	1,00
4	5	4	4	4	3	3	10	0,83
5	5	2	5	4	1	4	9	0,75
6	5	4	4	4	3	3	10	0,83
7	5	2	3	4	1	2	7	0,58
8	5	2	4	4	1	3	8	0,67
9	5	2	5	4	1	4	9	0,75
10	5	4	4	4	3	3	10	0,83
11	5	5	4	4	4	3	11	0,92
12	5	4	4	4	3	3	10	0,83
13	5	2	4	4	1	3	8	0,67
14	4	4	3	3	3	2	8	0,67
15	5	4	5	4	3	4	11	0,92

Untuk keseluruhan skala SRL, koefisiennya dapat dihitung dengan rumus yang sama. Hasil peritungan disajikan pada Tabel 3.17.

Tabel 3.17. Hasil perhitungan koefisien Aiken untuk angket Self Regulated Learning

Skala	Rater1	Rater2	Rater3	s1	s2	s3	Ss	V
Butir 1-15	74	52	61	59	37	46	142	0,79

Mencermati hasil yang disajikan pada Tabel 3.16, diperoleh hasil semua butir berada pada kategori valid atau sangat valid, karena indeks terendah 0,58 dan yang tertinggi 1. Interpretasi ini dilakukan dengan menggunakan kriteria kurang dari 0,4 maka dikatakan validitasnya rendah, diantara 0,4-0,8 dikatakan validitasnya sedang (*mediocare*) dan jika lebih dari 0,8 dikatakan tinggi, Untuk perangkat, berdasarkan Tabel 3.17 diperoleh bahwa indeks Aiken untuk perangkat SRL sebesar 0,79 dengan kategori sedang.

Setelah dibuktikan validitasnya, atau dengan kata lain instrumen telah terbukti valid, langkah yang dapat ditempuh peneliti adalah ujicoba keterbacaan. Setelah merevisinya peneliti dapat melakukan ujicoba instrumen dengan responden yang sesuai dengan tujuan dikembangkannya instrumen tersebut. Ujicoba ini diperlukan untuk memeroleh data empiris, yang kemudian dapat dianalisis lebih lanjut, misalnya membuktikan validitas konstruk instrumen, mengestimasi koefisien reliabilitas, dan mengetahui karakteristik butir.

Cara lain membuktikan validitas isi dengan kesepakatan ahli adalah dengan menggunakan indeks kesepakatan ahli yang disarankan Gregory (2007) yang diperluas (Heri Retnawati, 2015). Untuk keperluan ini, peneliti perlu menyiapkan lembar penilaian validator untuk menilai relevansi butir dengan indikator. Contoh lembar penilaian disajikan pada Tabel 3.18 sebagai berikut.

3.18. Lembar Penilaian Validator

No. Butir	Hasil Penilaian ahli dengan memberi tanda cek (√)			
	Tidak Relevan	Kurang Relevan	Cukup Relevan	Sangat Relevan
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				
11.				
12.				
13.				
14.				
15.				

Kemudian instrumen ini diberikan kepada 3 ahli, kemudian hasilnya ditabulasikan sebagai berikut.

Tabel 3.19. Hasil Penilaian dari 3 ahli sebagai Validator

Butir	Validator 1	Validator 2	Validator 3
1	4	4	2
2	4	4	4
3	4	4	4
4	4	4	4
5	4	2	4
6	4	4	4
7	4	2	3
8	4	2	4
9	4	2	4
10	4	4	4
11	4	4	4
12	4	4	4
13	4	2	4
14	4	4	3
15	4	4	4

Selanjutnya kategori pertama tidak relevan (skor 1) dan kurang relevan (skor 2) dikategorikan ulang menjadi kategori relevansi lemah, dan kategori kedua untuk yang cukup relevan (3) dan sangat relevan (4) yang dibuat kategori baru relevansi kuat.

Tabel 3.20. Kategorisasi ulang hasil peilaian ahli/validator

Butir	Validator 1	Validator 2	Validator 3
1	Kuat	Kuat	Lemah
2	Kuat	Kuat	Kuat
3	Kuat	Kuat	Kuat
4	Kuat	Kuat	Kuat
5	Kuat	Lemah	Kuat
6	Kuat	Kuat	Kuat
7	Kuat	Lemah	Kuat
8	Kuat	Lemah	Kuat
9	Kuat	Lemah	Kuat
10	Kuat	Kuat	Kuat
11	Kuat	Kuat	Kuat
12	Kuat	Kuat	Kuat
13	Kuat	Lemah	Kuat
14	Kuat	Kuat	Kuat
15	Kuat	Kuat	Kuat

Hasil tersebut dapat disajikan pada tabel kontingensi dengan banyaknya sel 2x2x2=8 sel, sebagai berikut.

Tabel 3.21. Hasil kategori ulang untuk menghitung Indeks Gregory

Ahli 1	Lemah	Lemah	Lemah	Lemah	Kuat	Kuat	Kuat	Kuat
Ahli 2	Lemah	Lemah	Kuat	Kuat	Lemah	Lemah	Kuat	Kuat
Ahli 3	Lemah	Kuat	Lemah	Kuat	Lemah	Kuat	Lemah	Kuat
Total	A	B	C	D	E	F	G	H

Pada kasus SRL dalam buku ini, banyaknya butir yang ada pada tiap sel kemudian dihitung. Salah satu cara agar lebih mudah yakni dengan menggunakan turus terlebih dahulu.

Tabel 3.22. Hasil kategori ulang untuk menghitung Indeks Gregory (kasus SRL)

Ahli 1	Lemah	Lemah	Lemah	Lemah	Kuat	Kuat	Kuat	Kuat
Ahli 2	Lemah	Lemah	Kuat	Kuat	Lemah	Lemah	Kuat	Kuat
Ahli 3	Lemah	Kuat	Lemah	Kuat	Lemah	Kuat	Lemah	Kuat
Total	0	0	0	0	0	5	1	9

Koefisien validitas isi dihitung dengan formula:

Koefisien validitas isi = $\frac{H}{(A+B+C+D+E+F+G+H)}$

Pada contoh kasus tersebut, koefisien validitas isi =9/(0+0+0+0+0+5+1+9) =0,6.

Selanjutnya hasil tersebut diinterpretasikan, Jika indeks kesepakatan tersebut kurang dari 0,4 maka dikatakan validitasnya rendah, diantara 0,4-0,8 dikatakan validitasnya sedang (*mediocare*) dan jika lebih dari 0,8 dikatakan tinggi. Pada kasus ini karena koefisien validitas isinya 0,6, maka dikatakan validitasnya sedang.

Setelah dibuktikan validitasnya, untuk perangkat tes ada perlakuan khusus. Agar tes yang disusun merupakan tes yang baik, perlu dilakukan telaah terhadap tes yang telah disusun. Telaah ini dapat dilakukan secara kualitatif. Telaah secara kualitatif dikenal pula sebagai telaah secara teoretis.

Telaah secara teori dapat dilakukan dengan memperhatika tiga aspek tes dalam proses penyusunannya. Ketiga aspek yang tidak terlepas dalam penyusunan butir soal yakni : (1) aspek materi, (2) aspek konstruksi, (3) aspek bahasa/ budaya. (Puspendik, 2006, Panduan Penyusunan Instrumen Tes).

Aspek Materi

Ada 3 hal yang terkait dengan aspek materi penyusunan butir soal, yakni : (1) butir-butir dalam paket tes tersebut telah sesuai dengan indikator pencapaian belajar yang diharapkan, (2) distraktor berfungsi sangat baik (untuk butir pilihan ganda), dan (3) kunci jawaban untuk tiap-tiap butir tes yang ada

hanya satu jawaban. Kesesuaian isi butir-butir soal dalam tes dapat diketahui dengan membandingkan butir-butir tes dengan indicator yang akan dicapai. Berdasarkan hal ini, dapat diketahui apakah butir-butir yang terangkai dalam tes telah representative terhadap materi yang diujikan atau belum.

Aspek Konstruksi

Dalam membuat butir soal, ada sembilan hal yang harus diperhatikan dalam menyusun butir soal berdasarkan aspek konstruksi, diperoleh informasi bahwa (1) pokok soal dirumuskan dengan singkat, jelas, dan tegas, (2) rumusan pokok soal dan pilihan jawaban merupakan pertanyaan yang diperlukan, (3) pokok soal tidak memberi petunjuk ke kunci jawaban, (4) pokok soal bebas dari pertanyaan yang bersifat negatif ganda, (5) gambar, grafik, tabel, diagram, wacana dan sejenisnya yang terdapat dalam soal ditampilkan secara jelas dan berfungsi, (6) panjang pilihan jawaban relatif sama, (7) pilihan jawaban tidak menggunakan pernyataan "semua jawaban di atas salah" atau "semua pilihan jawaban di atas benar" dan sejenisnya, (8) pilihan jawaban yang berbentuk angka atau waktu disusun berdasarkan urutan besar kecilnya angka tersebut atau kronologis, dan (9) butir-butir tidak tergantung pada jawaban butir sebelumnya.

Aspek Bahasa/Budaya

Apabila dilihat dari aspek bahasa/budaya, maka tes itu sebaiknya : (1) menggunakan bahasa yang sesuai dengan kaidah Bahasa Indonesia, (2) menggunakan bahasa yang komunikatif, (3) tidak menggunakan bahasa yang berlaku setempat, dan (4) pilihan jawaban tidak mengulang kata/kelompok kata yang sama.

Bab IV

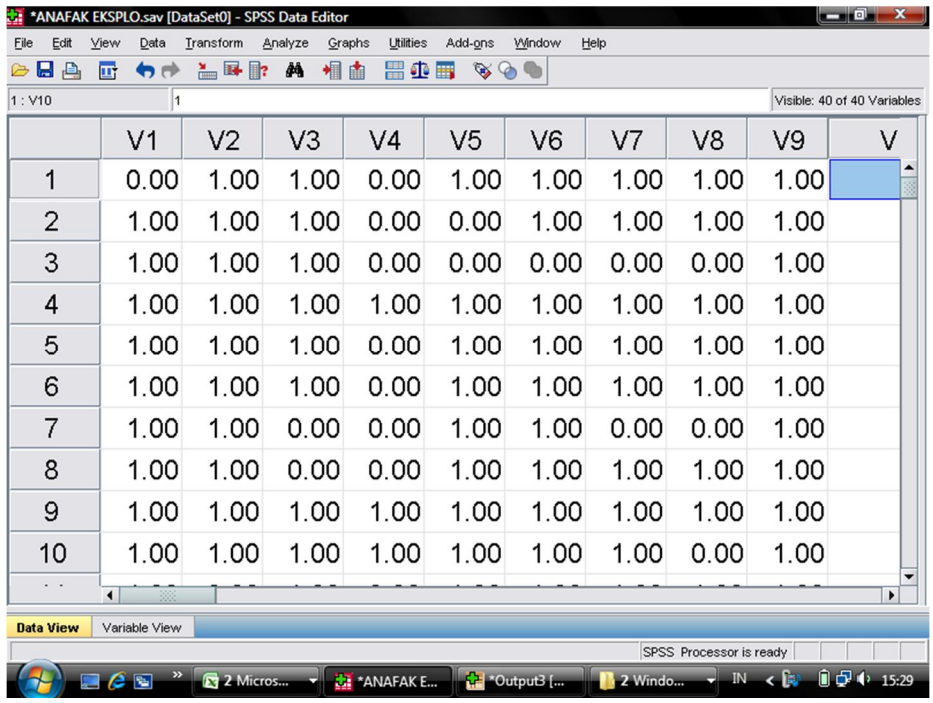
MEMBUKTIKAN VALIDITAS KONSTRUK INSTRUMEN

Membuktikan validitas konstruk instrumen dapat dilakukan dengan dua cara. Kedua cara tersebut yakni analisis faktor eksploratori (AFE) dan analisis faktor konfirmatori (AFK). Pada bab ini akan disajikan contoh melakukan analisis faktor eksploratori dan analisis faktor konfirmatori.

A. Analisis Faktor Eksploratori

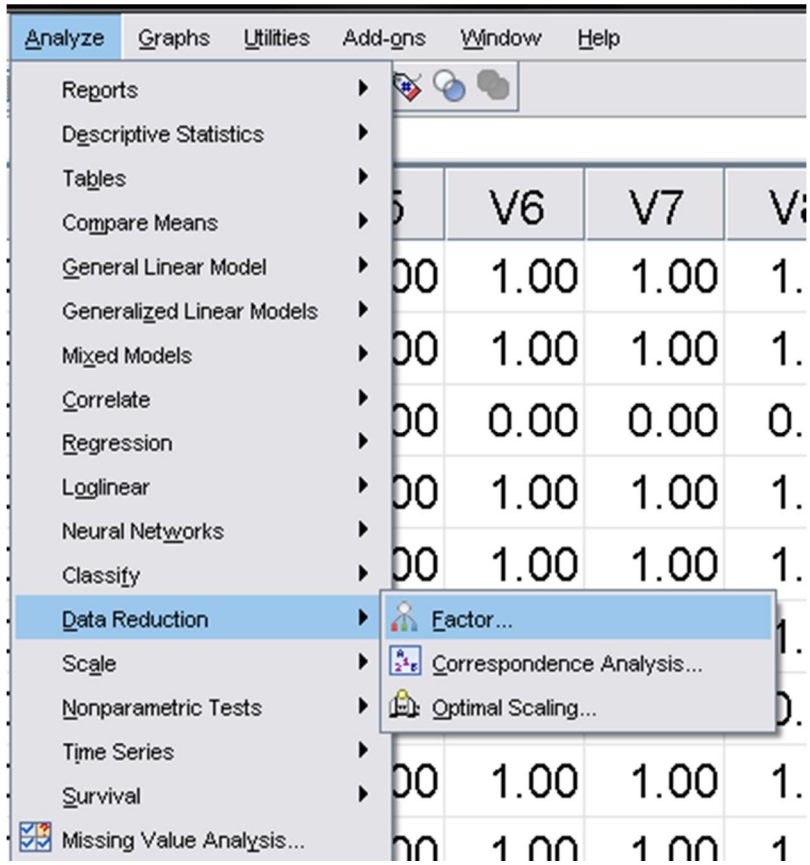
Analisis faktor eksploratori dapat dilakukan secara manual, maupun dengan dengan program komputer. Pada analisis faktor secara manual, jika peneliti sudah mempuyai data ujicobanya, dihitung terlebih dahulu matriks varians kovarians atau matriks korelasi. Kemudian dari matriks ini dihitung nilai eigen, untuk mengetahui persentase varians yang dapat dijelaskan. Namun analisis secara manual ini akan sangat sulit dilakukan jika respondennya banyak dan variabelnyapun juga banyak. Ketelitian perhitungan juga akan mempengaruhi hasil analisis.

Analisis faktor esploratori dengan bantuan computer dapat dilakukan dengan menggunakan berbagai program, seperti SPSS, SAS, MINITAB, R, MPLUS, dan lain-lain. Analisis ini dimulai dengan menguji kecukupan sampel yang digunakan dalam analisis. Selanjutnya computer menyusun matiks varians-kovarians, kemudian mengitung nilai eigen. Nilai eigen ini kemudian diigunakan untuk menghitung persentase varians yang terjelaskan, sekaligus menggambar scree-plotnya. Pada contoh kasus yang disajikan di buku ini dianalisis instrumen Ujian Nasional mata pelajaran matematika SMP 2006, dengan ukuran sampel sebanyak 3.012 siswa dengan panjang tes 40 butir (Heri Retnawati, 2008). Data ini kemudian dianalisis dengan software SPSS. Adapun langkahnya adalah menyiapkan data dalam file SPSS pada **Data View**, dengan variabel merupakan butir-butir instrumen pada **Variabel View**. Contoh hasil penyiapan data sebagai berikut.

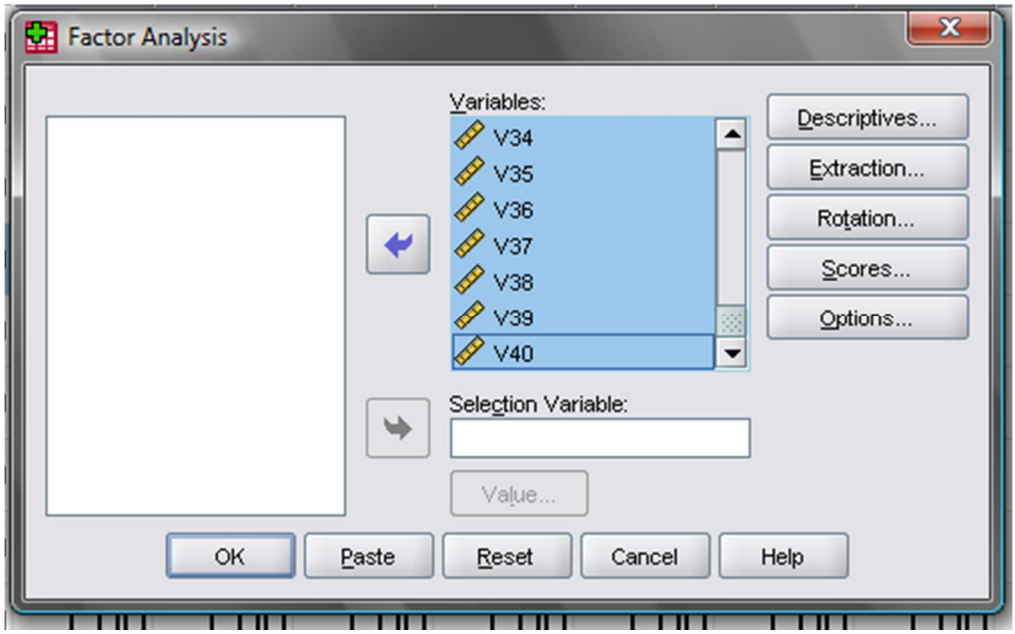


	V1	V2	V3	V4	V5	V6	V7	V8	V9	V
1	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	
2	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	
3	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
5	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	
6	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	
7	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	
8	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	

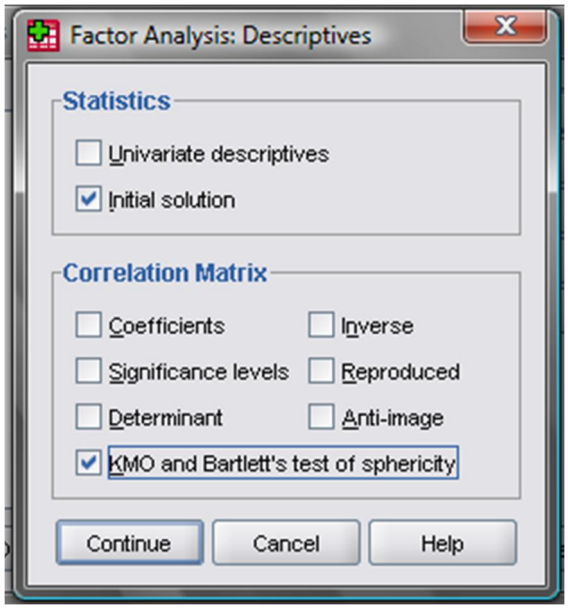
Selanjutnya klik **Analyse→Data Reduction→Factor**.



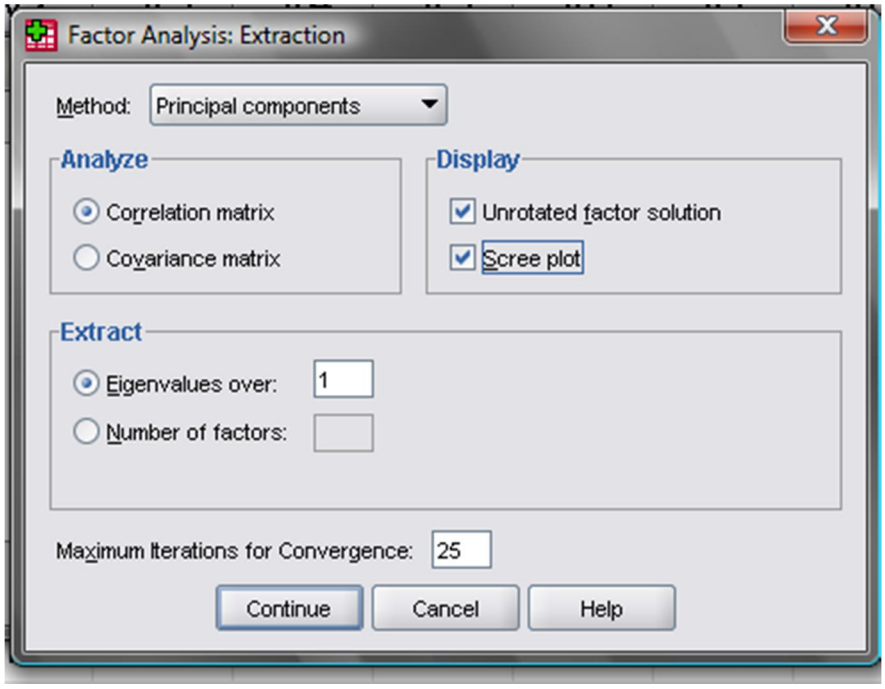
Selanjutnya masukkan variabelnya.



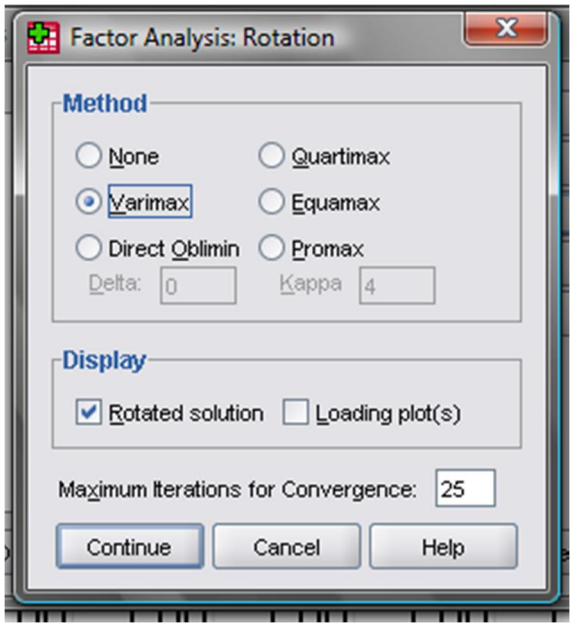
Klik **Descriptive**→klik output yang dikehendaki, misalnya **KMO and Bartlett's testof sphericity**.



Pada **Extraction**, klik **Scree plot**.



Pada **Rotation**, klik **Varimax**.



Selanjutnya akan diproleh hasil pada output, mulai dari KMO, nilai eigen, varians yang dapat dijelaskan, dan komponen faktor. Interpretasinya sebagai berikut.

Hasil analisis faktor tentang kecukupan sampel menunjukkan nilai Khi-kuadrat pada uji Bartlet sebesar 21863,839 dengan derajat kebebasan 435 dan nilai-p kurang dari 0,01. Hasil ini menunjukkan bahwa ukuran sampel sebesar 3.012 yang digunakan pada analisis faktor ini telah cukup, dan juga dikuatkan dengan *Kaiser-Meyer-Olkin measure of sampling adequacy* (KMO) sebesar 0,962 yang lebih besar dari 0,5. Hasil selengkapnya disajikan pada Tabel 4.1.

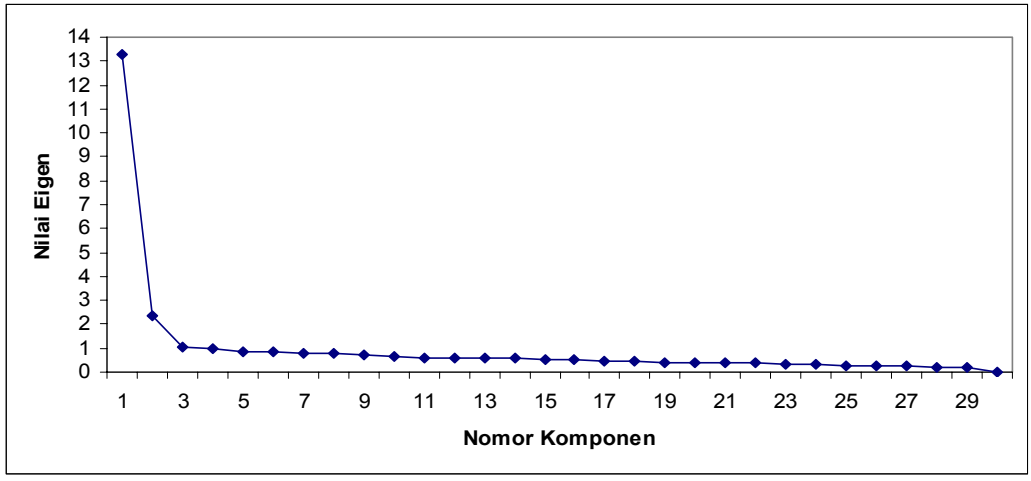
Tabel 4.1 Hasil Uji KMO dan Bartlett

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.962
Bartlett's Test of Sphericity	Approx. Chi-Square	21863.839
	df	435
	Sig.	.000

Berdasarkan nilai eigen dan komponen varians hasil analisis faktor dengan menggunakan SPSS maupun SAS/IML ataupun program lain, dapat diperoleh bahwa data respons siswa terhadap tes UN Matematika SMP memuat 4 nilai Eigen yang lebih besar dari 1, sehingga dapat dikatakan bahwa tes UN Matematika SMP memuat 4 faktor (Heri Retnawati, 2008). Dari keempat faktor ini, ada 59,14% varians yang dapat dijelaskan. Namun dengan memperhatikan scree-plot dari nilai eigen, diperoleh grafik, yang terdiri dari dua curaman saja, sedangkan yang lain menunjukkan grafik yang landai. Hasil ini menunjukkan bahwa bahwa ada 2 faktor dominan yang terukur dalam instrumen ini.

Banyaknya faktor yang termuat dalam instrumen dapat diketahui dari *scree-plot*, contohnya disajikan pada Gambar 4.1. Banyaknya faktor ditandai dengan curamnya grafik perolehan nilai eigen, Gambar tersebut menunjukkan bahwa ada 2 faktor yang terukur pada instrumen ujian nasional matematika tersebut. Dengan 2 faktor ini, instrumen telah dapat menjelaskan 52,19% varians hasil pengukuran. Hasil selengkapnya disajikan pada Tabel 4.2.



Gambar 4.1
Scree Plot Hasil Analisis Faktor Eksploratori

Nilai Eigen selanjutnya dapat disajikan dengan *scree plot* pada Gambar 4.1. Mencermati hasil *scree plot* tersebut, nampak nilai Eigen mulai landai pada faktor ke-3.Ini menunjukkan bahwa terdapat 1 faktor dominan pada perangkat tes matematika, 1 faktor lainnya juga memberikan sumbangan yang cukup besar terhadap komponen varians yang dapat dijelaskan. Mulai faktor ketiga dan seterusnya, pada grafik menunjukkan sudah mulai mendatar. Hal ini menunjukkan bahwa perangkat tes matematika mengukur paling tidak 2 faktor dengan faktor yang pertama merupakan faktor dominan.

Berdasarkan hasil menentukan banyaknya faktor yang termuat tersebut, selanjutnya dilakukan penamaan faktor. Penamaan faktor dilakukan dengan berdasarkan muatan faktor setelah dirotasi, dengan memperhatikan besarnya muatan faktor yang lebih dari 0,4. Penamaan faktor yang termuat dalam perangkat tes dilakukan peneliti dengan bantuan ahli matematika, praktisi (2 orang guru), ahli pendidikan matematika dan psikolog dalam forum *Focus Group Discussion* (*FGD*). Sebelumnya dilakukan analisis dengan 2 faktor menggunakan rotasi promaks. Rotasi ini termasuk pada rotasi nonortogonal. Hal ini dilakukan karena pada model 2 faktor, korelasi faktor pertama dan kedua sebesar 0,3559. Selanjutnya pakar menamai berdasarkan muatan faktor dari tiap butir yang lebih dari 0,4. Muatan faktor yang belum dirotasi disajikan pada Tabel 4.2 dan yang telah dirotasi disajikan pada Tabel 4.3.

Tabel 4.2
Nilai Eigen dan Komponen Varians Hasil Analisis Faktor

No. Komponen	Nilai Eigen	Perbedaan Nilai Eigen	Proporsi	Komulatif
1	13,2860	10,9153	0,4429	0,4429
2	2,3707	1,2986	0,0790	0,5219
3	1,0721	0,0587	0,0357	0,5576
4	1,0134	0,1385	0,0338	0,5914
5	0,8749	0,0512	0,0292	0,6206
6	0,8237	0,0467	0,0275	0,6480
7	0,7770	0,0216	0,0259	0,6739
8	0,7554	0,0400	0,0252	0,6991
9	0,7154	0,0422	0,0238	0,7230
10	0,6732	0,0607	0,0224	0,7454
11	0,6126	0,0116	0,0204	0,7658
12	0,6010	0,0191	0,0200	0,7859
13	0,5819	0,0232	0,0194	0,8052
14	0,5587	0,0285	0,0186	0,8239
15	0,5303	0,0269	0,0177	0,8415
16	0,5034	0,0220	0,0168	0,8583
17	0,4814	0,0430	0,0160	0,8744
18	0,4384	0,0261	0,0146	0,8890
19	0,4124	0,0206	0,0137	0,9027
20	0,3917	0,0233	0,0131	0,9158
21	0,3685	0,0067	0,0123	0,9281
22	0,3618	0,0219	0,0121	0,9401
23	0,3399	0,0141	0,0113	0,9515
24	0,3258	0,0443	0,0109	0,9623
25	0,2815	0,0253	0,0094	0,9717
26	0,2561	0,0032	0,0085	0,9802
27	0,2529	0,0580	0,0084	0,9887
28	0,1949	0,0174	0,0065	0,9952
29	0,1774	0,2098	0,0059	1,0011
30	-0,0323		-0,0011	1,0000

Tabel 4.3
Muatan Faktor pada Model 2 Faktor Sebelum Dirotasi

No. Butir	Nama Butir	Kemampuan Umum	Kemampuan Spasial
1	Persentase (soal cerita)	0,84034	-0,24293
2	Diagram Venn	0,76538	-0,20374
3	Persentase	0,56454	0,06392
4	HP Bilangan Bulat	0,49375	-0,02924
5	Jaring-jaring Kubus	0,71063	-0,3254
6	Simetri Lipat	0,53969	-0,25362
7	Sudut Segitiga	0,4322	0,44252
8	Pemetaan	0,7653	-0,12332
9	Akar dan Pangkat	0,79163	-0,26301
10	Sifat Garis Sejajar	0,59717	0,04047
11	Keliling Belah Ketupat	0,82338	0,02225
12	Luas Jajar Genjang	0,7618	0,07998
13	Perbandingan (Soal Cerita)	0,80968	-0,27777
14	Persamaan Garis Lurus	0,6086	0,11901
15	SPL (Soal Cerita)	0,81121	-0,16936
16	Median Data	0,66918	0,0411
17	Volume limas	0,69272	0,08342
18	Luas permukaan prisma	0,73462	0,07833
19	Refleksi	0,68367	0,18329
20	Dilatasi	0,64019	0,13857
21	Perbandingan Segitiga	0,1559	0,83857
22	Segitiga kongruen	0,61977	-0,19426
23	Juring Lingkaran	0,4935	0,353
24	Persekutuan Lingkaran	0,71413	0,11688
25	Suku dan Faktor	0,69395	0,03537
26	Fungsi kuadrat	0,69903	0,17732
27	Phytagoras dan luas segitiga	0,68345	0,15845
28	Barisan dan deret	0,86509	-0,29498
29	Trigonometri	0,30703	0,55874
30	Logaritma	0,38815	0,53147

Tabel 4.4
Muatan Faktor (Model 2 Faktor) Setelah Dirotasi (*Promax*)

No. Butir	Nama Butir	Kemampuan Umum	Kemampuan Spasial
1	Persentase (soal cerita)	0,8972	-0,06997
2	Diagram Venn	0,80705	-0,04538
3	Persentase	0,4715	0,19086
4	HP Bilangan Bulat	0,46158	0,07775
5	Jaring-jaring Kubus	0,82803	-0,1848
6	Simetri Lipat	0,63261	-0,14715
7	Sudut Segitiga	0,13355	0,55832
8	Pemetaan	0,7605	0,03883
9	Akar dan Pangkat	0,86494	-0,10169
10	Sifat Garis Sejajar	0,51444	0,17347
11	Keliling Belah Ketupat	0,7287	0,20403
12	Luas Jajar Genjang	0,63988	0,25097
13	Perbandingan (Soal Cerita)	0,88972	-0,11319
14	Persamaan Garis Lurus	0,47935	0,25822
15	SPL (Soal Cerita)	0,82845	0,00069
16	Median Data	0,57893	0,18993
17	Volume limas	0,57568	0,23942
18	Luas permukaan prisma	0,61636	0,24328
19	Refleksi	0,50982	0,34202
20	Dilatasi	0,4965	0,28564
21	Perbandingan Segitiga	-0,34413	0,91245
22	Segitiga kongruen	0,67043	-0,0674
23	Juring Lingkaran	0,24048	0,47802
24	Persekutuan Lingkaran	0,57562	0,27916
25	Suku dan Faktor	0,60455	0,18936
26	Fungsi kuadrat	0,52711	0,33914
27	Phytagoras dan luas segitiga	0,52398	0,31596
28	Barisan dan deret	0,94956	-0,11905
29	Trigonometri	-0,04633	0,65256
30	Logaritma	0,04248	0,64181

Langkah selanjutnya adalah penamaan faktor. Faktor ini dinamai berdasarkan butir-butir yang tercakup dalam faktor-faktor tersebut. Bila yang tercakup banyak hal, bisa dinamai dengan butir-butir yang paling dominan. Nama-nama faktor ini

didasarkan pada muatan faktor hasil analisis faktor setelah melakukan rotasi, baik ortogonal maupun nonortogonal.

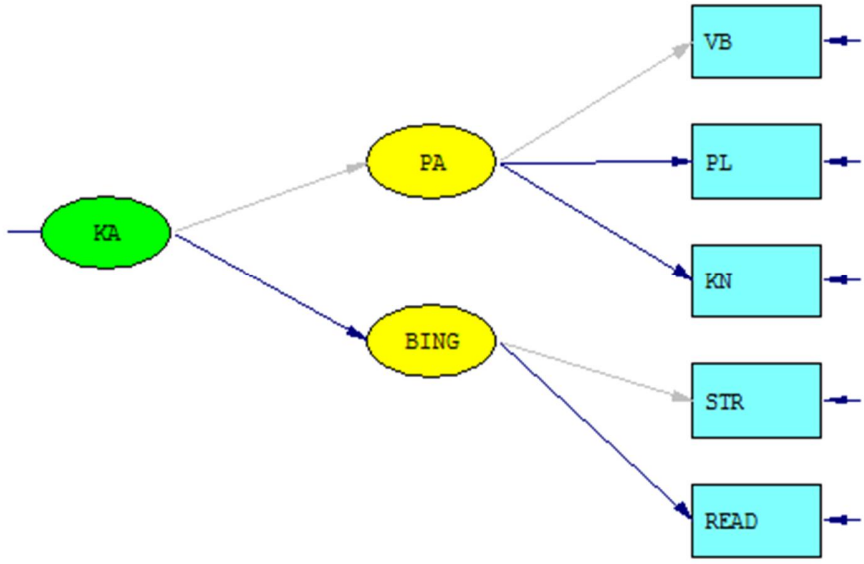
Faktor pertama pada model 2 faktor dinamai dengan kemampuan matematika umum karena muatan faktor hasil rotasi pada faktor pertama meliputi keseluruhan kompetensi dasar minimum yang harus dicapai peserta tes. Faktor kedua dinamai kemampuan spasial karena 4 dari 5 butir soal yang mempunyai muatan faktor lebih dari 0,4 merupakan butir yang terkait dengan kemampuan spasial yakni sudut segitiga, perbandingan segitiga, juring lingkaran, trigonometri dan yang lainnya butir tentang logaritma.

Berdasarkan analisis faktor eksploratori tersebut, dapat disimpulkan bahwa instrumen yang berupa perangkat ujian tersebut valid untuk mengukur kemampuan matematika pada umumnya dan kemampuan spasial dan terbukti secara empiris.

B. Analisis Faktor Konfirmatori

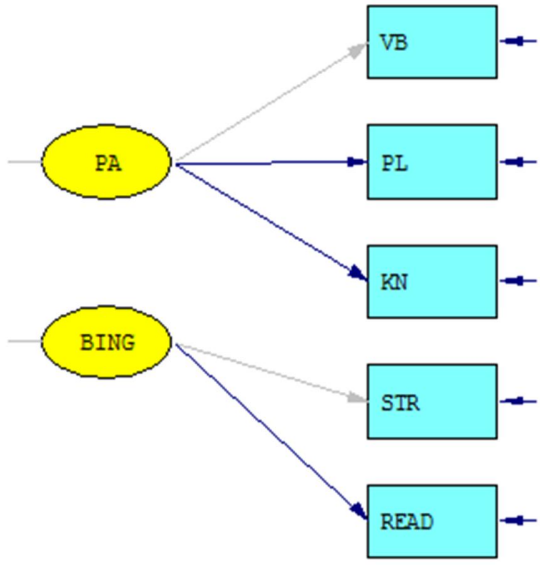
Analisis faktor ini dapat dilakukan dengan bantuan software, misalnya Lisrel, AMOS, MPLUS, PLS, maupun aplikasi lain. Analisis ini memerlukan input berupa variabel-variabel yang akan dibuktikan menjadi indikator suatu instrumen penelitian. Sebagai contoh, seorang peneliti akan membuktikan validitas konstruk instrumen untuk mengukur kemampuan akademik (KA) calon mahasiswa. Berdasarkan teori, kemampuan akademik mahasiswa ini ditentukan oleh potensi akademik (PA) dan kemampuan berbahasa, khususnya bahasa asing misalnya bahasa Inggris (BING). Potensi akademik diukur dengan butir-butir yang mengukur kemampuan verbal (VB), kemampuan penalaran (PN) dan kemampuan numerik (KN). Kemampuan bahasa Inggris yang diukur pada penelitian misalnya hanya structure (STR) dan kemampuan membaca (READ). Konstruk kemampuan akademis tersebut disajikan pada Gambar 4.2. Pada kasus ini, kemampuan verbal, penalaran, kemampuan numerik, struktur dan kemampuan membaca dapat langsung terukur dari tes. Variabel-variabel ini disebut dengan variabel observabel (dapat diamati), yang disimbolkan dengan kotak.

Potensi akademik, kemampuan bahasa Inggris, dan kemampuan akademik untuk melanjutkan studi merupakan variabel yang tidak terukur secara langsung atau disebut dengan variabel laten, yang disimbolkan dengan oval.



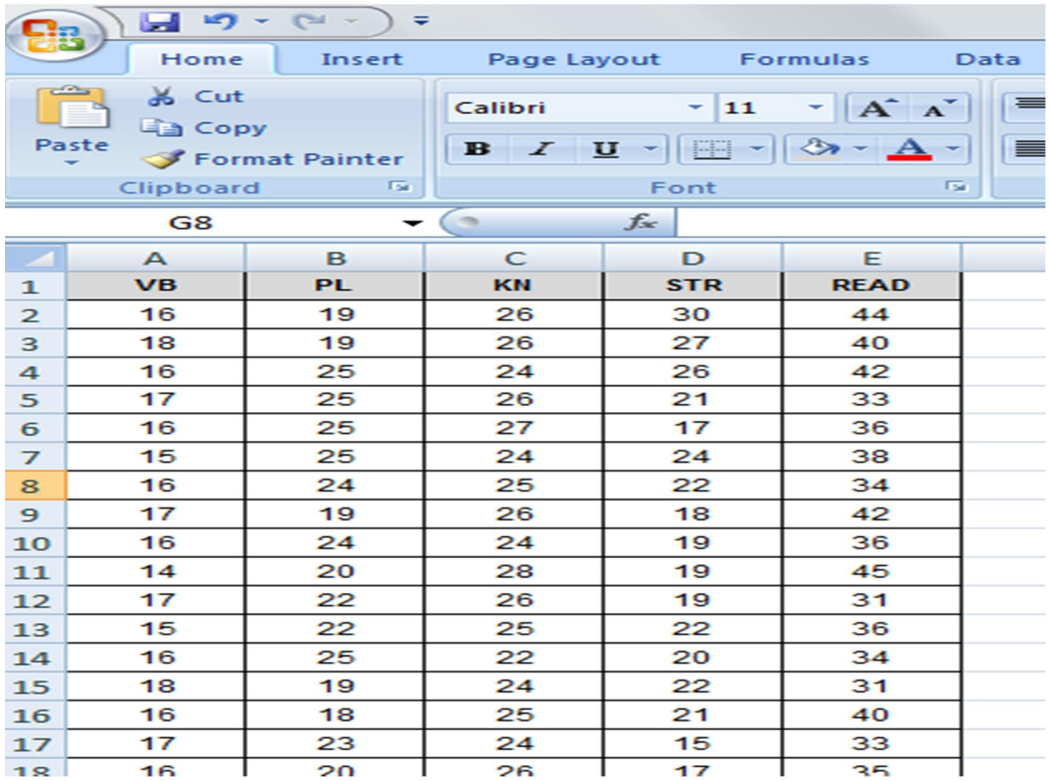
Gambar 4.2. Konstruk Teori Kemampuan Akademis Calon Mahasiswa
(*second order confirmatory factor analysis*)

Model pada Gambar 4.2 merupakan model analisis faktor dengan variabel laten dua tingkat, yang disebut dengan *second order confirmatory factor analysis*. Jika peneliti hanya memerlukan pembuktian dengan variabel laten satu tingkat saja, maka analilis yang diperlukan adalah analisis faktor orde pertama (*first order confirmatory factor analysis*), misalnya hanya sampai pada PA dan BING saja, seperti disajikan pada Gambar 4.3.



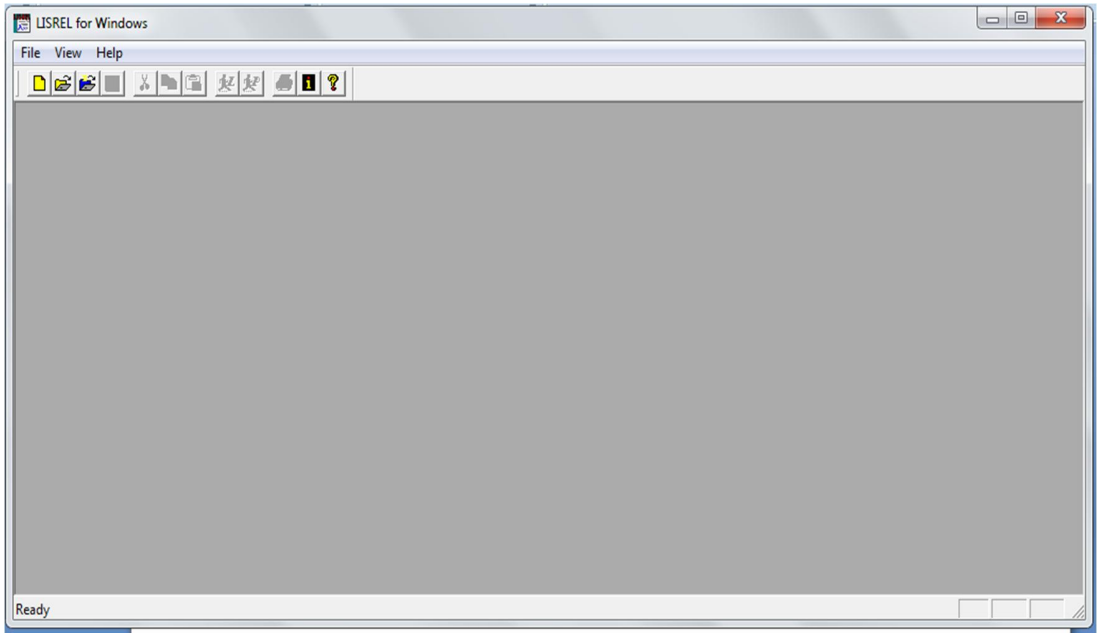
Gambar 4.3. Konstruk Teori Kemampuan Akademis Calon Mahasiswa
(*First order confirmatory factor analysis*)

Dengan menggunakan data empiris, kemudian dilakukan analisis faktor konfirmatori, diantaranya dengan menggunakan software Lisrel sampai memeroleh model yang fit. Tentunya untuk kasus ini, dikumpulkan data ujicoba terlebih dahulu. Data yang diperoleh dapat dientri dalam format teks, format SPSS, maupun format Excel. Contoh kasus yang disajikan di sini disjikan dalam format Excel, dan analisis CFA dilakukan dengan software Lisrel.

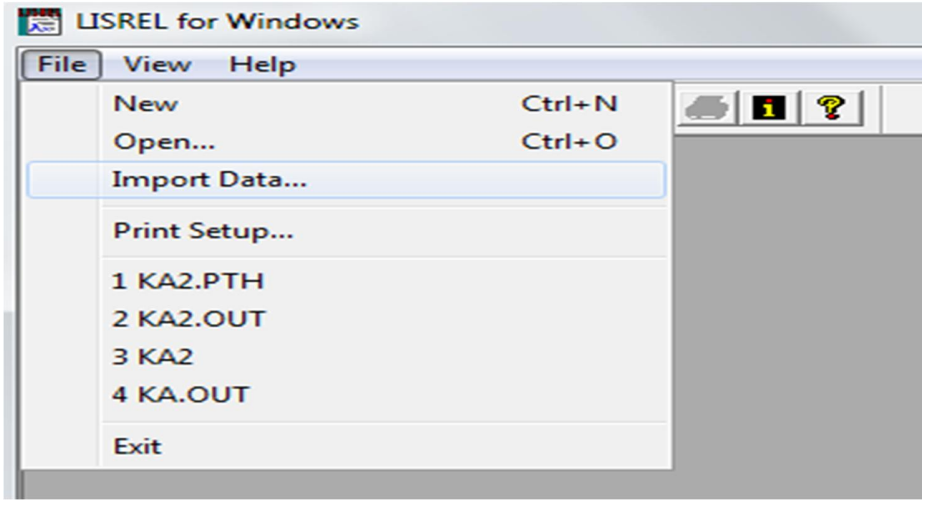


	A	B	C	D	E
1	VB	PL	KN	STR	READ
2	16	19	26	30	44
3	18	19	26	27	40
4	16	25	24	26	42
5	17	25	26	21	33
6	16	25	27	17	36
7	15	25	24	24	38
8	16	24	25	22	34
9	17	19	26	18	42
10	16	24	24	19	36
11	14	20	28	19	45
12	17	22	26	19	31
13	15	22	25	22	36
14	16	25	22	20	34
15	18	19	24	22	31
16	16	18	25	21	40
17	17	23	24	15	33
18	16	20	26	17	35

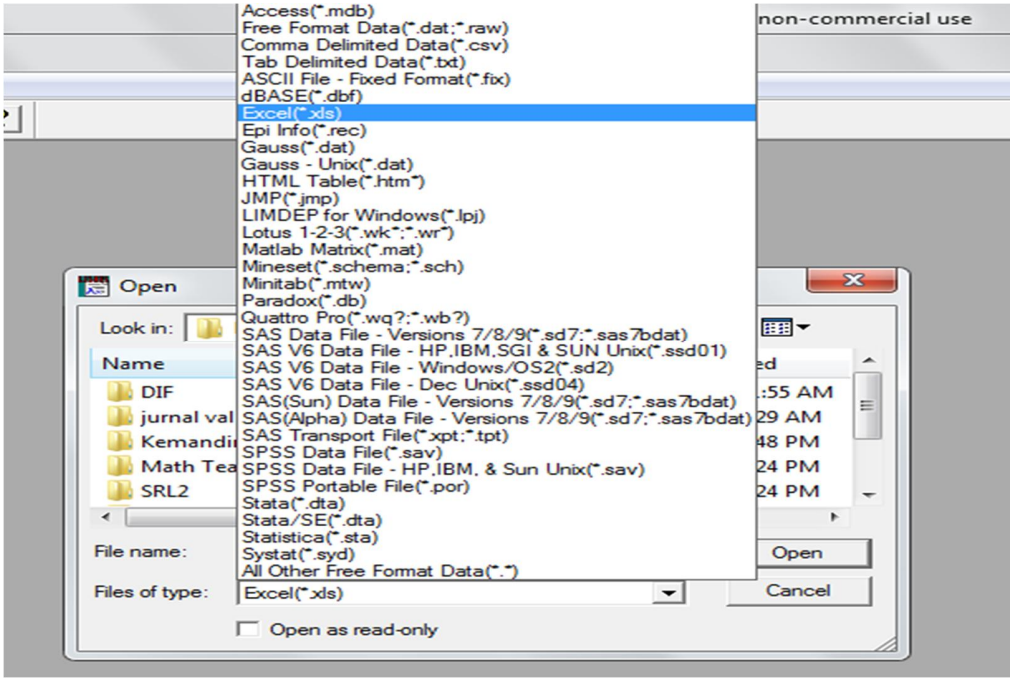
Selanjutkan untuk mengalisis dengan Lisrel, diaktifkan dulu programnya dengan mengklik ganda program Lisrel sehingga diperoleh tampilan berikut.



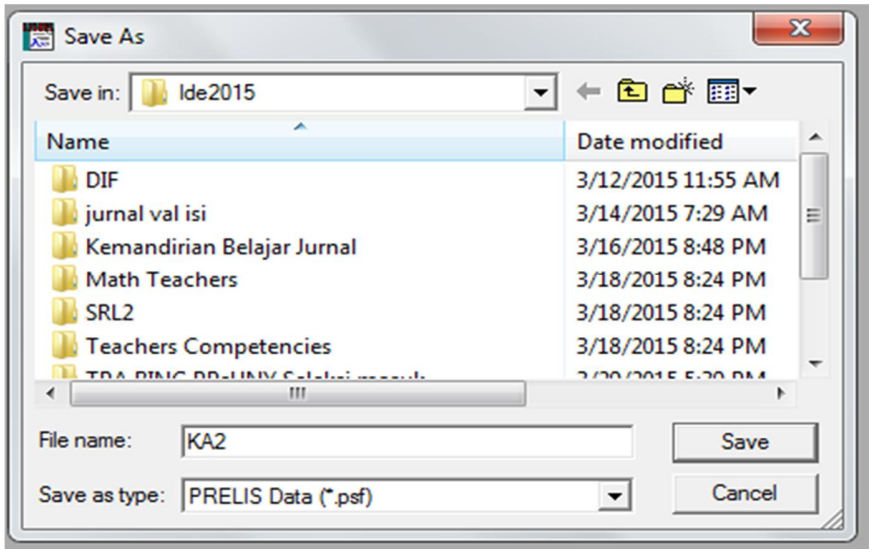
Untuk mulai menganalisis, klik menu **File→Import Data**.



Kemudian diklik merujuk pada tempat data disimpan, dengan memperhatikan tipe dari data yang dimiliki. Jika data tersimpan dalam format Excel dipilih *.xls, jika data tersimpan dalam format SPSS dipilih *.sav, dan lain-lain.



Selanjutnya data kita akan dibaca program Lisrel sebagai prelis, sehingga perlu disimpan dahulu dengan tipe/ekstensi *.psf.



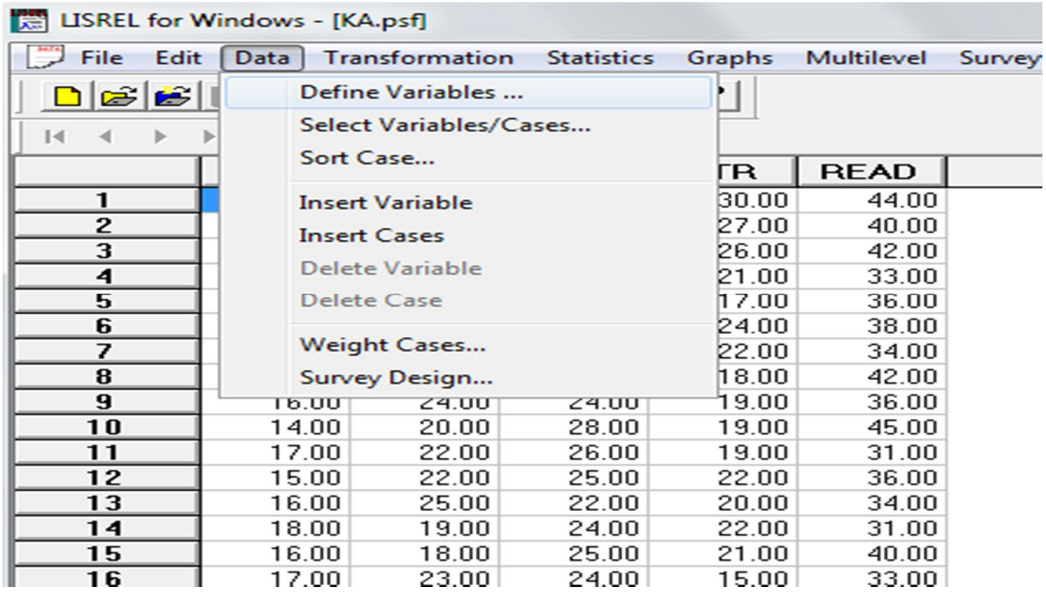
Selanjutnya data akan ditampilkan kembali dalam tampilan program Lisrel .

LISREL for Windows - [KA.psf]

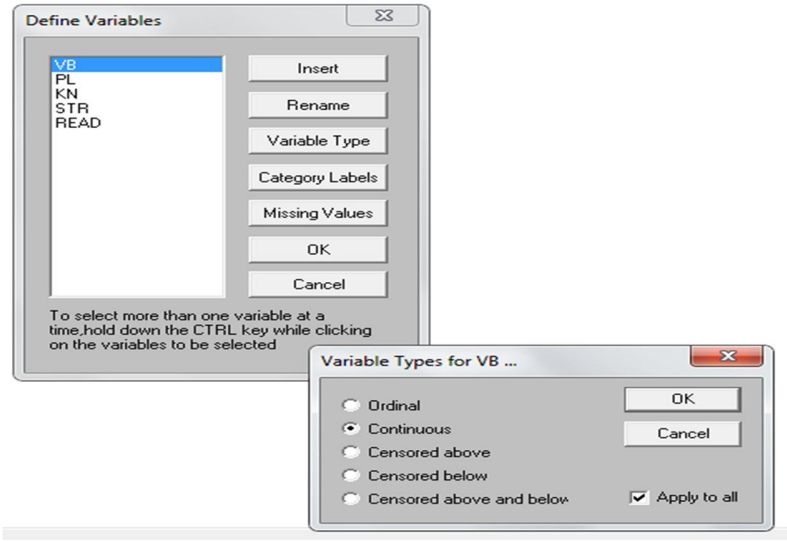
File Edit Data Transformation Statistics Graphs Multilevel

	VB	PL	KN	STR	READ
1	16.00	19.00	26.00	30.00	44.00
2	18.00	19.00	26.00	27.00	40.00
3	16.00	25.00	24.00	26.00	42.00
4	17.00	25.00	26.00	21.00	33.00
5	16.00	25.00	27.00	17.00	36.00
6	15.00	25.00	24.00	24.00	38.00
7	16.00	24.00	25.00	22.00	34.00
8	17.00	19.00	26.00	18.00	42.00
9	16.00	24.00	24.00	19.00	36.00
10	14.00	20.00	28.00	19.00	45.00
11	17.00	22.00	26.00	19.00	31.00
12	15.00	22.00	25.00	22.00	36.00
13	16.00	25.00	22.00	20.00	34.00
14	18.00	19.00	24.00	22.00	31.00
15	16.00	18.00	25.00	21.00	40.00
16	17.00	23.00	24.00	15.00	33.00
17	16.00	20.00	26.00	17.00	35.00
18	19.00	18.00	24.00	20.00	26.00
19	19.00	16.00	29.00	15.00	22.00
20	16.00	15.00	27.00	19.00	36.00
21	15.00	19.00	27.00	20.00	32.00
22	15.00	24.00	25.00	19.00	24.00
23	19.00	16.00	25.00	17.00	25.00

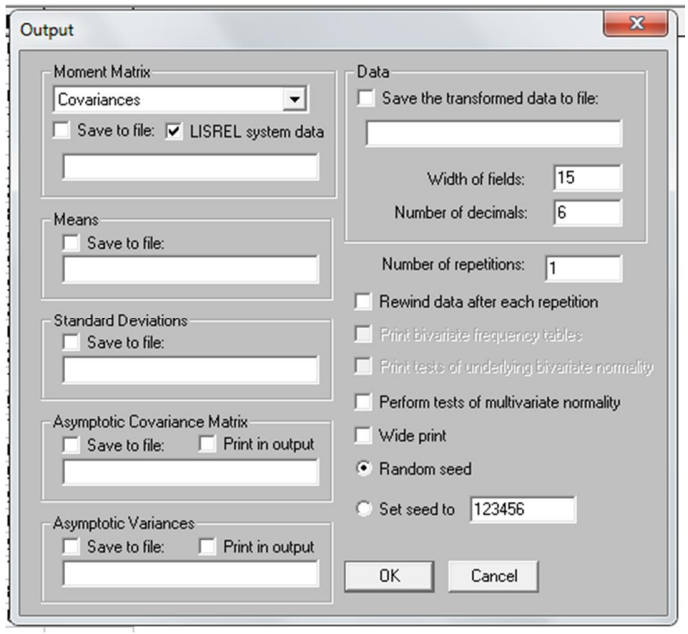
Langkah selanjutnya adalah mengubah tipe variabel, jika variabelnya bukan ordinal. Langkah ini ditempuh dengan mngeklik menu **Data→Define Variable**.



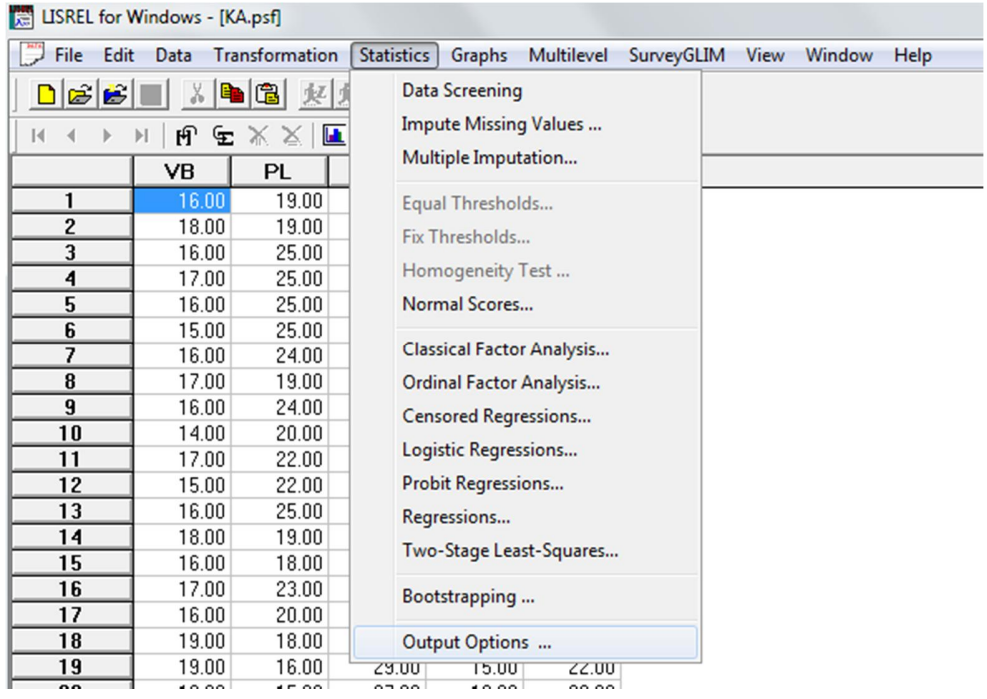
Pada **Define Variabel**, klik salah satu variabel kemudian ubahlah ke tipe yang sesuai dengan data yang dianalisis. Jika akan diberlakukan untuk semua variabel, klik **Apply to All**.



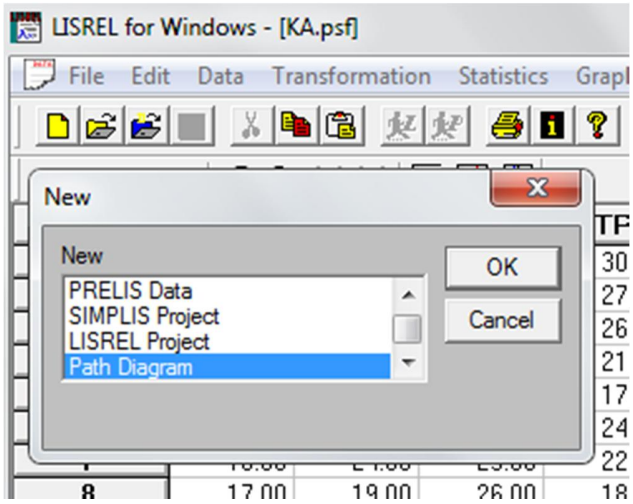
Setelah itu klik **OK→OK→Save**.



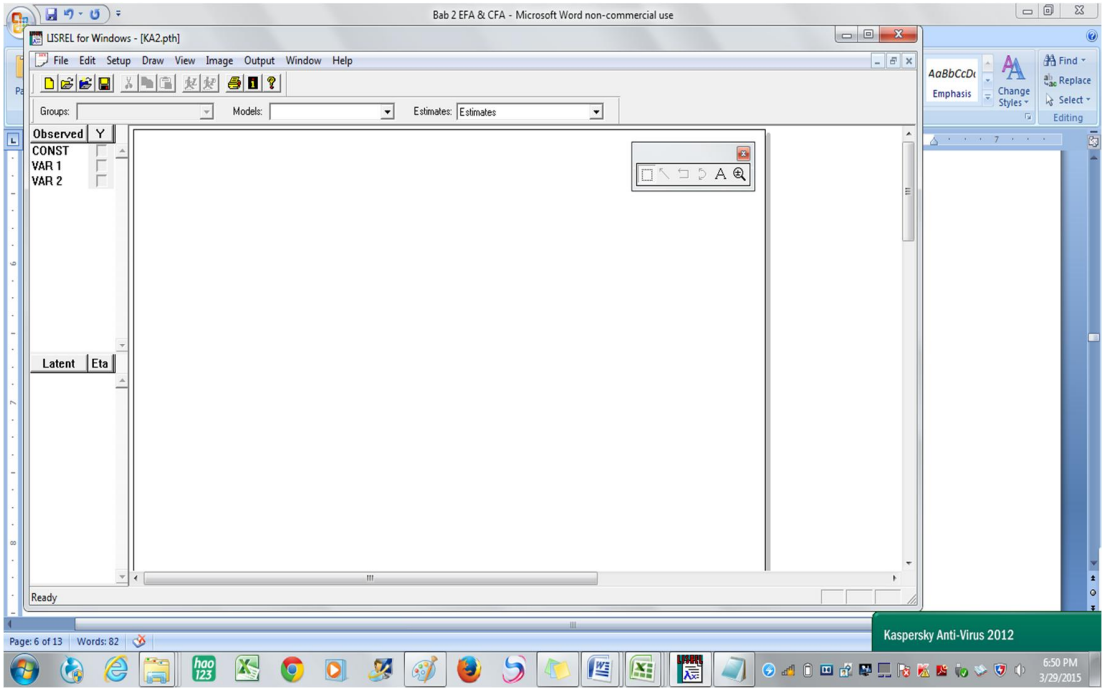
Setelah itu klik menu **Statistics→Output Options**.



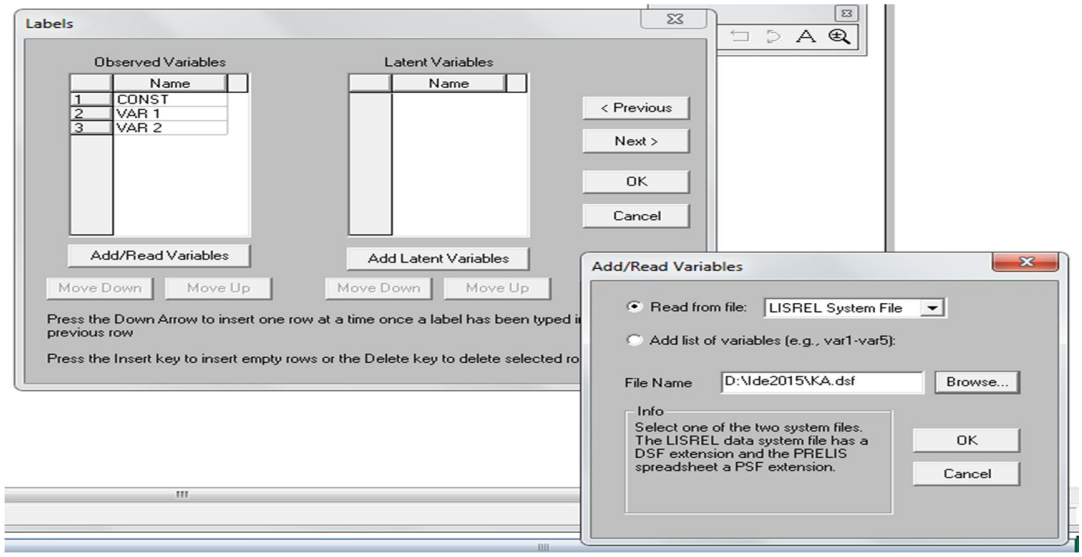
Selanjutnya klik **File→New→Path Diagram**.



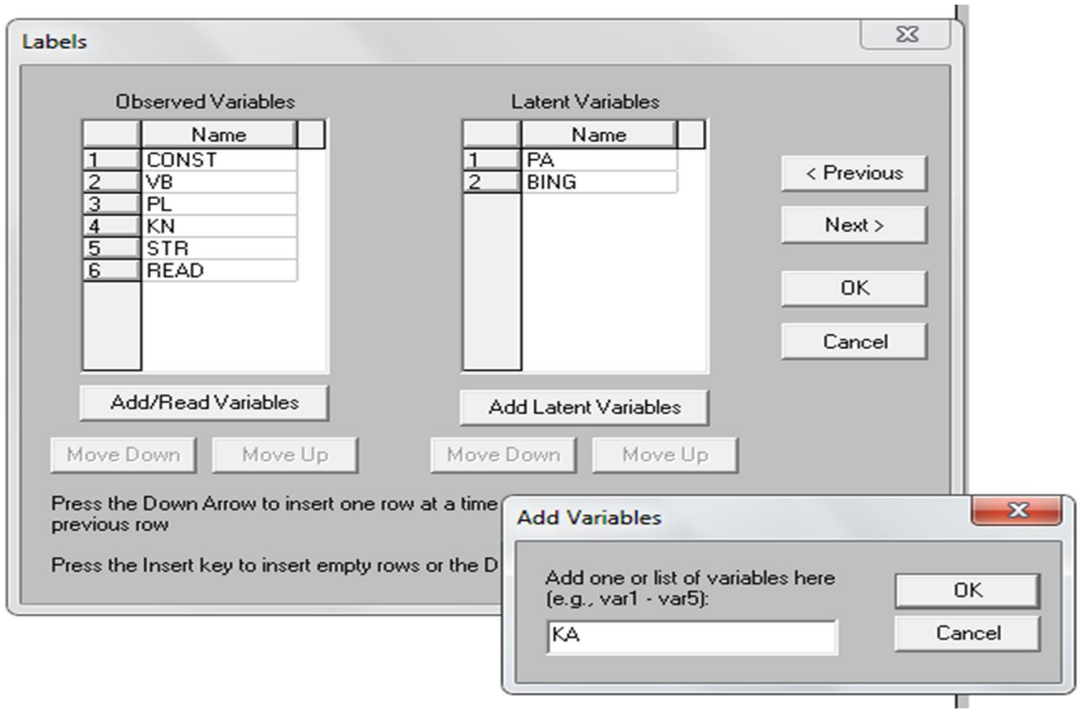
Selanjutya komputer akan meminta diagram disimpan dengan nama *.pth. Kemudan akan muncul bidang gambar seperti berikut.



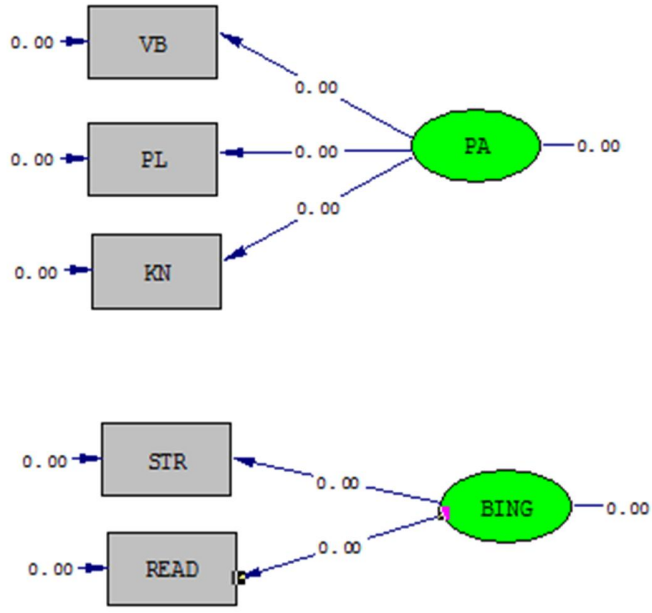
Klik menu **Setup→Variable→Add/Read Variables→Browse**. Kemudian diklik file *.dsf.



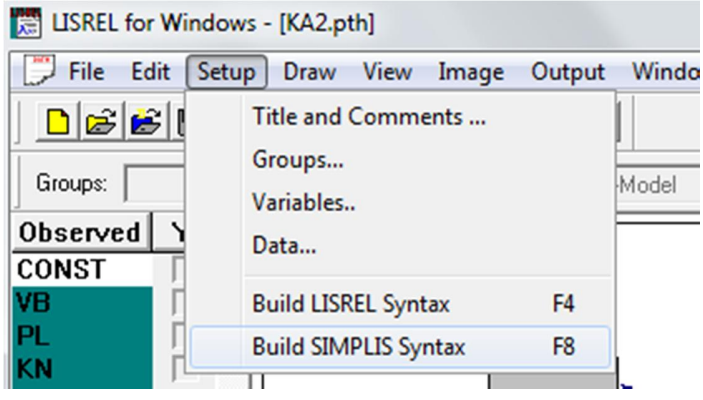
Variabel yang dapat diamati (*observable*) kemudian akan muncul sendiri. Untuk memasukkan laten variabel, diklik add latent variables ketikkan namanya, kemudian klik OK.



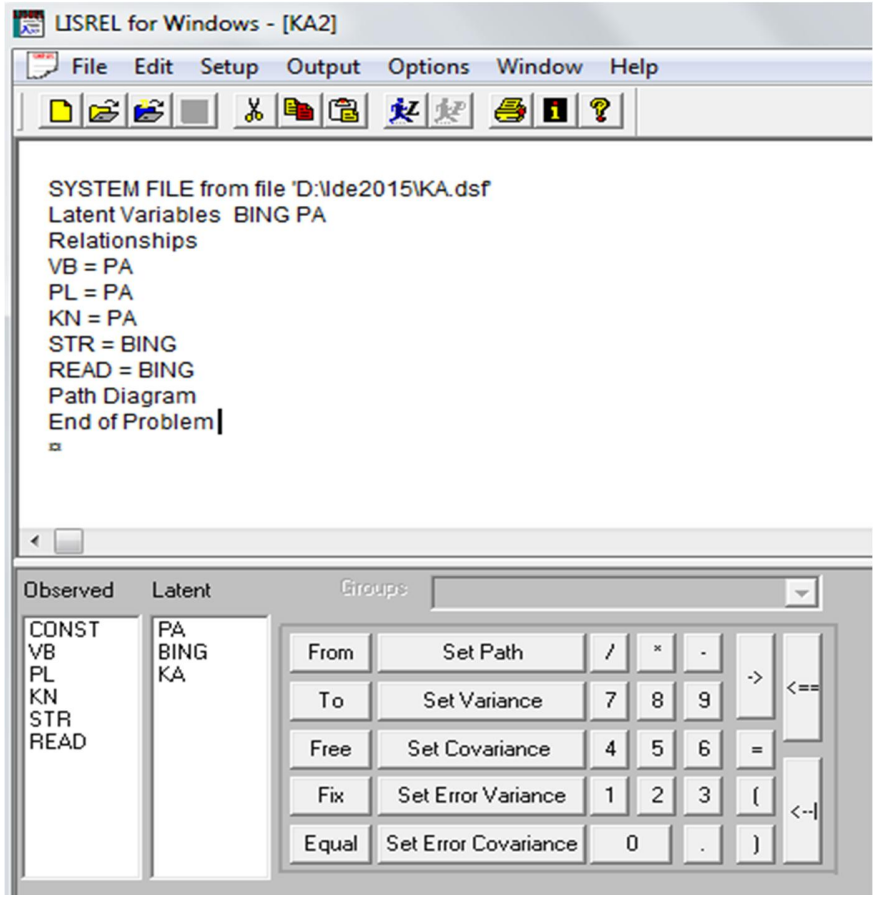
Kemudian dengan *drag and drop*, digambarkan model yang diinginkan. Pertama digambarkan dahulu variabel *observable* (kotak), kemudian variabel laten (oval), dan hubungan antara variabel laten dengan variable *observable*. Untuk first order analisis diperoleh gambar sebagai berikut.



Langkah selanjutnya adalah membangun perintah atau sintaks. Klik **Setup→Build LISREL Syntax (F4)** atau **Klik Setup→ Build Lisrel Syntax (F8).**



Ketika pilihannya F8, akan muncul sebagai berikut untuk *first order analysis* (CFA). Selanjutya dipilih menu **Run** (klik gambar orang berlari).

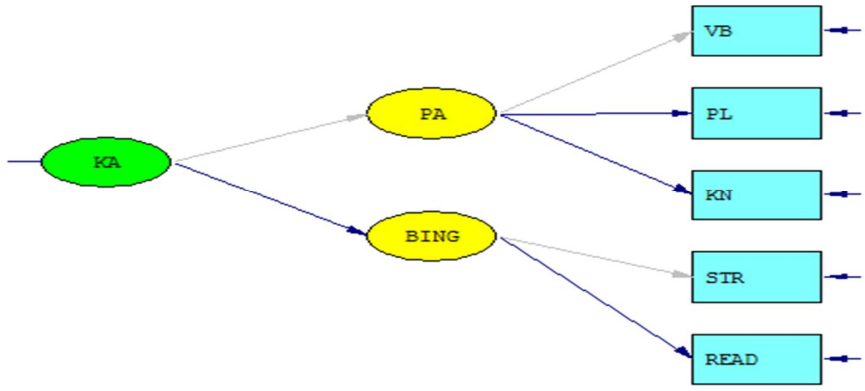


Jika akan melakukan analisis **second order confirmatory factor analysis**, sintaks diubah/ diketikkan sehingga menjadi seperti berikut ini.

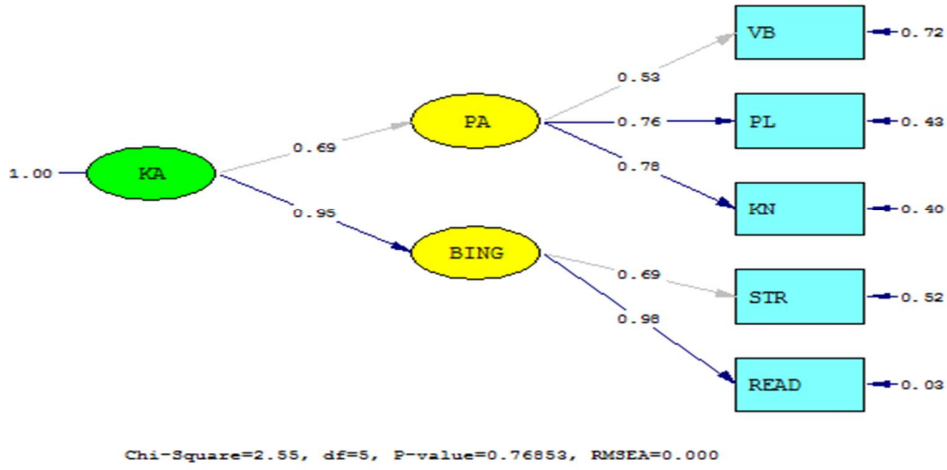
```
TI
SYSTEM FILE from file 'D:\Ide2015\KA.dsf'
Sample Size = 129
Latent Variables  PA BING KA
Relationships
VB = 1.000*PA
PL = PA
KN = PA
STR = 1.000*BING
READ = BING
PA = 1*KA
```

```
BING = KA
Set the Variance of PA to 1.00
Set the Variance of BING to 1.00
Path Diagram
Number of Decimals = 3
End of Problem
```

Setelah di klik gambar orang berlari (**Run**), diperoleh model konseptual berikut.

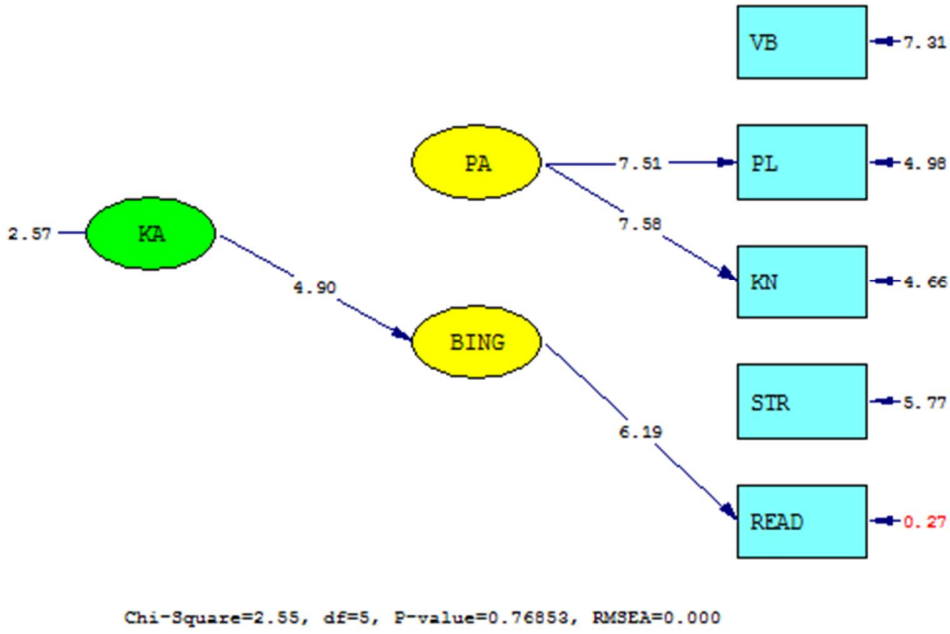


Adapun model terstandarnya (*standardized*) sebagai berikut. Model standar merupakan model yang diharapkan, karena menunjukkan muatan faktor (koefisien jalur dari variabel ke variabel). Beberapa ahli mengatkan, bahwa koefisien jalur ini memiliki arti (*meaningful*) jika besarnya tidak kurang dari 0,4 dan signifikan yang ditunjukkan nilai-T tidak berwarna merah (untuk taraf signifikansi 0,05, nilai-T tidak kurang dari 1,96). Hasil analisis menunjukkan bahwa semua variabel *observable* memberikan sumbangan yang berarti untuk mengukur variabel laten.

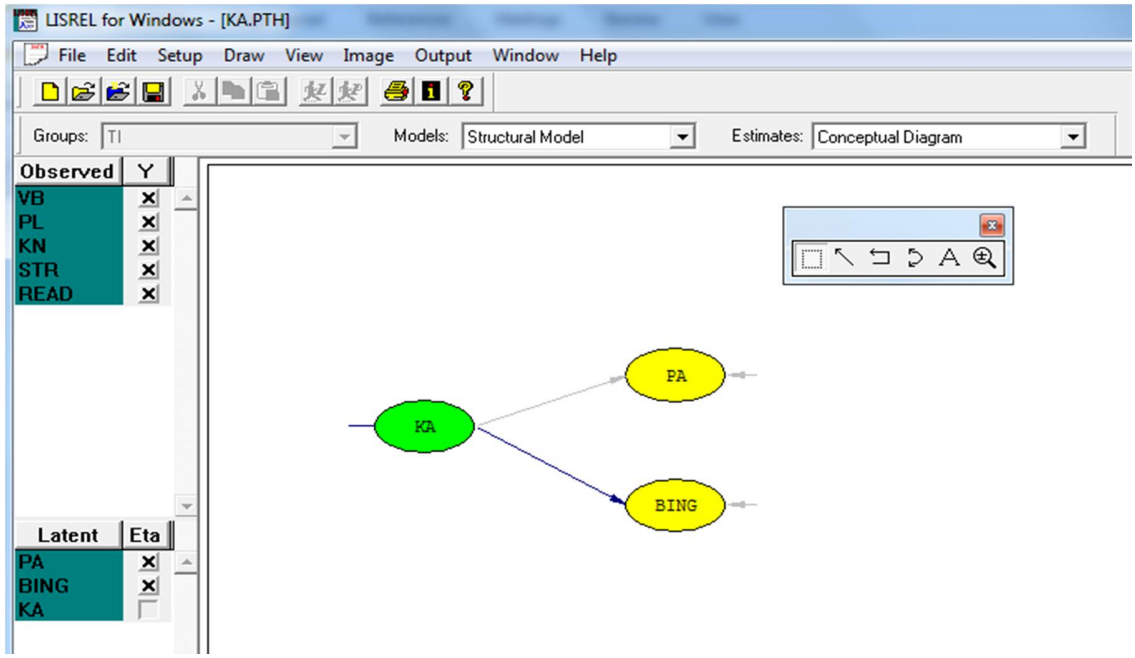


Bagian yang perlu diperhatikan dari gambar tersebut adalah kecocokan model. Ada banyak kriteria, namun yang utama model dikatakan cocok adalah *p-value* lebih dari α dan RMSEA mendekati 0.

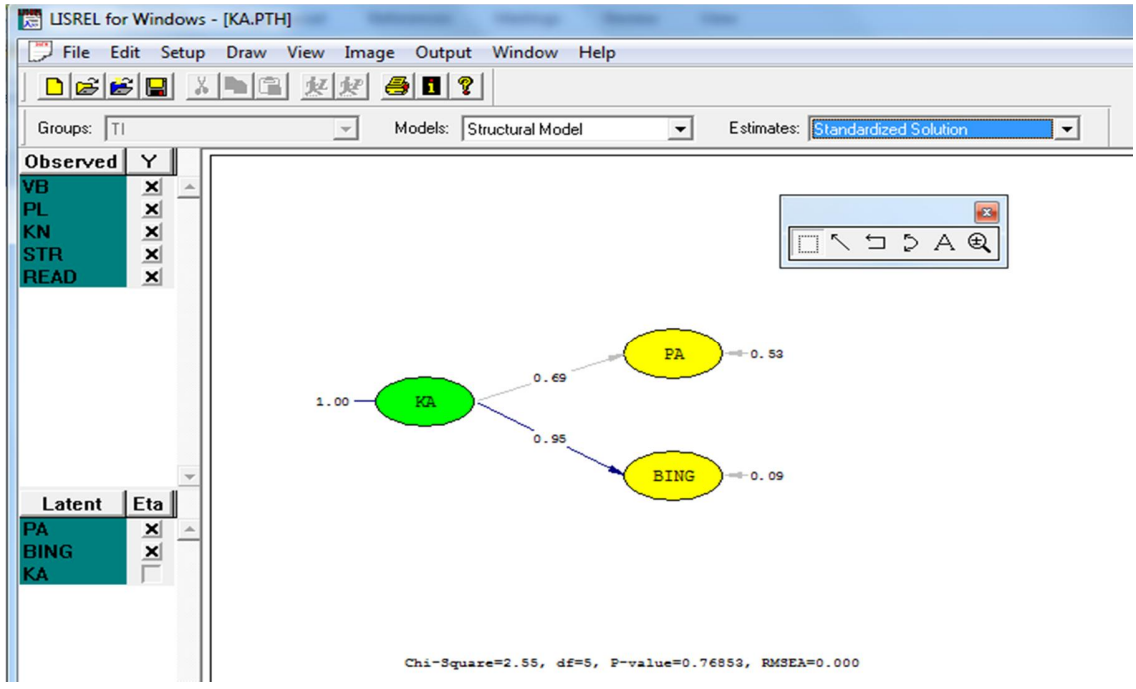
Kemudian untuk melihat signifikansi jalur juga akan diperoleh dengan memilih nilai-T.



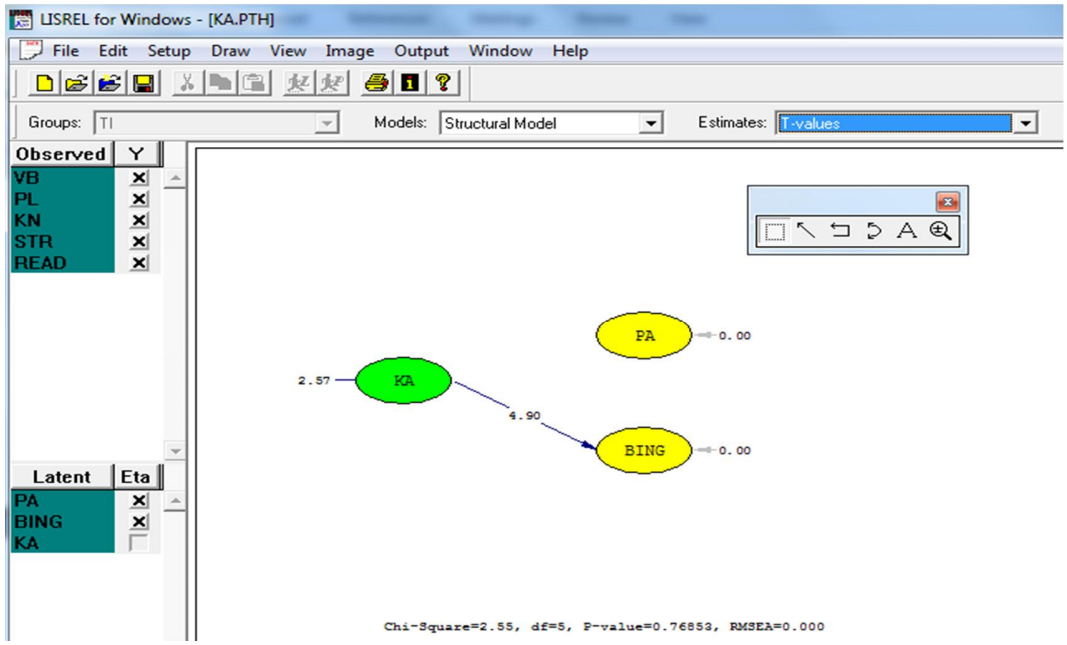
Jika pada **Models** dipilih **Structural Model** dengan **Estimates Conceptual Diagram**, akan diperoleh diagram jalur antar variabel laten.



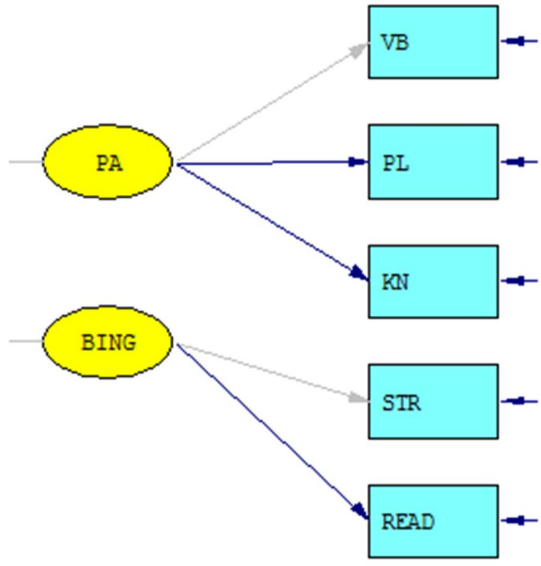
Pada **Standardized Solution** diperoleh koefisien jalurnya.



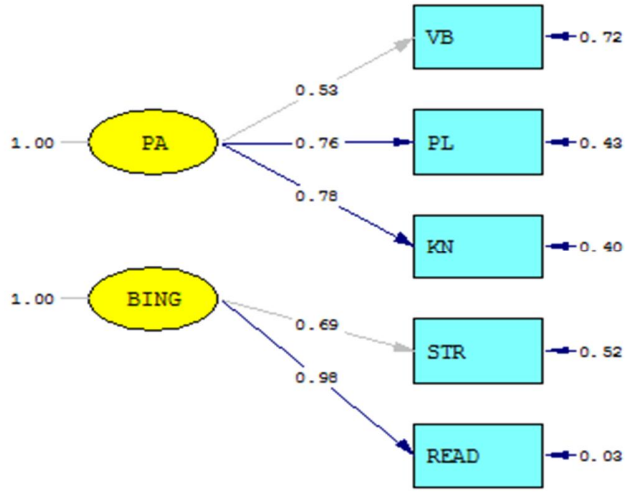
Demikian juga melihat signifikansi jalur dengan memilih nilai-T.



Demikian pula halnya pada **first order analysis**, dapat diperoleh model konseptualnya sebagai berikut.

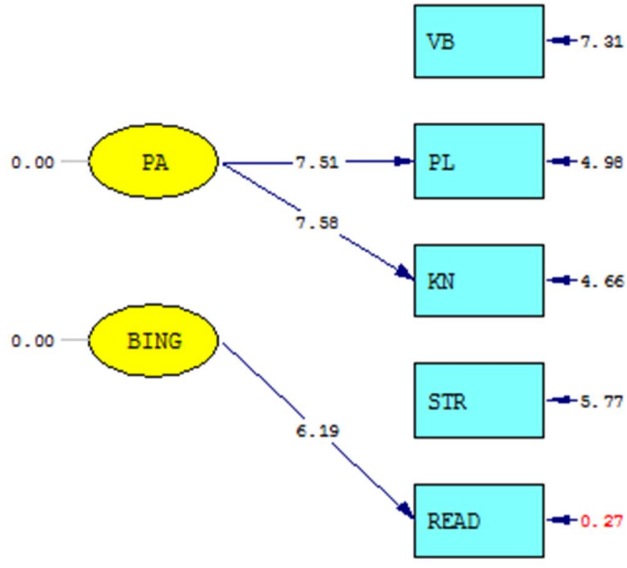


Adapun hasilnya yang terstandar (*standardized solution*) sebagai berikut.



Chi-Square=2.55, df=5, P-value=0.76853, RMSEA=0.000

Hasil tersebut menunjukkan bahwa semua variabel *observable* memberikan sumbangan yang signifikan terhadap variabel laten, demikian pula berdasarkan nilai-T semua jalur signifikan. Pada kasus ini, jalur dari PA ke VB dan BING ke STR dibuat tetap (*fixed*) sehingga jalurnya tidak muncul.



Chi-Square=2.55, df=5, P-value=0.76853, RMSEA=0.000

Bab V

MEMBUKTIKAN VALIDITAS KRITERIA

A. Bukti Validitas Kriteria

Untuk membuktikan validitas kriteria, diperlukan suatu kriteria (*criterion*). Jika validitas kriteria suatu tes atau suatu instrumen akan dibuktikan, diperlukan tes lain yang mengukur konsruk yang bersesuaian. Tes lain yang dijadikan kriteria ini biasanya tes yang dianggap lebih terstandar. Sebagai contoh, asosiasi guru bahasa Inggris sebagai bahasa asing di Indonesia (TEFLIN) mengembangkan *Test of English Proficiency* (TOEP). Untuk membuktikan validitas kriteria TOEP, digunakan tes yang dianggap lebih terstandar yaitu Test of *English as Foreign Language* (TOEFL) sebagai kriterianya (Suwarsih Madya, Ali Saukah, Heri Retnawati, Suharso,2008). Pada kasus ini, instrumen yang akan dibuktikan validitasnya adalah TOEP, sedangkan TOEFL dianggap sebagai kriteria, karena merupakan tes yang lebih terstandar.

Untuk membuktikan validitas kriteria, suatu sampel diminta mengerjakan tes/instrumen yang akan dibuktikan validitasnya dan juga mengerjakan tes/instrumen yang dijadikan kriteria. Jika waktu pengerjaan kedua tes atau instrumen tersebut berdekatan, maka proses pembuktian validitasnya dinamai dengan validitas kriteria jenis *concurrent validity*. Jika jarak waktu pengerjaan kedua tes cukup lama, maka pembuktian validitasnya dinamai dengan validitas kriteria jenis *predictive validity*.

Pada kasus membuktikan validitas kriteria dari TOEP, sejumlah peserta tes diminta mengerjakan TOEP. Hasilnya kemudian diskor, menghasilkan skor TOEP. Kemudian peserta tes ini didanai untuk mengikuti TOEFL, yang kemudian dari lembaga yang bersangkutan memeroleh skor TOEFL. Pada proses validasi ini diperlukan dana yang cukup besar, karena TOEFL yang dijadikan kriteria merupakan tes yang standar dan hasilnya diakui secara internasional.

Validitas kriteria dapat dibuktikan dengan menghitung koefisien korelasi antara skor peserta yang diperoleh dari mengerjakan perangkat yang divalidasi dengan skor yang diperoleh dari mengerjakan perangkat yang dianggap kriteria. Untuk mempermudah, skor peserta yang diperoleh dari mengerjakan perangkat yang divalidasi diberikan simbol x, sedangkan skor yang

diperoleh dari mengerjakan perangkat yang dianggap kriteria diberikan simbol y. Pada statistika, untuk x variabel bebas dan y variabel terikat, dan keduanya merupakan nilai dalam skala interval atau rasio, maka koefisien korelasi ρ dapat dicari dengan mengestimasinya, menggunakan persamaan :

$$\hat{\rho}_{X,Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 . \sum (Y - \bar{Y})^2}} \qquad (5.1)$$

atau untuk memudahkan perhitungan dapat ditulis sebagai

$$\hat{\rho}_{X,Y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} \qquad (5.2)$$

Kuadrat nilai korelasi ini diinterpretasikan sebagai sumbangan variabel X terhadap variabel Y atau yang sering disebut sebagai koefisien determinasi.

Besarnya koefisien korelasi dapat digunakan untuk menentukan besarnya koefisien validitas kriteria. Semakin besar koefisien $\hat{\rho}_{X,Y}$ maka semakin tinggi pula validitas kriterianya, baik untuk jenis prediktif maupun *concurrent*. Besarnya koefisien validitas ini digunakan untuk menghitung kesalahan standar estimasi prediksi, dengan menggunakan formula

$$SE_{estimasi} = SD_y\sqrt{1 - r_{xy}^2} \qquad (5.3)$$

Dengan r_{xy}^2 merupakan koefisien determinasi dari validitas, dan SD_y merupaka standar deviasi dari skor kriteria.

Manfaat lain yang diperoleh dengan adanya validitas kriteria adalah skor prediksi. Sebagai contoh, dengan memanfaatkan skor TOEP, kemampuan peserta jika diukur menggunakan TOEFL dapat diprediksikan. Skor prediksi ini dapat diperoleh setelah mengestimasi persamaan regresi

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (5.4)$$

Dengan Y = Peubah tak bebas pada skor kriteria, X = Peubah bebas pada skor tes atau instrumen yang akan divalidasi, β_0 = intersep/perpotongan dengan sumbu tegak, β_1 = Kemiringan/gradien, ε_i *error* yang saling bebas dan menyebar normal $N(0,\sigma^2)$ $i = 1, 2, \ldots, n$.

Jika untuk perangkat tes atau instrumen yang terdiri dari banyak subtes, maka dapat digunakan regresi linear ganda. Regresi linear ganda adalah persamaan regresi yang menggambarkan hubungan antara lebih dari satu peubah bebas (X) dan satu peubah tak bebas (Y). Hubungan peubah-peubah tersebut dapat dituliskan dalam bentuk persamaan:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (5.5)$$

Y = Peubah tak bebas (skor kriteria) , X = Peubah bebas (skor peserta terhadap instrumen yang divalidasi) , β_0 = intersep/perpotongan dengan sumbu tegak, β_1, β_2, ..., β_{p-1} = parameter model regresi, ε_i saling bebas dan menyebar normal $N(0,\sigma^2)$, $i = 1, 2, \ldots, n$

Persamaan regresi dugaannya adalah

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1} \qquad (5.6)$$

Berikut disajikan contoh membuktikan validitas ktiteria TOEP yang disarikan dari artikel Heri Retnawati (2009).

Contoh Membuktikan Validitas Kriteria TOEP

Contoh kasus untuk membuktikan validitas kriteria ini pada pembuktian validitas perangkat *Test of English Proficiency*, TOEP. Selama ini alat ukur yang digunakan untuk mendapatkan informasi tentang kemahiran berbahasa Inggris adalah tes bahasa Inggris yang dibuat oleh lembaga asing, misalnya TOEFL, TOEIC, dan IELTS. Biaya untuk mengikuti tes ini cukup mahal, tetapi memang hasilnya jelas diakui di semua Negara karena memang tes-tes tsb bersifat standar, yang telah dikembangkan melalui serentetan kegiatan yang ditujukan untuk menjaga agar tes yang dihasilkan memenuhi kriteria tes yang baik.

Sebagai contoh adalah membuktikan validitas kriteri (*Test of English Proficiency* atau TOEP). Diawali pada tahun 2007, Direktorat PSMA memandang perlu untuk segera mengembangkan tes profisiensi bahasa Inggris (*Test of English Proficiency* atau TOEP), yang mengukur kemahiran menggunakan bahasa Inggris dalam dunia nyata para lulusan SMA. Pada tahun 2007 telah dimulai pengembangan seperangkat instrumen pengukuran kemahiran menggunakan bahasa Inggris tersebut, yang dilanjutkan tahun 2008 dan 2009. Selama 3 tahun

(2007-2009) telah dikembangkan 7 perangkat TOEP yang diberi nama TOEP 1, 2A, 2B, 3A, 3B, 4, dan 5 yang saling paralel (Suwarsih Madya, dkk, 2008).

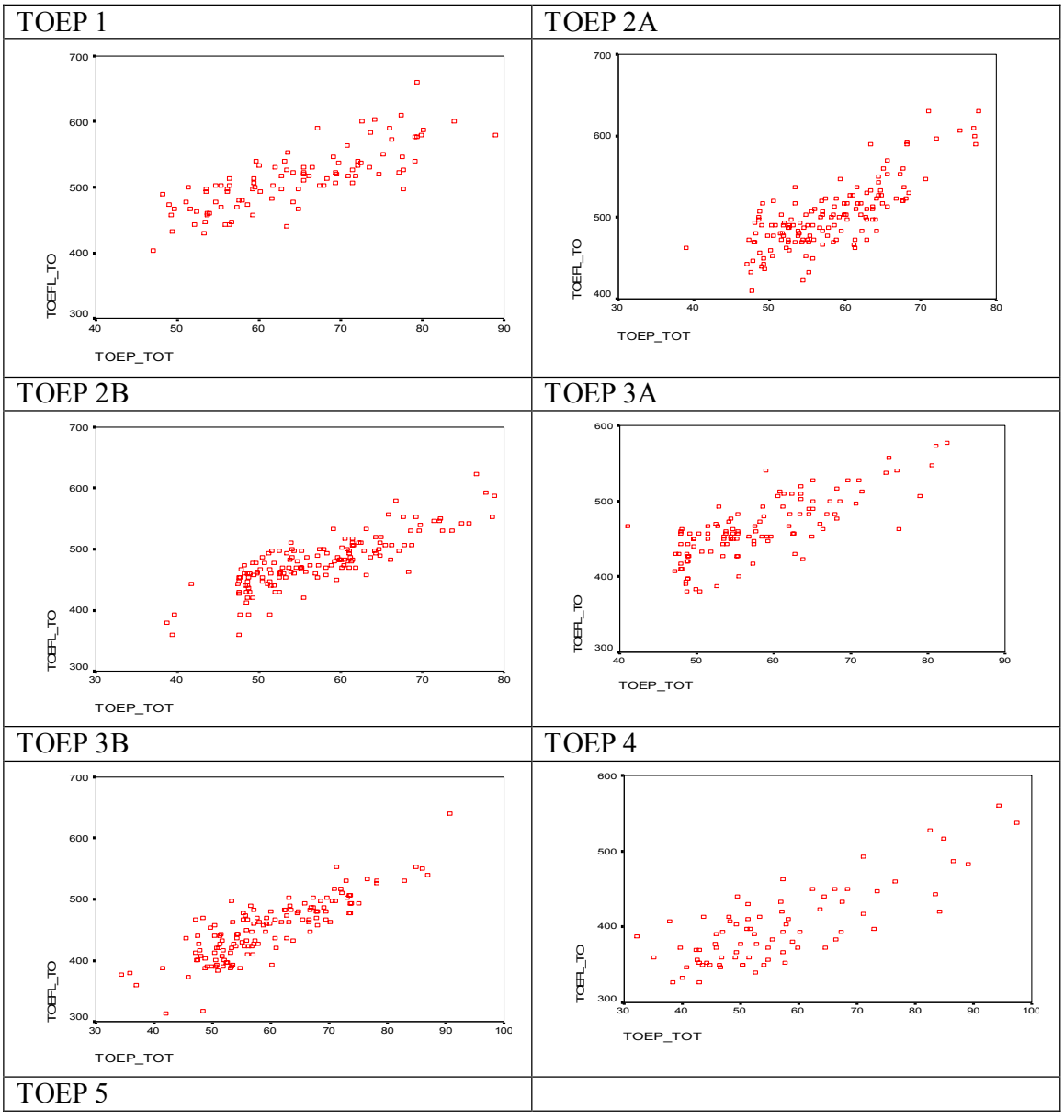
Tes Kemahiran Bahasa Inggris (*Test of English Proficiency, TOEP*) yang merupakan tes standar untuk mengukur kemahiran berbahasa Inggris siswa Sekolah Menengah Atas (SMA). TOEP yang dikembangkan merupakan tes tertulis (*paper and pencil test*) pada tahun 2007 dan 2008, dan selanjutnya dirintis tes untuk mengukur kemampuan *Speaking* dan *Reading* di tahun 2009 dan 2010. Penskoran tiap butir dilakukan dengan sistem dikotomi, benar diberi skor 1 dan jika salah diberi skor 0. Tes ini khusus mengukur kemahiran siswa SMA dalam menggunakan bahasa Inggris, khususnya *Reading* dan *Listening*. Tes terdiri dari 100 butir soal, dengan rincian 50 butir tes *Reading* dan 50 butir tes *Listening*. Terkait dengan tes yang dikembangkan merupakan tes standar internasional, pada kegiatan ini juga dihasilkan petunjuk pelaksanaan TOEP. Hal ini dimaksudkan agar setiap TOEP yang dilaksanakan benar-benar merupakan tes yang terstandar.

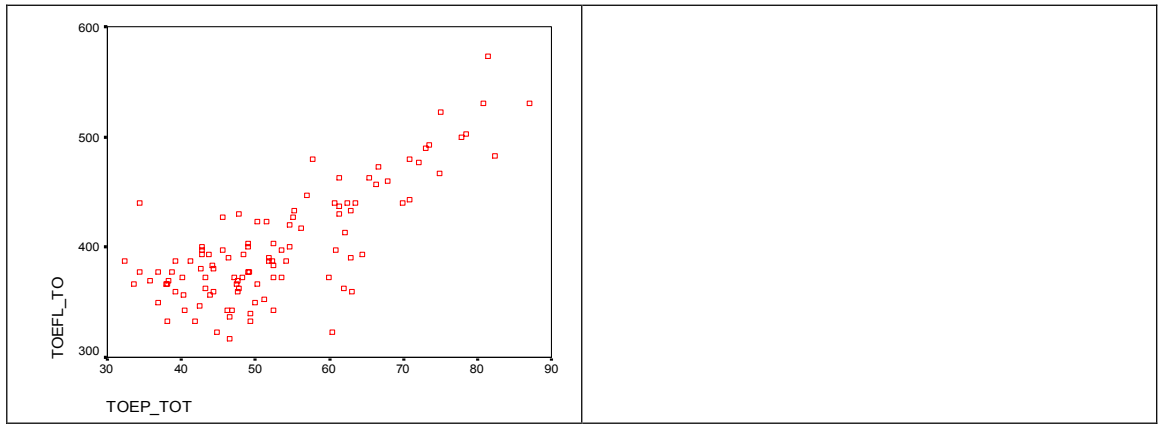
TOEP dikembangkan melalui proses menjabarkan tujuan menjadi indikator-indikator, yang kemudian dikembangkan menjadi butir. Ini berarti TOEP memenuhi syarat tes yang baik ditinjau dari validitas isinya. Validitas kenampakan (*face validity*) untuk menjadi tes yang baik juga terpenuhi, mengingat pengembangan tes ini mulai dari menyusun butir sampai dengan perakitan tes melibatkan ahli yang terkait, baik dari perguruan tinggi maupun dari praktisi di lapangan (guru). Validitas lain yang digunakan yakni validitas *criterion-related evidence of validity* jenis konkuren, yakni mengaitkan skor TOEP dengan skor TOEFL Institusional perolehan siswa (*benchmarking*).

Terkait dengan adanya validitas *criterion-related evidence of validity* jenis konkuren yang dimiliki TOEP, skor perolehan siswa SMA yang menempuh TOEP dapat dikonversikan ke skor tes lain, misalnya TOEFL. Sehubungan dengan adanya dua model regresi untuk memprediksi kemampuan peserta, yakni regresi tunggal dan regresi ganda, pada tulisan ini akan dibandingkan keakuratan kedua model ini dalam memprediksi skor TOEFL siswa SMA di Indonesia. Pada model regresi tunggal digunakan skor TOEP peserta untuk memprediksi skor TOEFL, dan pada regresi ganda digunakan skor Listening TOEP dan skor Reading TOEP untuk memprediksikan skor TOEFL peserta (Heri Retnawati, 2009).

Untuk mengetahui keberadaan hubungan linear antara variabel prediktor dengan variabel kriteria, dibuat diagram pencar (*Scatter Plot*) terlebih dahulu. Pada model regresi tunggal,

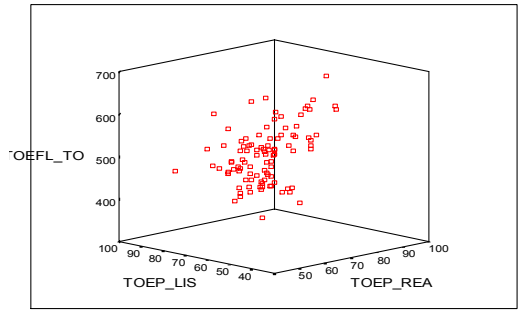
variabel prediktornya merupakan skor TOEP dan variabel kriterianya merupakan skor TOEFL. Hasilnya disajikan pada Gambar 1. Demikian pula pada model regresi ganda, variabel prediktornya merupakan skor TOEP *Listening* dan skor TOEP *Reading* dan variabel kriterianya merupakan skor TOEFL, dengan hasil pada Gambar 5.1.



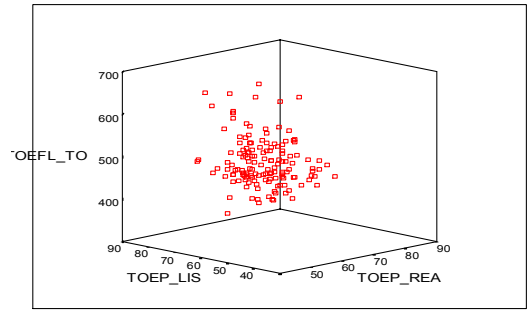


Gambar 5.1. Diagram Pencar Skor TOEP untuk Memprediksi Skor TOEFL

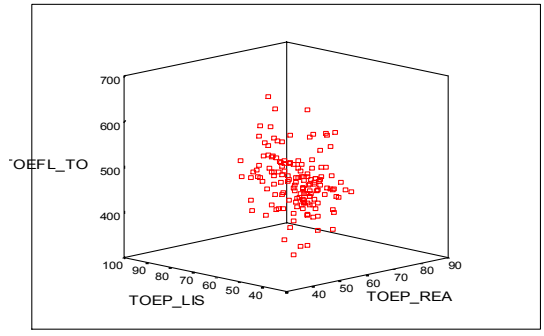
TOEP 1



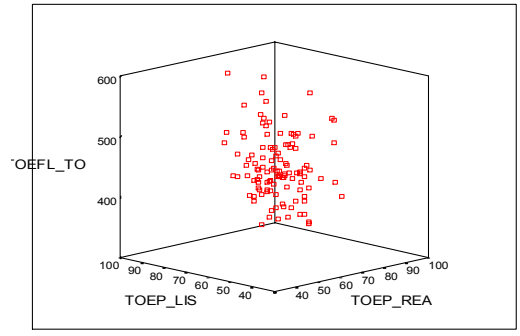
TOEP 2A



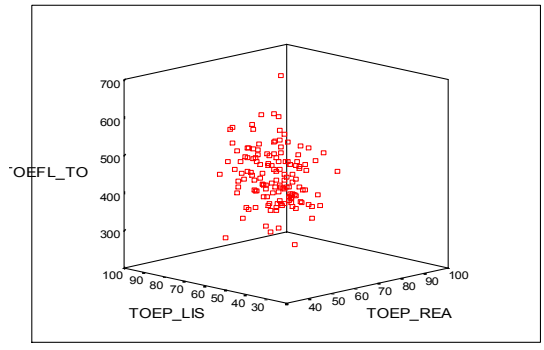
TOEP 2B



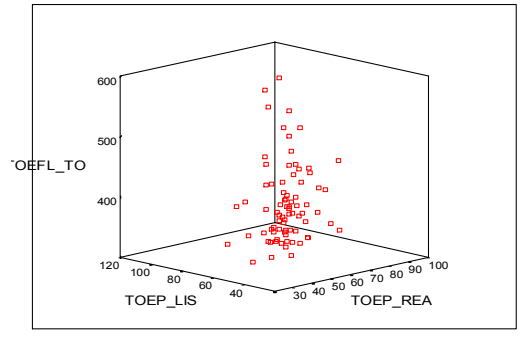
TOEP 3A



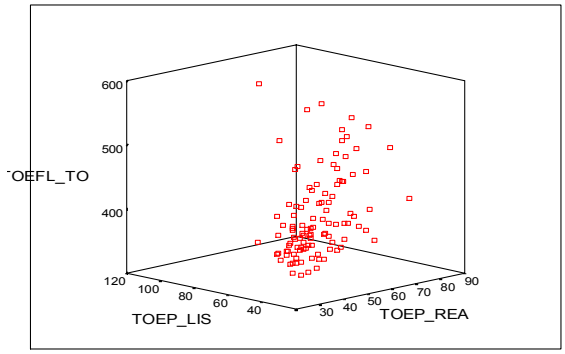
TOEP 3B



TOEP 4



TOEP 5



Gambar 5.2. Diagram Pencar TOEP *Listening* dan TOEP *Reading* untuk Memprediksi Skor TOEFL

Mencermati diagram pencar pada Gambar 5.1, diperoleh bahwa terdapat hubungan linear antara skor TOEP dengan skor TOEFL pada model regresi tunggal. Pada Gambar 5.2 juga menunjukkan adanya hubungan linear antara skor TOEP *Listening* dan TOEP *Reading* untuk Memprediksi skor TOEFL dengan menggunakan regresi ganda. Hasil estimasi korelasi baik pada model regresi tunggal maupun regrasi ganda disajikan pada Tabel 5.1.

Tabel 2.7 Hasil Estimasi Koefisien Korelasi dan Kontribusi

Perangkat TOEP	Model			
	$Y=b_0+b_1X$		$Y= b_0+b_1X_1+b_2X_2$	
	r	r^2	R	r^2
1	0.7943	0.6309	0.8060	0.6497
2A	0.7801	0.6085	0.8081	0.6530
2B	0.8349	0.6970	0.8357	0.6984
3A	0.7687	0.5908	0.7692	0.5917
3B	0.8445	0.7132	0.8467	0.7169
4	0.7910	0.6257	0.7910	0.6257
5	0.7765	0.6030	0.7773	0.6043

Keterangan : $\hat{Y}$ skor TOEFL prediksi, X skor TOEP (regresi tunggal)

$\hat{Y}$ skor TOEFL prediksi, X_1 skor TOEP *Listening*, X_2 skor TOEP *Reading*

(regresi ganda)

Hasil perhitungan korelasi tersebut menunjukkan kecenderungan bahwa korelasi dengan dua prediktor terhadap TOEFL lebih tinggi dibandingkan korelasi dengan prediktor tunggal. Demikian pula koefisien korelasi determinasi (r^2) yang menunjukkan persentase kontribusi TOEP dalam memprediksi TOEFL.

Dengan menggunakan data empiris, selanjutnya dapat diestimasi konstanta dan koefisien pada persamaan regresi, yang disajikan pada Tabel 2.8 untuk model regresi tunggal dan regresi ganda. Persamaan regresi ini digunakan untuk membuat prediksi, jika diperoleh salah satu skor, dapat diestimasi skor yang lain.

Tabel 5.2. Persamaan Regresi untuk Memrediksi Skor TOEFL dengan Skor TOEP

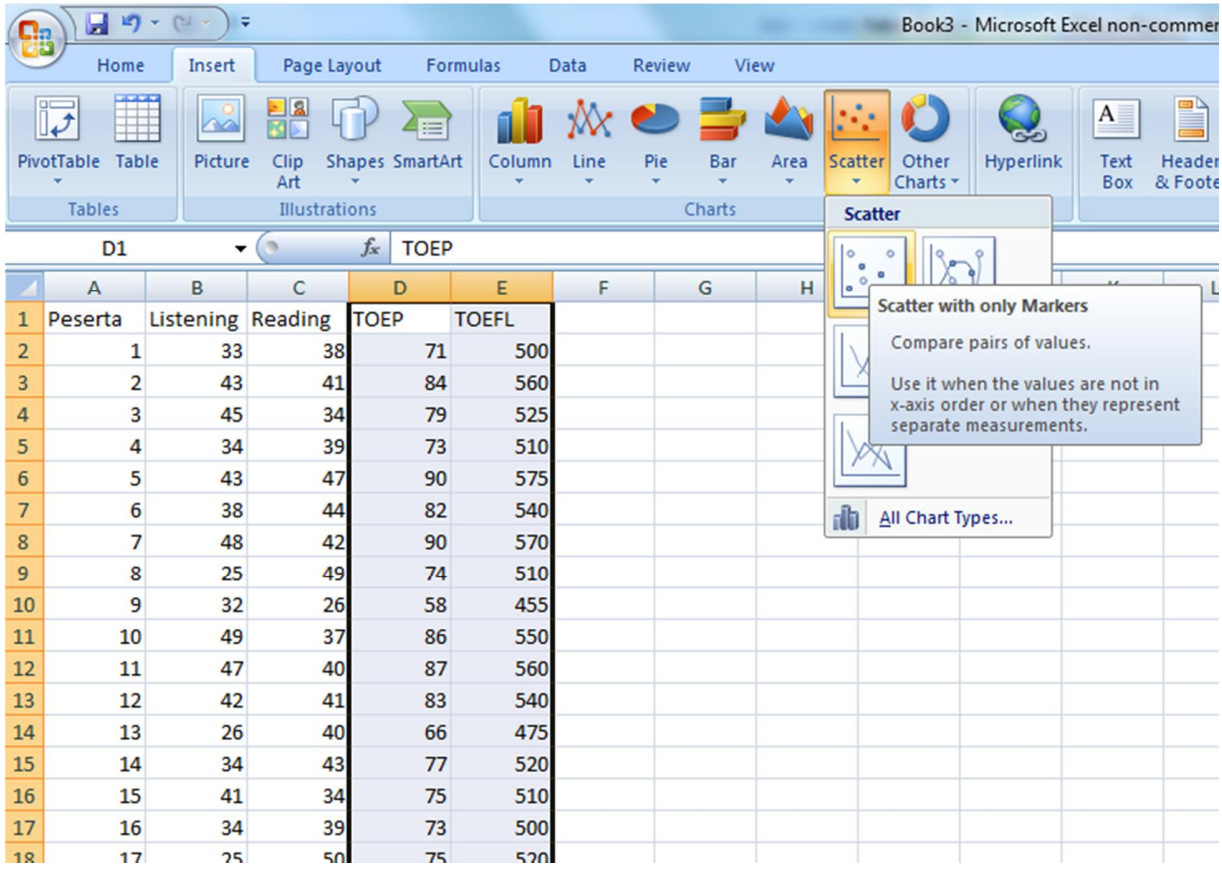
Perangkat TOEP	Persamaan Prediksi (Dengan $\hat{Y}$ skor TOEFL prediksi, X skor TOEP)	Persamaan Prediksi (Dengan $\hat{Y}$ skor TOEFL prediksi, X_1 skor TOEP *Listening*, X_2 skor TOEP *Reading*)
1	$\hat{Y}$ = 3,381 . X + 266,214	$\hat{Y}$ = 274,449+1,285. X_1 +2.401. X_2
2A	$\hat{Y}$ = 4,321 . X + 251,435	$\hat{Y}$ = 264,609+2,977. X_1 +1.120. X_2
2B	$\hat{Y}$ = 4,268 . X + 234,846	$\hat{Y}$ = 239,063+2,273. X_1 +1.922. X_2
3A	$\hat{Y}$ = 3,630 . X + 252,836	$\hat{Y}$ = 254,244+1,917. X_1 +1.692. X_2
3B	$\hat{Y}$ = 3,923 . X + 218,624	$\hat{Y}$ = 223,336+2,210. X_1 +1.634. X_2
4	$\hat{Y}$ = 4,321 . X + 251.435	$\hat{Y}$ = 243,872+1,377. X_1 +1.383. X_2
5	$\hat{Y}$ = 4,268 . X + 234.846	$\hat{Y}$ = 230,464+1,469. X_1 +1.759. X_2

Mencermati besarnya nilai korelasi dan indeks determinasi hasil analisis tersebut, dapat diperoleh bahwa telah terbukti secara empiris bahwa TOEP terbukti valid atau memiliki validitas kriteria.

B. Membuktikan Validitas Kriteria

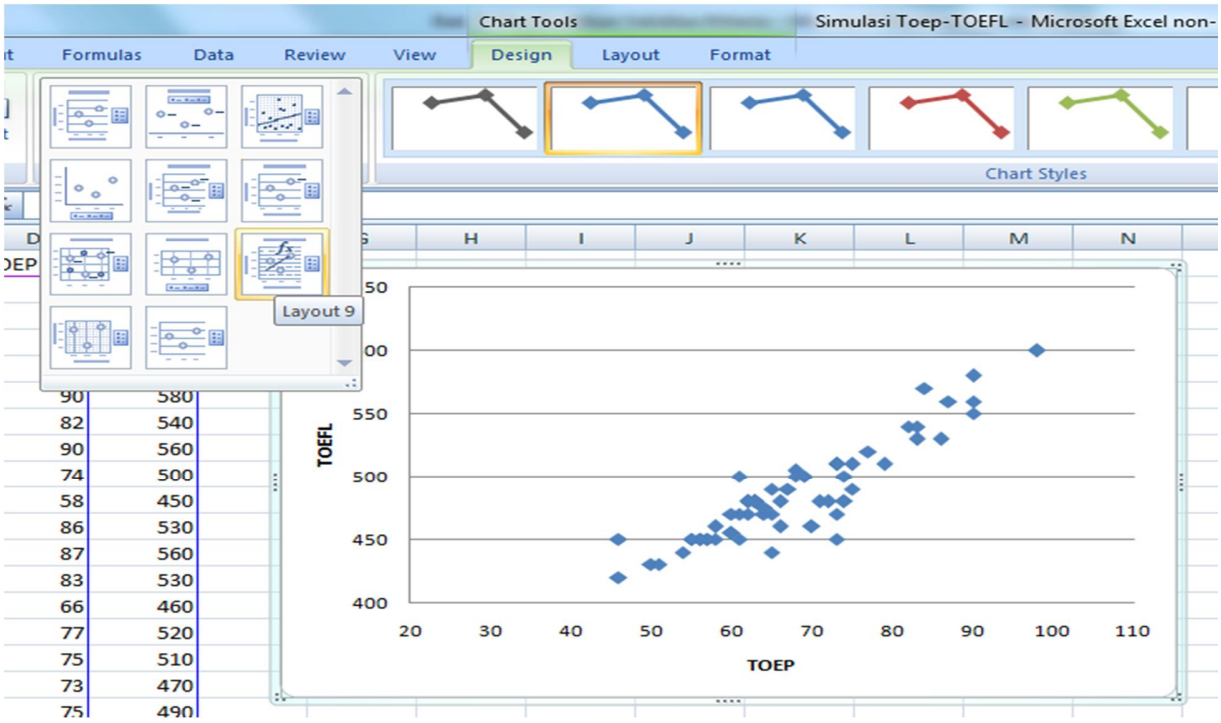
Untuk membuktikan validitas kriteria, dapat digunakan beberapa bantuan. Secara sederhana, menghitung koefisien korelasi dapat dilakukan dengan bantuan kalkulator maupun program Excel. Pada buku ini, akan disajikan cara membuktikan validitas kriteria dengan Excel untuk perangkat yang diskor tunggal (univariat), dan cara membuktikan validitas kriteria untuk perangkat dengan skor yang terdiri dari beberapa bagian (multivariat).

Input data dapat dilakukan langsung pada baris-baris dalam worksheet Excel. Untuk mengetahui pola hubungan antara skor perangkat yang akan divalidasi dengan skor perangkat kriteria, diblok kolom yang memuat kedua skor tersebut, klik **Insert**, kemudian **Scatter**, dipilih model yang diinginkan.

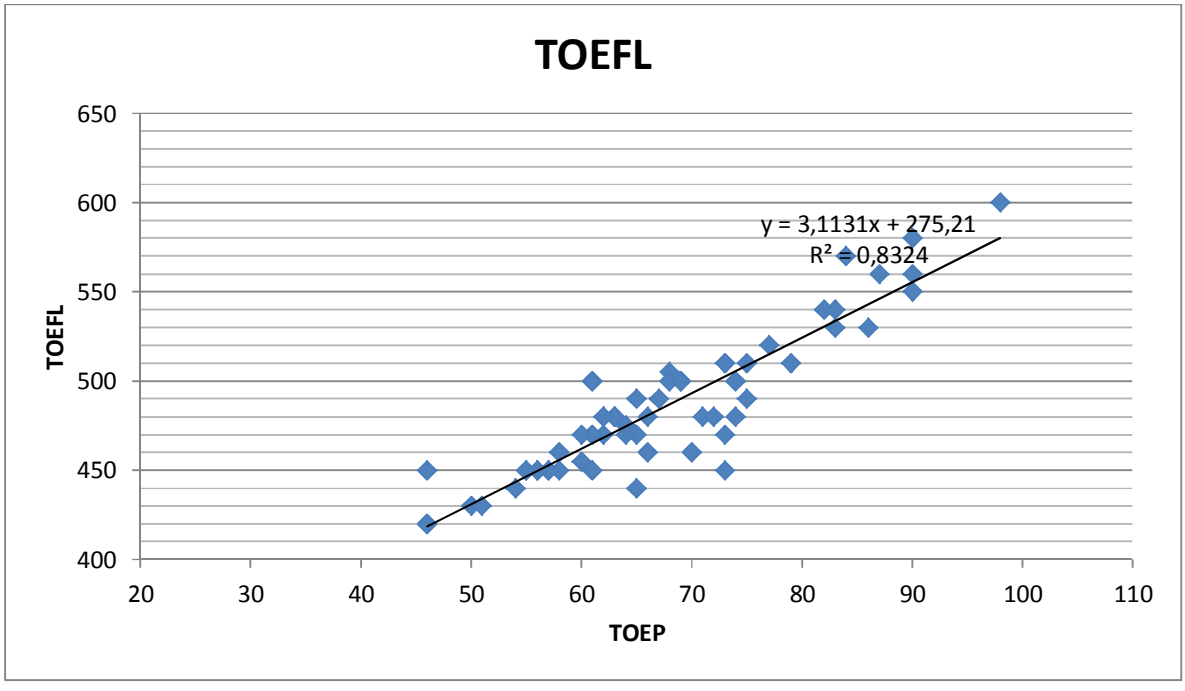


	A	B	C	D	E
1	Peserta	Listening	Reading	TOEP	TOEFL
2	1	33	38	71	500
3	2	43	41	84	560
4	3	45	34	79	525
5	4	34	39	73	510
6	5	43	47	90	575
7	6	38	44	82	540
8	7	48	42	90	570
9	8	25	49	74	510
10	9	32	26	58	455
11	10	49	37	86	550
12	11	47	40	87	560
13	12	42	41	83	540
14	13	26	40	66	475
15	14	34	43	77	520
16	15	41	34	75	510
17	16	34	39	73	500
18	17	25	50	75	520

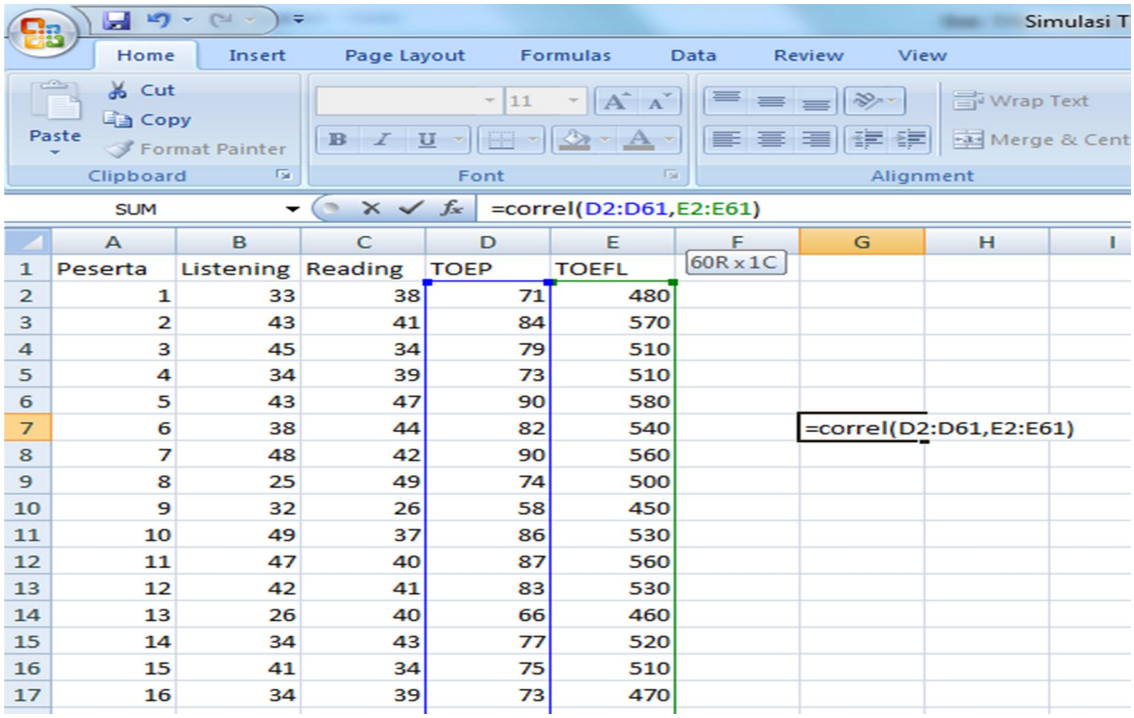
Scater-plot yang diperoleh dapat dimodifikasi untuk mendapatkan persamaan regresi dan koefisien determinasi. Dengan mengeklik **Chart Tools**, kemudian dipilih yang ada grafiknya dan ada fungsinya f_x.



Selanjutnya akan diperoleh persamaan prediksi untuk variabel y dengan menggunakan variabel x, yaitu y=3,113 x + 275,2 dan juga koefisien determinasinya 0,832 atau 83,2%.

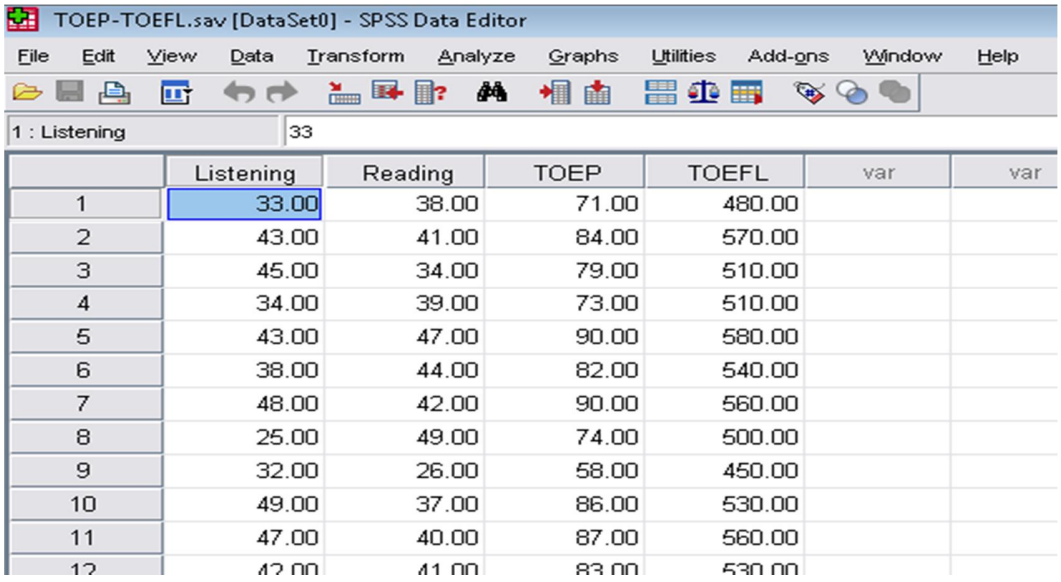


Untuk menghitung korelasi, dapat digunakan fungsi **=CORREL**, kemudian array nya atau kolom yang akan dihitung variabelnya untuk variabel x dan variabel y dimasukkan. Jika dienter, akan diperoleh koefisien korelasi.



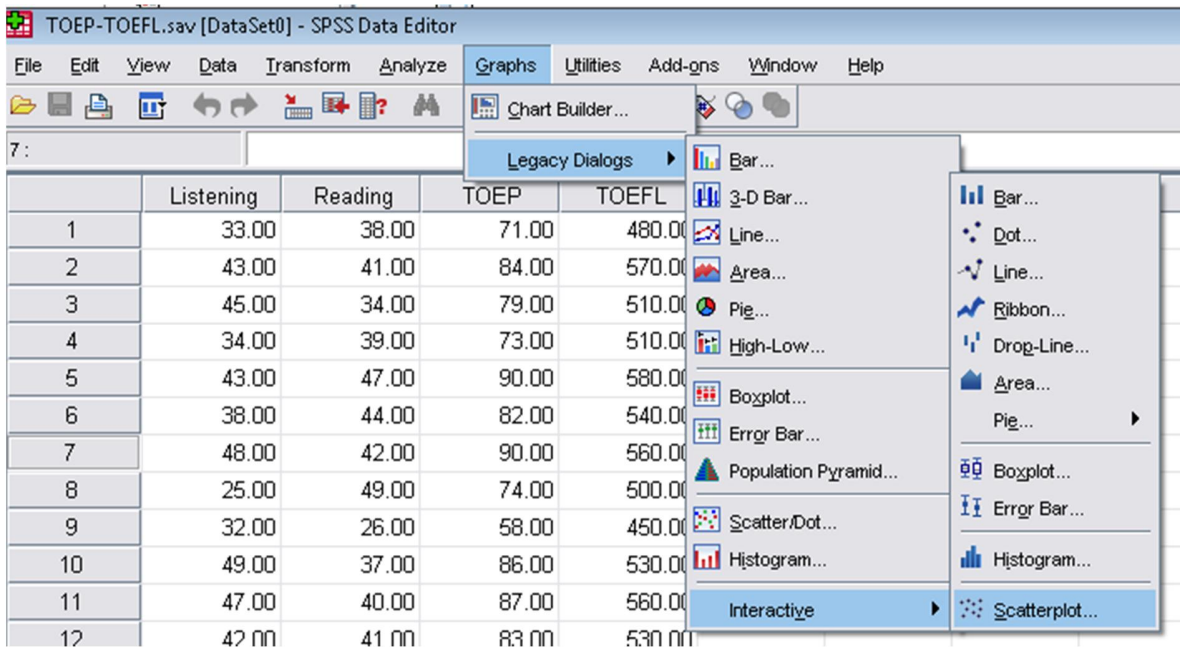
	A	B	C	D	E	F	G	H	I
1	Peserta	Listening	Reading	TOEP	TOEFL				
2	1	33	38	71	480				
3	2	43	41	84	570				
4	3	45	34	79	510				
5	4	34	39	73	510				
6	5	43	47	90	580				
7	6	38	44	82	540				
8	7	48	42	90	560				
9	8	25	49	74	500				
10	9	32	26	58	450				
11	10	49	37	86	530				
12	11	47	40	87	560				
13	12	42	41	83	530				
14	13	26	40	66	460				
15	14	34	43	77	520				
16	15	41	34	75	510				
17	16	34	39	73	470				

Cara lain untuk melakukan analisis pembuktian validitas prediktif adalah menggunakan SPSS. Cara ini dilakukan dengan menginputkan data terlebih dahulu. Pada variabel view, dimasukkan nama variabel yang akan digunakan.

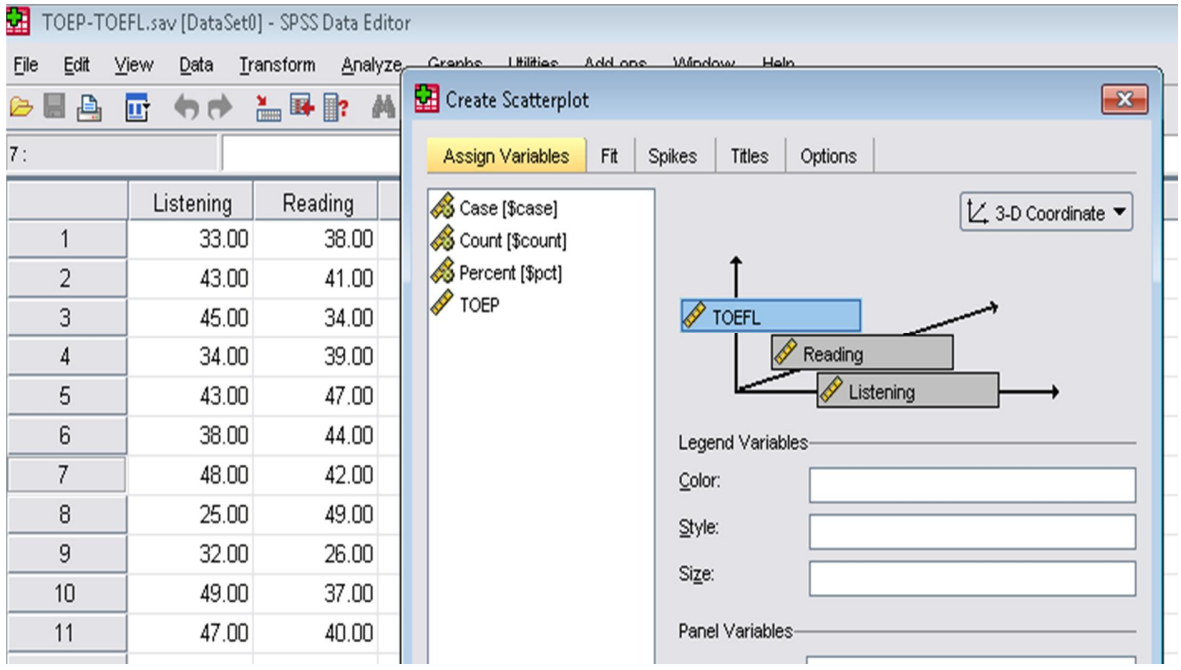


	Listening	Reading	TOEP	TOEFL	var	var
1	33.00	38.00	71.00	480.00		
2	43.00	41.00	84.00	570.00		
3	45.00	34.00	79.00	510.00		
4	34.00	39.00	73.00	510.00		
5	43.00	47.00	90.00	580.00		
6	38.00	44.00	82.00	540.00		
7	48.00	42.00	90.00	560.00		
8	25.00	49.00	74.00	500.00		
9	32.00	26.00	58.00	450.00		
10	49.00	37.00	86.00	530.00		
11	47.00	40.00	87.00	560.00		
12	42.00	41.00	83.00	530.00		

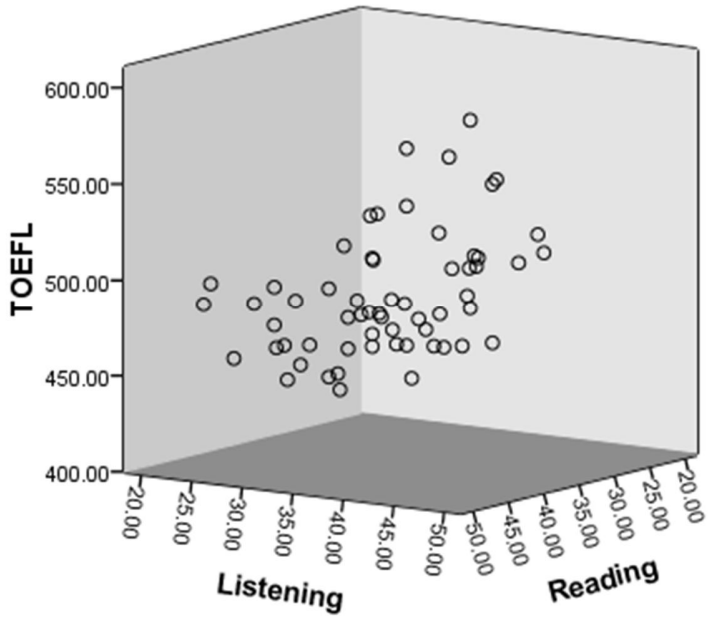
Untuk menggambar grafik *scatterplot*, dapat dipilih menu **Graphs, Legacy Dialogs, Interactive, Scatterplot**.



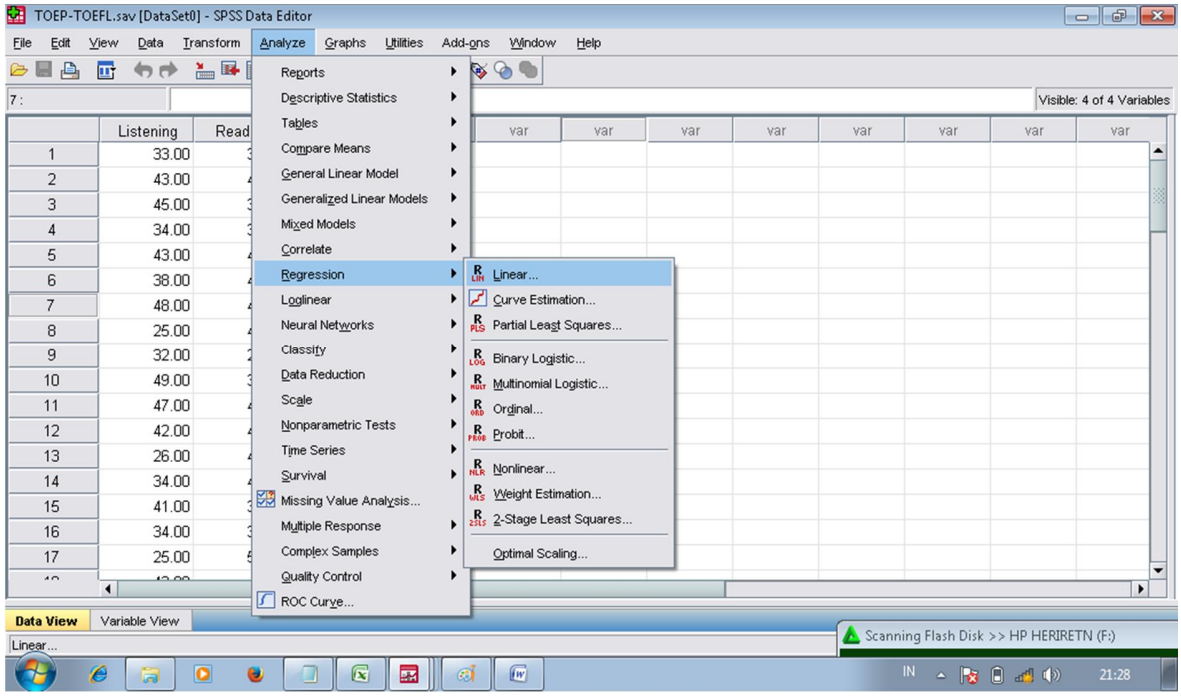
Jika yang akan dibuat adalah grafik 3 dimensi, maka dipilih **3-D Coordinate** di menu kanan atas. Pada list variabel, di drag and drop variabel yang akan disajikan pada sumbu-x, sumbu-y, dan sumbu-z.



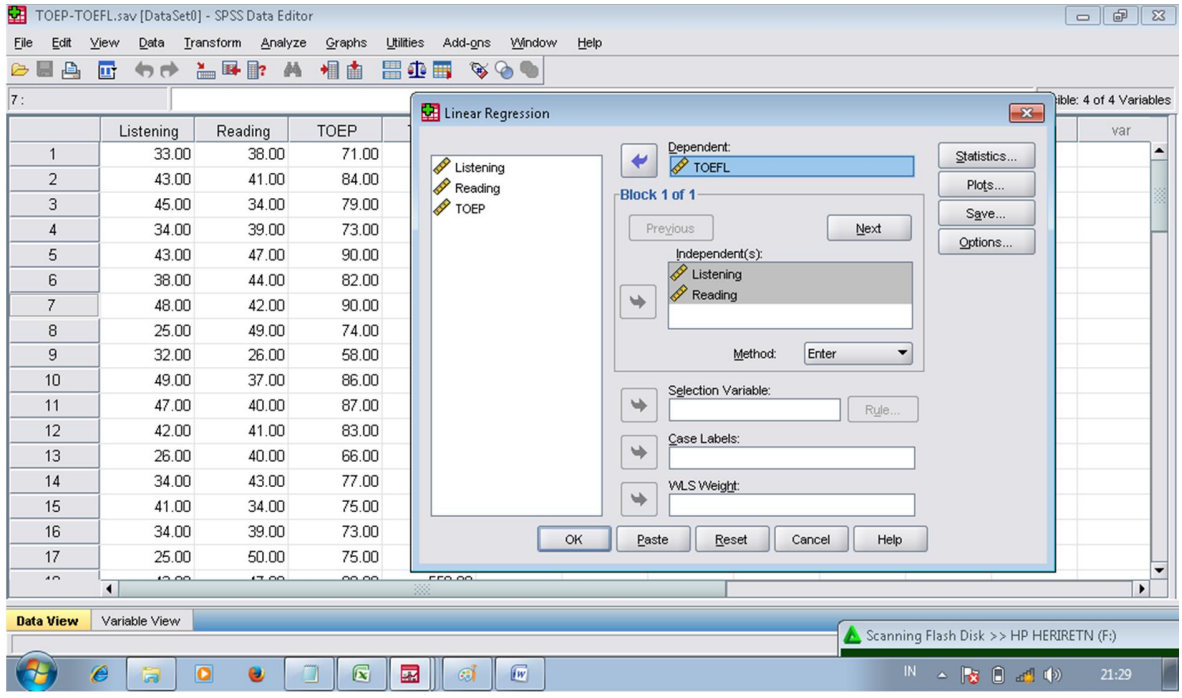
Selanjutnya akan diperoleh scater-plot 3 dimensi. Dengan melihat grafik ini, dapat diketahui adanya pola linear dari variabel prediktor ke variabel kriteria.



Untuk memeroleh persamaan prediksi, dapat dilakukan analisis regresi. Langkah-langkahnya adalah dengan mengeklik **Analyze, Regression, Linear**.



Kemudian dimasukkan skor yang diperoleh pada kriteria ke **Dependent**, kemudian skor dari perangkat tes yang akan diketahui validitas prediktifnya atau variabel prediktor dimasukkan pada **Independent(s)**. Selanjutnya dipilih **OK**.



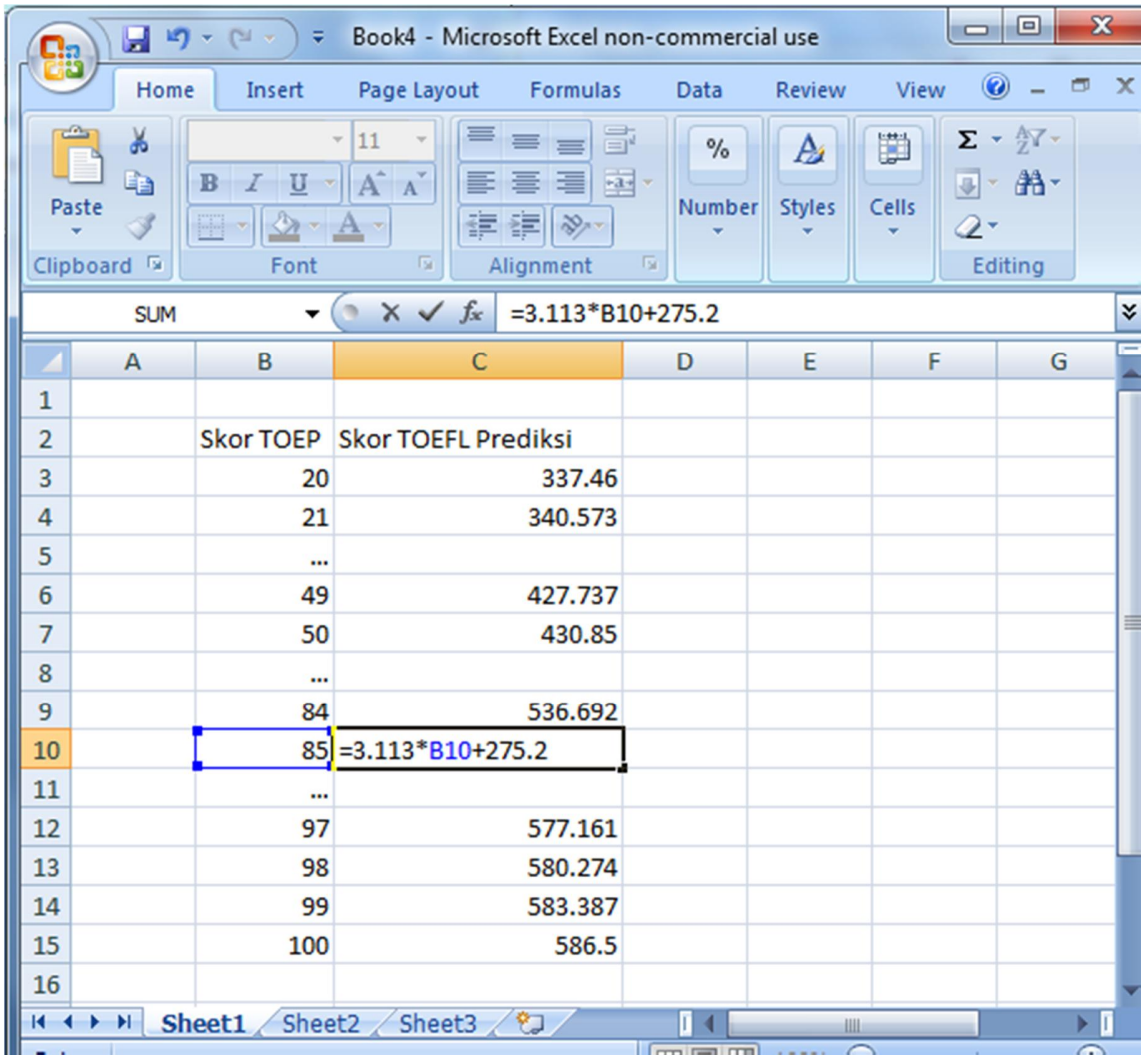
Output yang diperoleh berupa koefisien untuk persamaan prediksi. Pada kasus ini diperoleh TOEFL = 278,494 + 3,523 TOEP Listening + 2,637 TOEP Reading.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	278.494	12.440		22.387	.000
	Listening	3.523	.280	.683	12.562	.000
	Reading	2.637	.308	.465	8.555	.000

a. Dependent Variable: TOEFL

Dengan menggunakan persamaan prediksi, selanjutnya dapat disusun tabel skor prediksi dengan menggunakan skor prediktor. Menyusun tabel ini dapat menggunakan bantuan program Excel.



Demikian pula halnya jika menggunakan 2 variabel prediktor.

Bab VI
RELIABILITAS

A. Estimasi Reliabilitas

Pada suatu instrumen yang digunakan untuk mengumpulkan data, reliabilitas skor hasil tes merupakan informasi yang diperlukan dalam pengembangan tes. Reliabilitas merupakan derajat keajegan *(consistency)* di antara dua skor hasil pengukuran pada objek yang sama, meskipun menggunakan alat pengukur yang berbeda dan skala yang berbeda (Mehrens & Lehmann,1973; Reynold, Livingstone, & Wilson, 2010). Dalam kaitannya dengan penilaian pendidikan, prestasi atau kemampuan seorang siswa dikatakan reliabel jika dilakukan pengukuran, hasil pengukuran akan sama informasinya, walaupun penguji berbeda, korektornya berbeda atau butir soal yang berbeda tetapi memiliki karakteristik yang sama.

Allen & Yen (1979) menyatakan bahwa tes dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Selanjutnya dinyatakan bahwa reliabilitas merupakan koefisien korelasi antara dua skor amatan yang diperoleh dari hasil pengukuran menggunakan tes yang paralel. Dengan demikian, pengertian yang dapat diperoleh dari pernyatan tersebut adalah suatu tes itu reliabel jika hasil pengukuran mendekati keadaan peserta tes yang sebenarnya.

Dalam pendidikan, pengukuran tidak dapat langsung dilakukan pada ciri atau karakter yang akan diukur. Ciri atau karakter ini bersifat abstrak, yang dapat diukur melalui suatu indikator. Hal ini menyebabkan sulitnya memperoleh alat ukur yang stabil untuk mengukur karakteristik seseorang (Mehrens & Lehmann, 1973). Kestabilan ini yang dikatakan sebagai reliabilitas. Untuk melihat reliabilitas suatu alat ukur, yang berupa suatu nilai, dapat dilakukan perhitungan statistik. Nilai ini biasa dinamakan dengan koefisien reliabilitas *(reliability coefficient*).

Koefisien reliabilitas dapat diartikan sebagai koefisien keajegan atau kestabilan hasil pengukuran. Alat ukur yang reliabel akan memberikan hasil pengukuran yang stabil (Lawrence, 1994) dan konsisten (Mehrens & Lehmann, 1973). Artinya suatu alat ukur dikatakan memiliki koefisien reliabilitas tinggi manakala digunakan untuk mengukur hal yang sama pada waktu berbeda hasilnya sama atau mendekati sama. Dalam hal ini,

reliabilitas merupakan sifat dari sekumpulan skor (Frisbie, 2005). Dalam kaitannya dengan dunia pendidikan, dengan alat ukur yang reliabel, hasil pengukuran akan sama informasinya walaupun penguji berbeda, korektornya berbeda atau butir soal yang berbeda tetapi mengukur hal yang sama dan memiliki karakteristik butir yang sama.

Allen & Yen (1979) menyatakan bahwa tes dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Selanjutnya dinyatakan bahwa koefisien reliabilitas merupakan koefisien korelasi antara dua skor amatan yang diperoleh dari hasil pengukuran menggunakan tes yang paralel. Dengan demikian, pengertian yang dapat diperoleh dari pernyatan tersebut adalah suatu tes itu reliabel jika hasil pengukuran mendekati keadaan peserta tes yang sebenarnya.

Reliabilitas (ρ) suatu tes pada umumnya diekspresikan secara numerik dalam bentuk koefisien yang besarnya $-1{,}00 \leq \rho \leq +1{,}00$. Koefisien tinggi menunjukkan reliabilitas tinggi. Sebaliknya, jika koefisien suatu skor tes rendah maka reliabilitas tes rendah. Jika suatu reliabilitas sempurna, berarti koefisien reliabilitas tersebut +1,00. Harapannya, koefisien reliabilitas bersifat positif.

Reliabilitas terkait pula dengan kesalahan pengukuran. Reliabilitas tinggi menunjukkan kesalahan yang kecil dalam memeroleh hasil pengukuran. Semakin besar reabilitas suatu instrumen, akan semakin kecil kesalahan pengukuran, demikian pula sebaliknya, semakin kecil reliabilitas skor, akan semakin besar hasil pengukurannya. Kesalahan pengukuran dapat disebabkan oleh beberapa faktor, diantaranya karakteristik instrumen yang digunakan sendiri, misalnya penyusunan dan pelaksanaan pengukuran yang tidak mengikuti aturan baku, kualitas butir dalam instrumen tidak baik, adanya kerjasama selama melaksanakan tes atau mengisi instrumen, butir-butir instrumen yang meragukan, keadaan peserta selama merespons instrumen, seperti peserta yang sedang lelah baik fisik maupun psikis, mempunyai problem pribadi, peserta yang mempunyai motivasi kurang, lingkungan tempat penyelenggaraan pengukuran yang kurang mendukung atau kombinasi dari segala permasalahan tersebut.

Mehrens & Lehmann (1973) menyatakan bahwa meskipun tidak ada perjanjian secara umum, tetapi secara luas dapat diterima bahwa untuk tes yang digunakan untuk membuat keputusan pada siswa secara perorangan harus memiliki koefisien reliabilitas

minimal sebesar 0,85. Dengan demikian, pada penelitian ini, tes seleksi digunakan untuk menentukan keputusan pada siswa secara perorangan, sehingga indeks koefisien reliabilitasnya diharapkan minimal sebesar 0,85.

Proses penghitungan reliabilitas disebut dengan estimasi. Estimasi reliabilitas tes yang dapat dilakukan dengan beberapa cara, konsistensi eksternal, konsistensi internal, reliabilitas komposit, reliabilitas konstruk, reliabilitas interrater, dan estimasi reliabilitas dengan teori generalisabilitas (*Generalizability theory*).

B. Estimasi Konsistensi Eksternal

Estimasi reliabilitas eksternal diperoleh dengan menggunakan skor hasil pengukuran yang berbeda, baik dari instrumen yang berbeda maupun yang sama. Ada dua cara untuk mengestimasi reliabilitas eksternal suatu instrumen yaitu dengan teknik pengukuran ulang (*test-retest-method*) dan teknik paralel.

1. Metode Tes Ulang (*Test-Retest-Method*)

Untuk mengetahui keterandalan atau reliabilitas skor hasil pengukuran, pengukuran perlu dilakukan dua kali, pengukuran pertama dan pengukuran kedua atau ulanganya. Kedua pengukuran ini dapat dilakukan oleh orang yang sama atau berbeda, namun pada proses pengukuran yang kedua, keadaan yang diukur itu harus benar-benar berada pada kondisi yang sama dengan pengukuran pertama. Selanjutnya hasil pengukuran yang pertama dan yang kedua dikorelasikan dan hasilnya menunjukkan reliabilitas skor perangkat pengukuran.

Teknik ***Test-Retest-Method*** ini akan dapat sesuai dengan tujuannya jika keadaan subjek yang diukur tetap dan tidak mengalami perubahan pada saat pengukuran yang pertama maupun pada pengukuran yang kedua. Pada dasarnya keadaan responden itu selalu berkembang, tidak statis ataupun berubah-ubah, maka sebenarnya teknik ini kurang tepat digunakan. Di samping itu pada pengukuran yang kedua akan terjadi adanya *carry-over-effect* atau *testing effect*, responden pengukuran atau penelitian telah mendapat tambahan pengetahuan karena sudah mengalami tes yang pertama ataupun belajar setelah pengukuran yang pertama.

Ada beberapa hal yang harus dipertimbangkan dalam mengestimasi koefisisen dengan teknik tes-retes ini. Jangka waktu antara kedua pengukuran dengan menggunakan instrumen tersebut perlu menjadi pertimbangan. Jika jarak pengukuran terlalu dekat,maka *carry-over-effect* masih ada. Sementara jika jarak pengukuran terlalu jauh, korelasi kedua skor akan menjadi semakin rendah. Faktor kedua yang menjadi pertimbangan adalah stabilitas yang diharapkan dari kinerja yang diukur dengan instrumen tersebut. Semakin lama interval pelaksanaan pengukuran kedua instrumen, akan semakin rendah koefisien reliabilitasnya. Untuk mengatasi hal ini, jarak kedua pengukuran sebaiknya tidak terlalu jauh, misalnya tidak sampai satu bulan.

Estimasi reliabilitas dengan teknik tes-retes akan menghasilkan koefisien stabilitas. Untuk memperoleh koefisien reliabilitas melalui pendekatan tes-retes dapat dilakukan dengan menghitung koefisien korelasi linier antara skor pada pengukuran pertama (X) dengan skor hasil pengukuran kedua (Y).

$$r_i = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}} \qquad (6.1)$$

2. Metode Bentuk Paralel (*Equivalent*)

Teknik kedua untuk mengestimasi reliabilitas secara eksternal adalah dengan metode bentuk paralel. Pada teknik ini, diperlukan dua instrumen yang dikatakan paralel untuk mengestimasi koefisien reliabilitas. Dua buah tes dikatakan paralel atau *equivalent* adalah dua buah instrumen yang mempunyai kesamaan tujuan dalam pengukuran, tingkat kesukaran dan susunan juga sama, namun butir-butir soalnya berbeda, atau dikenal dengan istilah *alternate-forms method* atau *parallel forms*.

Dengan metode bentuk paralel ini, dua buah instrumen yang paralel, misalnya instrumen paket A yang akan diestimasi reliabilitasnya dan instrumen paket B merupakan instrumen yang paralel dengan paket A, keduanya diberikan kepada sekelompok responden yang sama, kemudian kedua skor tersebut dikorelasikan. Koefisien korelasi dari kedua skor respon responden terhadap instrumen inilah yang menunjukkan koefisien reliabilitas skor instrumen paket A. Jika koefisiennya reliabilitas skor instrumen tinggi, maka perangkat tersebut sudah dapat dikatakan reliabel dan dapat digunakan sebagai instrumen pengukur suatu konstruk yang terandalkan.

Dari sisi responden, estimasi reliabilitas dengan teknik ini ada kelemahannya. Dalam menggunakan teknik ini, diperlukan dua buah instrumen, dan masing-masing diberikan kepada sekelompok responden yang sama. Teknik ini responden tidak mengalami *practice-effect* dan *carry-over-effect* atau responden tidak mengingat pengerjaan instrumen sebelumnya.

Estimasi reliabilitas dengan cara ini merupakan pekerjaan yang cukup berat. Pada cara ini, diperlukan dua instrumen untuk digunakan, sehingga harus mengembangkan 2 instrumen dan juga mengujicobakan 2 instrumen. Membuktikan kedua instrumen tersebut merupakan tes yang paralel atau ekivalen memerlukan ilmu yang tersendiri (konsep penyetaraan tes atau *equating*).

Langkah-langkah yang ditempuh pada pembuktian reliabilitas dengan cara ini adalah: (1) menyiapkan dua instrumen yang paralel, (2) menentukan subjek untuk mengujicobakan instrumen, (3) melaksanakan pengukuran dengan mengadministrasikan instrumen tersebut, (4) melakukan penyekoran pada setiap jawaban responden terhadap kedua perangkat tersebut, (5) menghitung koefisien korelasi dari skor kedua perangkat tersebut.

Hasil koefisien korelasi yang tinggi dari skor jawaban responden pada kedua instrumen yang digunakan menunjukkan bahwa reliabilitas paralel dari perangkat tersebut berada pada kategori yang baik. Namun sebaliknya, jika ternyata koefesien korelasinya rendah, maka reliabilitas skor perangkat ekivalen adalah rendah.

C. Konsistensi Internal

Dengan teknik konsistensi internal ini, hanya dengan melakukan satu kali pengumpulan data, reliabilitas skor perangkat pengukuran dapat diestimasi. Pada pembuktian instrumen dengan cara ini ada beberapa cara, yang masing-masing dapat memerlukan persyaratan-persyaratan atau asumsi tertentu yang harus dipenuhi oleh peneliti. Beberapa cara yang dapat digunakan untuk mengestimasi reliabilitas dengan konsistensi internal diantaranya sebagai berikut.

1. Metode Belah Dua (*Split Half Method*)

Dalam teknik belah dua ini, dalam satu instrumen dikerjakan satu kali oleh sejumlah subjek (*sample*) suatu penelitian. Butir-butir pada perangkat dibagi menjadi dua.Pembagian dapat menggunakan nomor ganjil-genap pada instrumen, atau separuh pertama maupun separuh kedua, maupun membelah dengan menggunakan nomor acak atau tanpa pola tertentu. Skor responden merespons setengah perangkat bagian yang pertama dikorelasikan dengan skor setengah perangkat pada bagian yang kedua. Teknik ini berpegang pada asumsi, belahan pertama dan belahan kedua mengukur konstruk yang sama, banyaknya butir dalam instrumen belahan pertama dan kedua harus dapat dibandingkan dari sisi banyaknya butir, atau paling tidak jumlahnya hampir sama.

Ada beberapa formula untuk mengestimasi reliabilitas dengan metode belah dua, antara lain rumus Spearman-Brown, rumus Flanagan, dan rumus Rulon. Masing formula disjikan berikut ini.

a. Reliabilitas dengan Rumus Spearman-Brown

Adapun rumus Spearman-Brown yang digunakan adalah :

$$r_i = \frac{2r_b}{1+r_b} \qquad (6.2)$$

Dengan $r_b = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}}$

Dengan r_i = koefisien reliabilitas skor instrumen; r_b = koefisien korelasi antara dua belahan instrumen, N = banyaknya responden, X = belahan pertama, Y = belahan kedua.

b. Reliabilitas dengan Rumus Flanagan

Untuk mengestimasi reliabilitas dengan rumus Rulon, peneliti perlu meghitung kovarians dari skor belahan pertama dan skor belahan kedua (σ_{12}) dan varians totalnya. Koefisien reliabilitas disajikan dengan formula sebagai berikut.

$$r_i = \frac{2\sigma_{12}}{\sigma_X^2} = \frac{4r_{12}\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2+2r_{12}\sigma_{12}} \text{ (Walker, 2006)} \qquad (6.3)$$

Dengan $\boldsymbol{r_i}$ = koefisien reliabilitas skor instrumen; $\boldsymbol{r_{12}}$ = koefisien korelasi antara dua belahan instrumen, $\boldsymbol{\sigma_1^2}$ = varians belahan pertama, $\boldsymbol{\sigma_2^2}$ = varians belahan kedua, $\boldsymbol{\sigma_X^2}$ = varians skor total.

c. Reliabilitas dengan Rumus Rulon

Rulon merumuskan suatu formula untuk mengestimasi reliabilitas belah dua tanpa perlu berasumsi bahwa kedua belahan mempunyai varians yang sama. Menurut Rulon, perbedaan skor subjek pada kedua belahan instrumen akan membentuk distribusi perbedaan skor dengan varians yang besarnya ditentukan oleh varians *error* masing-masing belahan menentukan varians *error* keseluruhan instrumen, maka varians eror istrumen ini dapat diestimasi lewat besarnya varians perbedaan skor diantara kedua belahan. Dalam melakukan estimasi reliabilitas skor instrumen, varians perbedaan skor diperhitungkan sebagai sumber *error*. Untuk melakukan estimasi reliabilitas instrumen dengan rumus Rulon, peneliti juga harus menghitung dahulu varians selisih belahan pertama dan kedua dan juga varians total. Formula Rulon untuk mengestimasi reliabilitas sebagai berikut.

$$r_i = 1 - \frac{\sigma^2 d}{\sigma^2 t} \qquad (6.4)$$

Dengan $\boldsymbol{r_i}$ = reliabilitas instrumen; $\boldsymbol{\sigma^2{}_t}$ = varians total atau varians skor total; $\boldsymbol{\sigma^2{}_d}$ = varians dari perbedaan skor kedua belahan (*varians difference*); d = skor pada belahan awal dikurangi skor pada belahan akhir.

D. Reliabilitas Komposit

Pada suatu instrumen, sering peneliti menggunakan instrumen yang terdiri dari banyak butir. Jika butir-butir ini merupakan butir yang berbeda-beda namun membangun suatu konstruk yang sama, maka analisis untuk mengestimasi reliabilitas dapat digunakan rumus reliabilitas komposit. Komposit yang dimaksudkan yakni skor akhir merupakan gabungan dari skor butir-butir penyusun instrumen. Ada 3 formula yang dapat digunakan untuk mengestimasi reliabilitas dengan cara ini, yaitu dengan menghitung koefisien α dari Cronbach, koefisien KR-20, dan koefisien KR-21.

1. Rumus Alpha dari Cronbanch

Rumus Alpha digunakan untuk mengestimasi reliabilitas instrumen yang skornya bukan hanya1 dan 0, namun juga skala politomus, misal misalnya angket (skala Likert 1-2-3-4-5) atau soal bentuk uraian (skor maksimum dapat tergantung peneliti). Rumus Alpha sebagai berikut.

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum {\sigma_i}^2}{{\sigma_t}^2}\right) \qquad (6.5)$$

Dengan α = koefisien reliabilitas instrumen; k = banyaknya butir pertanyaan dalam instrumen; $\sum {\sigma_i}^2$ = jumlah varians butir instrumen; ${\sigma_t}^2$= varians skor total.

2. Rumus Kuder-Richardson

Cara lain mengestimasi reliabilitas dengan reliabilitas komposit adalah dengan menggunakan formula Kuder dan Richardson yang disingkat dengan formula KR. Ada dua jenis formula KR, yaitu Kuder Richardson formula 20 (KR-20) dan Kuder Richardson formula 21 (KR-21).

Formula KR-20 dapat digunakan untuk analisis butir dikotomi. Pada butir instrumen dengan penskoran dikotomi, misal1-0, benar-salah, ya-tidak, hidup-mati, dan lain-lain, estimasi reliabilitas dapat dilakukan dengan menggunakan rumus KR-20.

Rumus KR-20 sebagai berikut:

$$r_{ii} = \frac{k}{(k-1)}\left\{\frac{s_t^2 - \sum p_i q_i}{s_t^2}\right\} \qquad (6.6)$$

Dengan r_{ii}= reliabilitas skor instrumen; k=banyaknya butir pertanyaan atau banyaknya soal; s_t^2= varians skor total; p_i= proporsi subjek yang menjawab betul pada suatu butir (proporsi subjek yang mendapat skor 1) yang dihitung dengan

$$p_i = \frac{\text{banyaknya subjek yang skornya 1}}{\text{N}}; \text{ dan } q_i = 1 - p_i$$

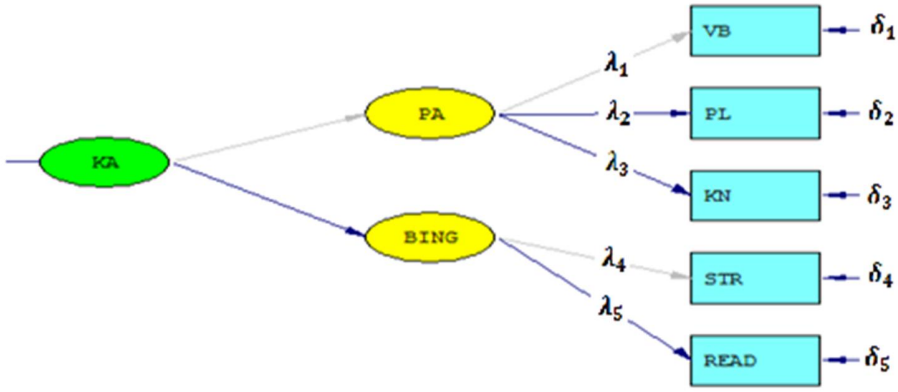
Rumus KR-21dapat digunakan untuk instrumen dengan skornya tiap butirnya 1 dan 0, dan juga skala politomus, misal misalnya angket (skala Likert 1-2-3-4-5) atau soal bentuk uraian. Formula KR-21 sebagai berikut.

$$r_{ii} = \frac{k}{k-1}\left(1 - \frac{\bar{X}(k-\bar{X})}{k\sigma_t^2}\right) \qquad (6.7)$$

Dengan r_{ii}= koefisien reliabilitas skor instrumen; k= banyaknya butir pertanyaan atau banyaknya soal; σ_t^2= varians total; $\bar{X}$= skor rata-rata (Allen & Yen, 1979).

E. Reliabilitas Konstruk

Cara keempat untuk mengestimasi reliabilitas adalah dengan mengestimasi reliabilitas konstruk. Reliabilitas konstruk ini dapat diestimasi setelah peneliti membuktikan validitas konstruk dengan analisis faktor konfirmatori sampai memeroleh model yang cocok (model yang fit). Dengan analisis faktor ini, peneliti dapat memeroleh muatan faktor (*factor loading*) tiap indikator yang menyusun instrumen (λ) dan indeks kesalahan unik dari tiap indikator (δ). Sebagai contoh pada pembuktian validitas konstruk kemampuan akademik, diperoleh model yang fit yang disajikan pada Gambar 6.1.



Gambar 6.1. Hasil Analisis Faktor Konfirmatori

Estimasi reliabilitas dapat dilakukan dengan tiga cara, estimasi reliabilitas dengan reliabilitas konstruk (*construct reliability, CR*), reliabilitas ω, dan reliabilitas maksimal (Ω).

Estimasi CR menggunakan muatan faktor (*factor loading*) tiap indikator yang menyusun instrumen (λ) dan indeks kesalahan unik dari tiap indikator (δ). Formulanya sebagai berikut (Geldhof, Preacher, Zyphur, 2014).

$$CR = \frac{\left(\sum_{i=1}^{i} \lambda_i\right)^2}{\left(\sum_{i=1}^{i} \lambda_i\right)^2 + \left(\sum_{i=1}^{i} \delta\right)} \qquad (6.8)$$

Estimasi dengan reliabilitas ω dilakukan hanya dengan menggunakan muatan faktor (λ) saja. Formula untuk estimasi reliabilitas ω sebagai berikut (Kamata, Turhan, Darandari, 2003).

$$\omega = \frac{\left(\sum_{i=1}^{i} \lambda_i\right)^2}{\left(\sum_{i=1}^{i} \lambda_i\right)^2 + \left(\sum_{i=1}^{i} 1 - \lambda_i^2\right)} \qquad (6.9)$$

Pada estimasi reliabilitas maksimal, muatan faktor disimbulkan dengan ℓ. Formulanya sebagai berikut (Penev & Raykov, 2006).

$$\Omega_w = \frac{\sum_{i=1}^{p} \frac{l_i^2}{(1 - l_i^2)}}{1 + \sum_{i=1}^{p} \frac{l_i^2}{(1 - l_i^2)}} \qquad (6.10)$$

F. Reliabilitas Inter-rater

Jika dalam suatu instrumen penskoran butir dilakukan dengan memanfaatkan dua orang rater, peneliti dapat mengestimasi reliabilitas dengan *inter-rater agreement*. Hasil estimasi reliabilitas dengan cara ini disebut dengan reliabilitas *inter-rater*. Adapun cara mengestimasinya dengan menghitung terlebih dahulu banyaknya butir atau kasus yang cocok atau butir atau kasus yang diskor sama oleh kedua rater. Banyaknya butir yang cocok ini kemudian dibandingkan dengan butir total, kemudian disajikan dalam

persentase. Estimasi reliabilitas skor dengan inter-rater dapat disajikan dengan formula sebagai berikut.

$$inter\text{-}rater\ agreement = \frac{\textbf{Banyaknya kasus yang diskor sama oleh kedua rater}}{\textbf{Banyaknya kasus}} \boldsymbol{x}\ \mathbf{100} \qquad (6.11)$$

Cara tersebut mudah dilakukan untuk penyekoran dengan skala yang mudah, misalnya 1-5 saja, itupun hasil penskoran berupa bilangan bulat. Namun jika hasil penskoran berada pada rentang yang panjang, misalnya 1-100, rumus tersebut akan menghasilkan koefisien kesepakatan interater yang kecil. Cara lain yang dapat dilakukan adalah dengan membuat urutan skor setiap rater dari yang rendah ke yang tinggi. Persentase banyaknya kasus yang sama peringkatnya dari kedua rater merupakan koefisien kesepakatan rater. Formulanya sebagai berikut.

$$inter\text{-}rater\ agreement = \frac{\textbf{Banyaknya kasus yang peringkatnya sama oleh kedua rater}}{\textbf{Banyaknya kasus}} \boldsymbol{x}\ \mathbf{100} \qquad (6.12)$$

G. Reliabilitas dengan Teori Generalizabilitas

Teori generalisabilitas *(Generalizability Theory)* terkait dengan 2 hal, generalizability (G) study dan decision (D) study. Peneliti yang melakukan *G-Study* mengutamakan generalisasi dari suatu sampel pengukuran ke keseluruhan pengukuran. Studi tentang stabilitas respons antarwaktu, equivalensi skor dari 2 atau lebih instrumen yang berbeda, hubungan antara skor sub-kemampuan dengan skor butir terkait dengan G-study. Pada *D-study*, data dikumpulkan untuk tujuan khusus terkait dengan membuat keputusan. Studi ini menyediakan data mendeskripsikan peserta tes, baik seleksi atau penempatan, maupun menyelidiki hubungan 2 variabel atau lebih (Crocker & Algina, 2008). Sebagai contoh, pda suatu tes seleksi, panitia akan menggunakan dua penilai atau lebih perlu diperiksa terlebih dahulu efisiensinya. Untuk hal tersebut, perlu dilakukan *D-study*.

Koefisien reliabilitas dalam teori ini disebut dengan koefisien *generalizability*. Dalam mengestimasi koefisien *generalizability*, ada beberapa desain, termasuk banyaknya bentuk tes, kesempatan melakukan tes atau administrasi tes, banyaknya rater,

yang sering disebut dengan facet. Banyaknya variabel yang digunakan menentukan bayaknya facet. Desain yang dapat dipilih misalnya desain facet tunggal (*single facet design*) dan facet ganda.

1. Desain Facet Tunggal

Desain faet tunggal terdiri dari 4 desain, yakni 1) setiap peserta tes atau jawaban peserta tes dinilai oleh satu rater, dan rater ini menilai semua peserta tes, 2) setiap peserta dinilai oleh beberapa rater, dan semua rater menilai peserta tes, 3) setiap peserta tes dinilai oleh rater yang berbeda, hanya satu rater untuk setiap peserta tes, dan 4) setiap penilaia dinilai oleh beberapa rater, ada rater yang berbeda-beda untuk setipa peserta tes (Crocker & Algina, 2008).

$$\boldsymbol{\rho_i^2 = \frac{\sigma_p^2}{\sigma_p^2+\sigma_e^2}} \qquad (6.13)$$

Pelaksanaan mengestimasi dilakukan dengan bantuan analisis varians (ANAVA). Misalkan MS adalah mean square pada ANAVA, dengan sumber variasi peserta tes (persons, P), dan rater (R).Rangkuman tabel ANAVA disajikan pada Tabel 6.1.

Tabel 6.1. Rangkuman Tabel Anava

SV	*SS*	*df*	*MS*	*EMS*
Examinee (*P*)	$\boldsymbol{n_i \sum_p (X_{pI} - X_{PI})^2}$	$\boldsymbol{n_p - 1}$	$\boldsymbol{\frac{SS_p}{n_p - 1}}$	$\boldsymbol{\sigma_e^2 + n_i\sigma_p^2}$
Raters (*I*)	$\boldsymbol{n_p \sum_i (X_{Pi} - X_{PI})^2}$	$\boldsymbol{n_i - 1}$	$\boldsymbol{\frac{SS_i}{n_i - 1}}$	$\boldsymbol{\sigma_e^2 + n_p\sigma_i^2}$
Residual (*R*)	$\boldsymbol{\sum_i \sum_j (X_{pi} - X_{PI})^2 - SS_p - SS_i}$	$\boldsymbol{(n_p - 1)(n_i - 1)}$	$\boldsymbol{\frac{SS_r}{(n_p - 1)(n_i - 1)}}$	$\boldsymbol{\sigma_e^2}$
	$\boldsymbol{X_{pI} = \sum_i \frac{X_{pi}}{n_i}}$	$\boldsymbol{X_{Pi} = \sum_p \frac{X_{pi}}{n_p}}$	$\boldsymbol{X_{PI} = \sum_i \sum_p \frac{X_{pi}}{n_i n_p}}$	

$\boldsymbol{EMS_p}$ dapat disubstitusikan dari

$\boldsymbol{\sigma_e^2 + n_p\sigma_i^2}$ dan EMS_r yaitu

$\boldsymbol{\sigma_e^2}$ untuk memeroleh

$\boldsymbol{n_i\sigma_p^2 = \left(EMS_p - EMS_r\right)}$ sehingga $\boldsymbol{\sigma_p^2 = \frac{(EMS_p - EMS_r)}{n_i}}$ dan juga

$$\boldsymbol{\hat{\sigma}_p^2 = \frac{(MS_p - MS_r)}{n_i}}$$

Sehingga diperoleh

$$\boldsymbol{\hat{\rho}_i^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2}}$$

$$\boldsymbol{\hat{\rho}_i^2 = \frac{MS_p - MS_r}{MS_p + (n_i - 1)MS_r}} \qquad (6.14)$$

yang merupakan formula koefisien generalizability untuk desain facet tunggal yang pertama.

Pada desain facet tunggal yang kedua, banyaknya rater ditingkatkan menjadi n_i' untuk mengetahui banyaknya rater yang sesuai pada D-study. Koefisien generalizability diestimasi dengan

$\boldsymbol{\hat{\rho}_{i*}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2/n_i'}}$ yang nilainya diestimasi dengan rumus

$$\boldsymbol{\hat{\rho}_{i*}^2 = \frac{MS_p - MS_r}{MS_p + (n_i - n_i')MS_r/n_i'}} \qquad (6.15)$$

Desain facet tunggal yang ketiga, setiap peserta tes dinilai oleh rater yang berbeda, hanya satu rater untuk setiap peserta tes.

$\boldsymbol{\hat{\rho}_i^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 + \hat{\sigma}_e^2}}$ yang nilainya diestimasi dengan rumus

$$\boldsymbol{\hat{\rho}_i^2 = \frac{MS_p - MS_r}{MS_p + n_iMS_i/n_p + (n_in_p - n_i - n_p)MS_r/n_p}} \qquad (6.16)$$

Desain facet tunggal yang keempat, setiap penilaian dinilai oleh beberapa rater, ada rater yang berbeda-beda untuk setiap peserta tes.

$$\widehat{\rho}_i^2 = \frac{\widehat{\sigma}_p^2}{\widehat{\sigma}_p^2+(\widehat{\sigma}_i^2+\widehat{\sigma}_e^2)/n_i'}$$

$$\widehat{\rho}_i^2 = \frac{MS_p - MS_r}{MS_p+n_iMS_i/n_pn_i'+(n_in_p-n_pn_i'-n_i)MS_e/n_pn_i'} \qquad (6.17)$$

2. Generalisabilitas untuk Desain Facet Ganda

Seperti pada ANAVA, pada teori ini dikenal dengan istilah tersilang (*crosses*) dan istilah tersarang (*nested*). Tersilang jika setiap kondisi pengukuran pada *facet* pertama terjadi dalam kombinasi dengan setiap pengukuran pada faktor yang kedua. Suatu facet dikatakan tersarang dalam *facet* kedua jika himpunan yang berbeda dari kondisi pengukuran pada *facet* pertama terjadi dalam kombinasi dengan setiap kondisi pengukuran pada *facet* yang kedua.

Untuk mengestimasi koefisien generalisabilitas (*generalizability coefficient*), ada beberapa varians skor yang digunakan. Pada desain dua *facet*, melibatkan varians peserta tes, varians kondisi *facet* I, varians kondisi *facet* J, varians interaksi, dan varians residual. Mengestimasi koefisien ini dapat dilakukan dengan menggunakan bantuan analisis varians (ANAVA) yang disajikan pada Tabel 6.2, dengan sumber varians *person* (p), butir (*question*, q), dan *rater* (r).

Tabel 6.2 Rangkuman ANAVA 3 jalur untuk generalizability 2 facet

Sumber Varians	Derajat Kebebasan	Sum Square (SS)	Mean Square (MS)	F	Sig
p (person)	p-1	SS_p	MS_p	.	.
q	q-1	SS_q	MS_q	.	.
R	r-1	SS_r	MS_r	.	.
pq	pq-1	SS_{pq}	MS_{pq}	.	.
pr	pr-1	SS_{pr}	MS_{pr}	.	.
qr	qr-1	SS_{qr}	MS_{qr}	.	.
pqr	pqr-1	SS_{pqr}	MS_{pqr}	.	.

Dari tabel rangkuman ANAVA tersebut, kolom MS yang digunakan untuk mengestimasi reliabilitas 2 *facet*. Adapun varians-varians yang digunakan menghitung reliabilitas yaitu

varians peserta tes, varians kondisi facet I (*question, q*), varians kondisi facet J (*rater, r*), varians interaksi, dan varians residual. Masing-masing disajikan sebagai berikut (Lord, 1973).

$$\boldsymbol{\sigma^2_{pqr} = MS_{pqr}}$$

$$\boldsymbol{\sigma^2_{pq} = \frac{1}{n_r}\left(MS_{pq} - MS_{pqr}\right)}$$

$$\boldsymbol{\sigma^2_{pr} = \frac{1}{n_q}\left(MS_{pr} - MS_{pqr}\right)}$$

$$\boldsymbol{\sigma^2_{qr} = \frac{1}{n_p}\left(MS_{qr} - MS_{pqr}\right)}$$

$$\boldsymbol{\sigma^2_{p} = \frac{1}{n_q n_r}\left(MS_{p} - MS_{pq} - MS_{pr} + MS_{pqr}\right)}$$

$$\boldsymbol{\sigma^2_{q} = \frac{1}{n_p n_r}\left(MS_{q} - MS_{pq} - MS_{qr} + MS_{pqr}\right)}$$

$$\boldsymbol{\sigma^2_{r} = \frac{1}{n_p n_q}\left(MS_{r} - MS_{pr} - MS_{qr} + MS_{pqr}\right)} \qquad (6.17)$$

Selanjutnya koefisien generalizability diestimasi dengan formula

$$r_{xx'} = \frac{\sigma^2_{true}}{\sigma^2_{obs}} \qquad (6.18)$$

Komponen varians skor murni dan varians skor amatan dimaksud dapat di-uraikan sebagai berikut:

$$\frac{\sigma^2_{true}}{\sigma^2_{obs}} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_q}{n_q} + \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_{pq}}{n_q} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{qr}}{n_q n_r} + \frac{\sigma^2_{pqr}}{n_q n_r}} \qquad (6.19)$$

H. Kesalahan Pengukuran Standar (Standard Error of Measurement, SEM)

Kesalahan Baku Pengukuran *(Standard Error of Measurement, SEM)* dapat digunakan untuk mamahami kesalahan yang bersifat acak/random yang mempengaruhi skor responden dalam merespons instrumen. Kesalahan pengukuran, yang disimbulkan dengan σ_E, dapat dihitung dengan rumus pada persaman 5, yang diturunkan dari rumus reliabilitas (Allen & Yen, 1979).

$$\sigma_E = \sigma_x \sqrt{1-\rho_{xx'}} \quad \text{.....(6.20)}$$

Penafsiran SEM dilakukan karena tidak adanya prosedur penilaian yang sangat konsisten, interpretasi skor dapat ditingkatkan dengan mempertimbangkan ukuran kemungkinan kesalahan pengukuran.

Interpretasi dari SEM digunakan untuk memprediksikan rentang skor sebenarnya (*true score*) yang diperoleh respoden. Skor sebenarnya (*true score*, τ) dari hasil penggukuran $X \pm SEM$, atau dengan simbol matematika sebagai berikut:

$$X - SEM < \tau < X + SEM \qquad \text{(6.210}$$

I. Faktor-Faktor yang mempengaruhi Reliabilitas

Ada beberapa faktor yang mempengaruhi reliabilitas. Faktor tersebut dapat dikategorikan faktor-faktor yang mempengaruhi secara langsung dan secara tidak langsung. Faktor-faktor tersebut yaitu :

1) Panjang tes dan kualitas butir-butir instrumen. Instrumen yang terdiri dari banyak butir, tentu lebih reliabel dibandingkan dengan instrumen yang hanya terdiri dari beberapa butir. Jika panjang instrumen ditambah dengan menambah butir-butir yang baik maka semakin panjang suatu instrumen maka reliabilitas skornya semakin tinggi. Namun jika instrumen terlalu panjang, maka responden akan terlalu lelah mengerjakannya. Faktor kelelahan responden ini yang akan menurunkan reliabilitas.
2) Kondisi penyelenggaraan pengumpulan data atau administrasi.
   a. Sebagai contoh pada pelaksanaan tes, petunjuk yang diberikan sebelum tes dimulai dan petunjuk ini disajikan dengan jelas, penyelenggaraan tes akan berjalan lancar dan tidak akan banyak terdapat pertanyaan atau komentar dari responden. Hal ini akan menjamin pelaksanaan tes yang tertib dan tenang sehingga skor yang diperoleh lebih reliabel.

b. Pengawas yang tertib akan mempengaruhi skor hasil perolehan responden. Pengawasan yang terlalu ketat ketika pengumpulan data menyebabkan responden merasa kurang nyaman atau merasa takut dan tidak dapat dengan leluasa dalam merespon instrumen, namun jika pengawasan kurang, maka peserta akan bekerjasama sehingga hasil pengumpulan data kurang dapat dipercaya.

c. Suasana lingkungan dan tempat pengumpulan data (tempat duduk yang tidak teratur, suasana disekelilingnya gaduh atau tidak tenang, dan sebagainya) akan mempengaruhi reliabilitas. Sebagai contoh pada pelaksanaan tes, suasana yang panas dan dekat sumber kegaduhan akan mempengaruhi hasil tes.

Adapun faktor-faktor yang mempengaruhi secara langsung hasil estimasi reliabilitas adalah

a. waktu penyelenggaraan pengumpulan data pertama dan kedua. Faktor ini terutama pada estimasi reliabiltas degan menggunakan teknik tes-retes. Interval penyelenggaraan yang terlalu dekat atau terlalu jauh, akan mempengaruhi koefisien reliabilitas.
b. Panjang instrumen, semakin panjang suatu instrumen pengumpul data, semakin banyak butir yang termuat di dalamnya. Hal ini akan memberikan dampak hasil pengumpulan data akan semakin mendekati keadaan yang sebenarnya, yang akan mempertinggi koefisien reliabilitas.

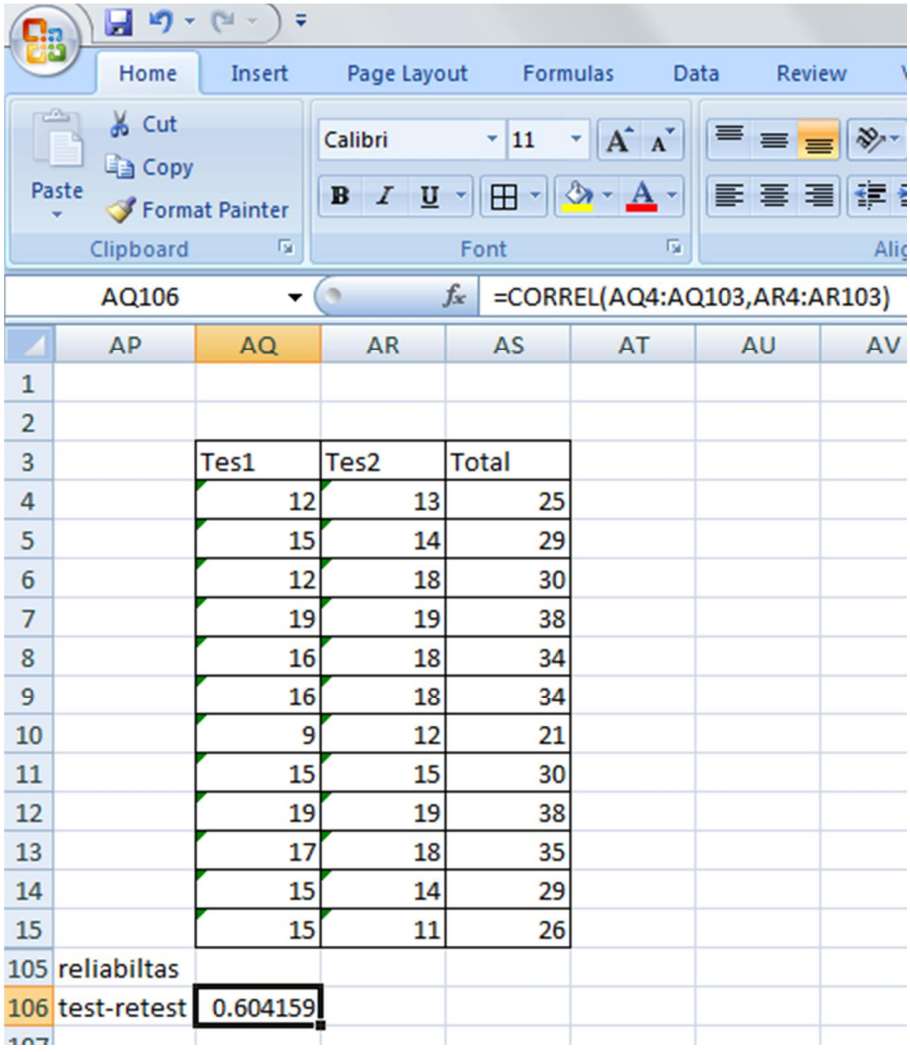
$$r_n = \frac{nr}{(n-1)r+1} \qquad (6.22)$$

c. Penyebaran skor perolehan responden. Koefisien reliabilitas secara langsung dipengaruhi oleh bentuk sebaran skor (variansi) dalam kelompok responden yang diukur. Semakin tinggi varians skor hasil pengukuran, semakin tinggi estimasi koefisien reliabilitas.
d. Tingkat kesulitan butir instrumen. Butir yang terlalu mudah dan butir terlalu sulit tidak memberikan tambahan variansi sebaran skor hasil pengukuran, sehingga akan mempengaruhi reliabilitas.
e. Objektivitas penskoran. Objektivitas penskoran terhadap respons responden terhadap instrumen akan mempengaruhi reliabilitas. Semakin objektif penskoran suat instrumen, maka skor perolehannya akan menjadi semakin reliabel.

J. Mengestimasi Reliabilitas

1. Tes-retes dan paralel

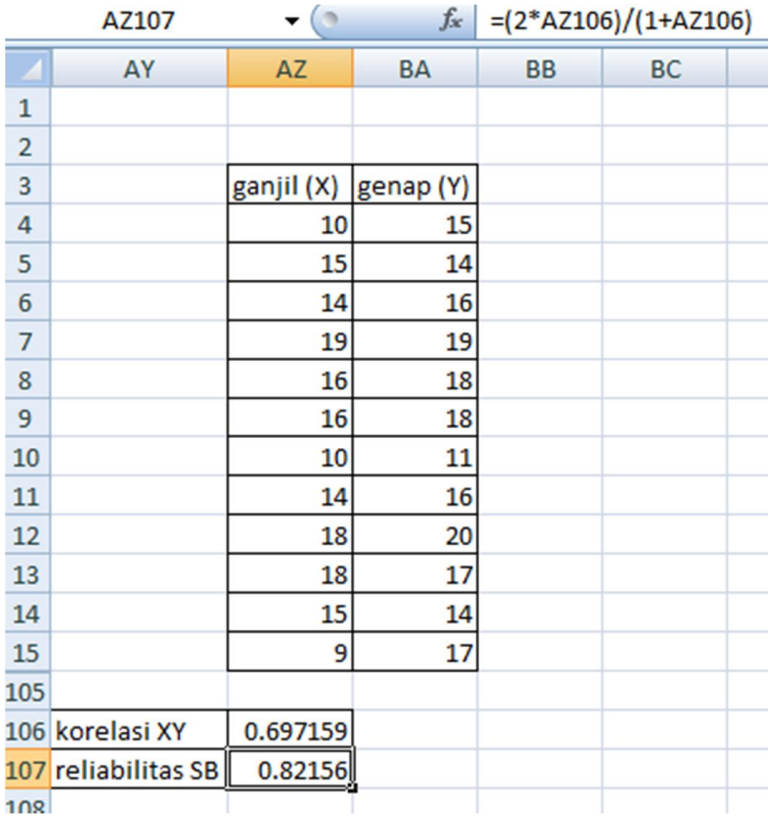
Untuk mengestimasi reliabilitas dengan formula ini, diperlukan 2 skor tes, baik untuk teknik tes-retes (tes ulang) maupun tes yang paralel. Skor kedua tes tersebut dikorelasikan dan hasil perhitungan korelasi tersebut merupakan koefisien reliabilitas tes ulang atau koefisien reliabilitas bentuk paralel. Teknik ini akan sangat mudah dilakukan dengan menggunakan program excel dengan menggunakan rumus korelasi **=CORREL**.



	AP	AQ	AR	AS	AT	AU	AV
1							
2							
3		Tes1	Tes2	Total			
4		12	13	25			
5		15	14	29			
6		12	18	30			
7		19	19	38			
8		16	18	34			
9		16	18	34			
10		9	12	21			
11		15	15	30			
12		19	19	38			
13		17	18	35			
14		15	14	29			
15		15	11	26			
105	reliabiltas						
106	test-retest	0.604159					

2. Belah dua

Pada estimasi dengan teknik belah dua ini, dipilih dulu skor belahan pertama dan skor belahan kedua, misalnya pada contoh dalam buku ini, skor belahan ganjil dan skor belahan genap. Setelah dikorelasikan keduanya dengan fungsi **(=CORREL(ARRAY1,ARRAI2)),** reliabilitas dengan rumus belah dua atau dikenal juga dengan rumus Spearmen-Brown dapat dihitung menggunakan rumus (6.2).

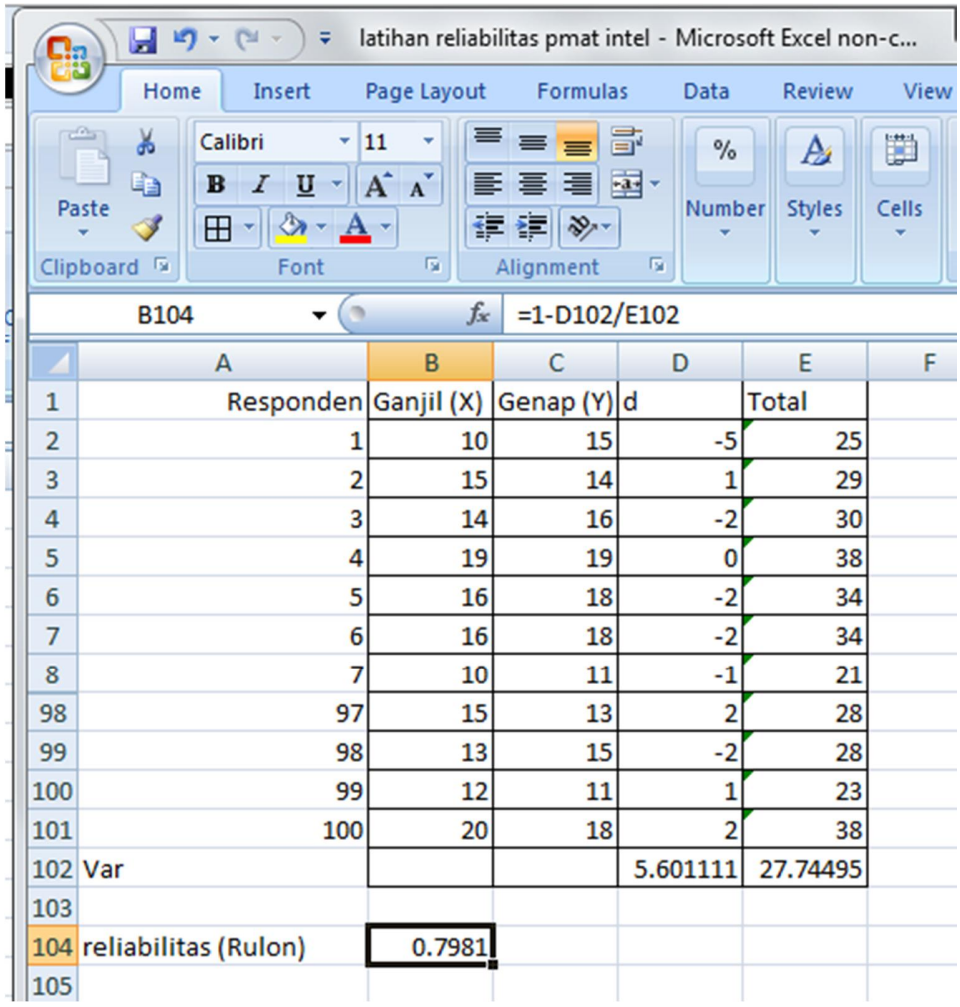


AZ107 | =(2*AZ106)/(1+AZ106)

	AY	AZ	BA	BB	BC
1					
2					
3		ganjil (X)	genap (Y)		
4		10	15		
5		15	14		
6		14	16		
7		19	19		
8		16	18		
9		16	18		
10		10	11		
11		14	16		
12		18	20		
13		18	17		
14		15	14		
15		9	17		
105					
106	korelasi XY	0.697159			
107	reliabilitas SB	0.82156			
108					

3. Rulon

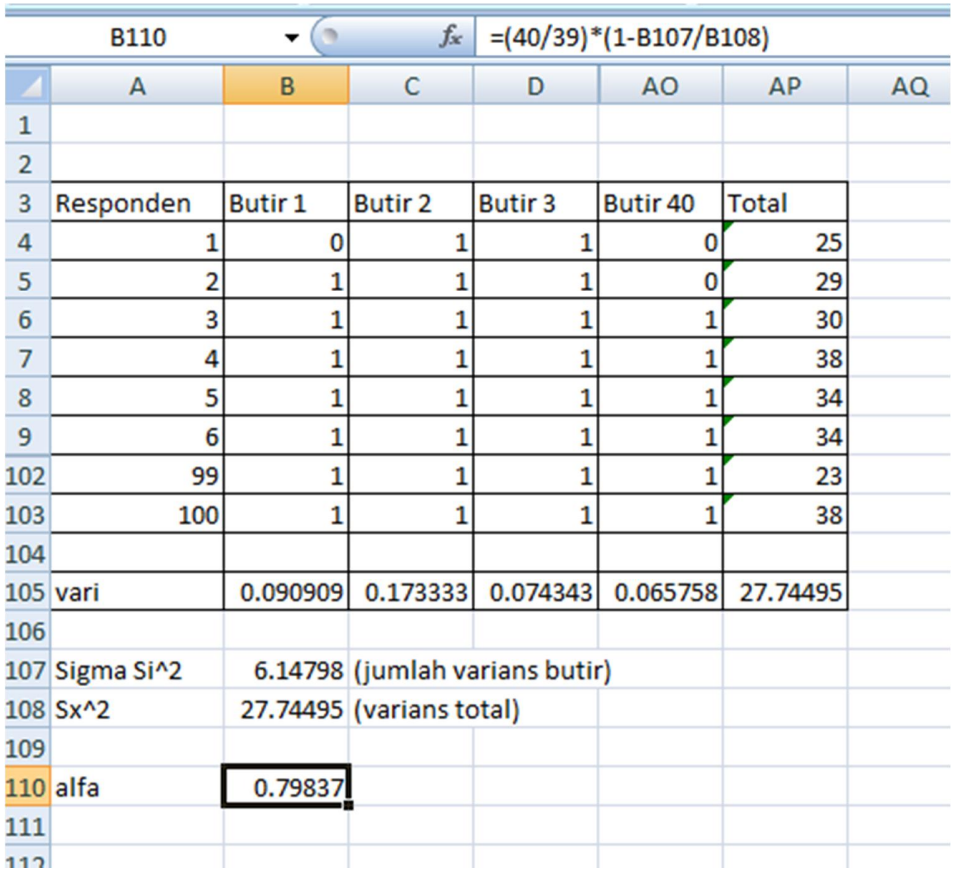
Pada estimasi reliabilitas dengan rumus Rulon, skor butir ganjil dan genap tetap dihitung, berikut nilai d_i = skor butir ganjil –skor butir genap (boleh sebaliknya asal konsisten). Variansi d_i keudian dihitung demikian pula variansi total (skor butir ganjil + skor butir genap). Hasil-hasil ini digunakan untuk mengestimasi reliabilitas dengan menggunakan rumus Rulon.



	A	B	C	D	E	F
1	Responden	Ganjil (X)	Genap (Y)	d	Total	
2	1	10	15	-5	25	
3	2	15	14	1	29	
4	3	14	16	-2	30	
5	4	19	19	0	38	
6	5	16	18	-2	34	
7	6	16	18	-2	34	
8	7	10	11	-1	21	
98	97	15	13	2	28	
99	98	13	15	-2	28	
100	99	12	11	1	23	
101	100	20	18	2	38	
102	Var			5.601111	27.74495	
103						
104	reliabilitas (Rulon)	0.7981				
105						

4. Alpha

Pada estimasi reliabilitas dengan rumus allfa dari Cronbach, diestimasi dulu varians tiap butir dan varians total. Koefisien reliabilitas selanjutnya dapat dihitung.



B110 | =(40/39)*(1-B107/B108)

	A	B	C	D	AO	AP	AQ
1							
2							
3	Responden	Butir 1	Butir 2	Butir 3	Butir 40	Total	
4	1	0	1	1	0	25	
5	2	1	1	1	0	29	
6	3	1	1	1	1	30	
7	4	1	1	1	1	38	
8	5	1	1	1	1	34	
9	6	1	1	1	1	34	
102	99	1	1	1	1	23	
103	100	1	1	1	1	38	
104							
105	vari	0.090909	0.173333	0.074343	0.065758	27.74495	
106							
107	Sigma Si^2	6.14798	(jumlah varians butir)				
108	Sx^2	27.74495	(varians total)				
109							
110	alfa	0.79837					
111							

5. KR-20

Pada estimasi reliabilitas dengan rumus KR-20, perlu dihitung terlebih dahulu proporsi menjawab benar tiap butir (p = banyaknya responden yang benar dibagi banyaknya siswa yang mengikuti tes), proporsi menjawab salah (q = 1–p), dan varians skor butir, pq (p dikali q). Selanjutnya dapat dihitung reliabilitas dengan KR20.

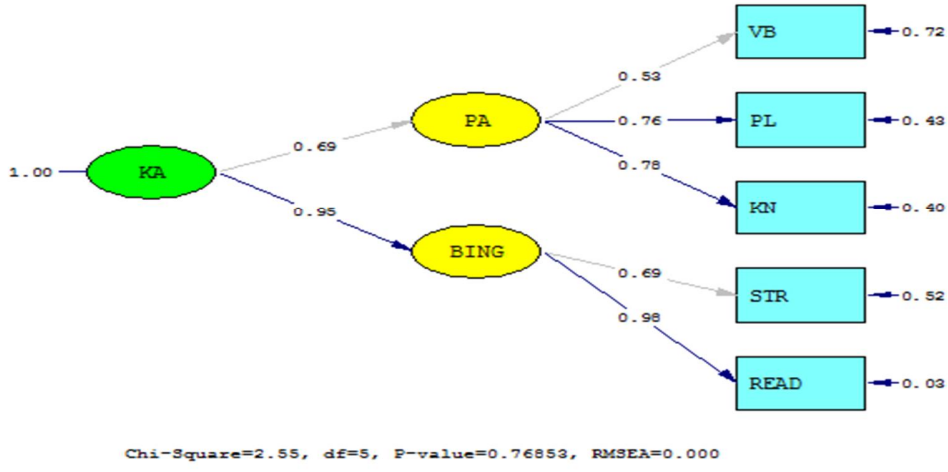
B112 =(40/39)*(1-(B109/27))

	A	B	C	D	E	AN	AO	AP
1								
2								
3	Responden	Butir 1	Butir 2	Butir 3	Butir 4	Butir 39	Butir 40	Total
4	1	0	1	1	0	1	0	25
5	2	1	1	1	0	1	0	29
6	3	1	1	1	0	1	1	30
7	4	1	1	1	1	1	1	38
8	5	1	1	1	0	1	1	34
9	6	1	1	1	0	1	1	34
10	7	1	1	0	0	1	1	21
102	99	1	1	1	0	1	1	23
103	100	1	1	1	0	1	1	38
104								
105	pi	0.9	0.78	0.92	0.42	0.72	0.93	27.74495
106	qi	0.1	0.22	0.08	0.58	0.28	0.07	
107	pi.qi	0.09	0.1716	0.0736	0.2436	0.2016	0.0651	
108								
109	Sigma pi.qi	6.0865	(jumlah pi.qi)					
110	Sx^2	27.74495	(varians total)					
111								
112	KR20	0.7944						
113								

6. Reliabiltas Konstruk

Reliabitas konstruk dapat dihitung setelah muatan faktor pada analisis faktor konfirmatori (λ) dan juga *error* dari tiap variable *observable* (δ). Dengan bantuan program Excel, menggunakan output dari analisis faktor konfirmatori dapat dihitung reliabilitas konstruk, reliabilitas ω, dan reliabilitas maksimal.

Hasil output analisis faktor konfirmatori:



Hasil analisis faktor konfirmatori tersebut kemudian dimasukkan dalam daftar di Excel untuk menghitung reliabilitas konstruk menggunakan rumus 6.8, 6.9, dan 6.10.

No.	λ	δ	λ^2	$1-\lambda^2$	$\lambda^2/(1-\lambda^2)$
1	0,53	0,72	0,2809	0,7191	0,390627
2	0,76	0,43	0,5776	0,4224	1,367424
3	0,78	0,4	0,6084	0,3916	1,553626
4	0,69	0,52	0,4761	0,5239	0,908761
5	0,98	0,03	0,9604	0,0396	24,25253
Jumlah	3,74	2,1	2,9034		28,47296
Reliabilitas	0,869465				
Konstruk					
		Reliabilitas	0,82811		
		Omega		Reliabitas	
				Maksimal	0,966071

7. Reliabilitas Inter-rater

Reliabilitas interrater dapat diestimasi dengan menghitung persentase kecocokan skor hasil penilaian oleh rater 1 dan rater 2. Teknik ini hanya dapat digunakan untuk melihat kecocokan 2 rater saja.

Responden	Rater 1	Rater 2
1	3	3
2	2	3
3	1	1
4	3	2
5	4	4
6	2	2
7	5	5
8	5	4
9	4	4
10	5	4

Responden	Rater 1	Rater 2
11	5	5
12	4	3
13	4	4
14	5	5
15	4	4
16	3	3
17	2	2
18	1	1
19	1	2
20	5	5

Dari skor penilaian rater 1 dan rater 2 dibuat daftar, kemudian dihitung skor-skor yang sama hasil penilaian 2 rater yang berbeda. Pada kasus tersebut, yang cocok ada 14 dari 20 butir, sehingga reliabilitas skornya 70% atau 0,7.

Inter-rater agreement= (14/20)x100% = 70%.

8. Koefisien Generalisabilitas

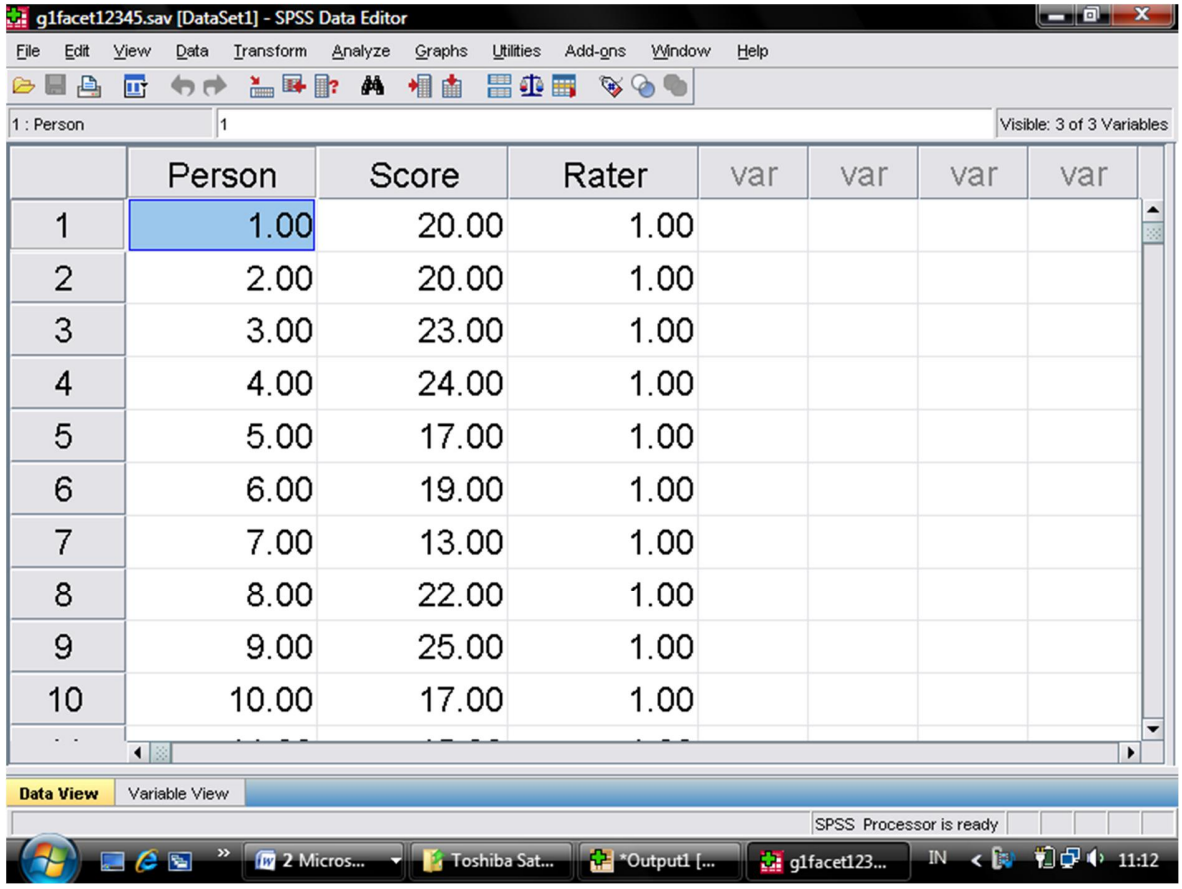
Estimasi reliabilitas dengan menggunakan koefisien generalisabilitas (*generalizability coefficient*) melibatkan analisis varians. Analisis varians ini dalam rangka mempermudah menghitung mean square (MS). Pada contoh ini, rangkuman analisis varians untuk memeroleh MS dilakukan dengan menggunakan SPSS. Sebelumnya, diinputkan datanya dengan variabel *person* (peserta tes), *score* (perolehan skor), dan *rater* (penilai).
Variabel untuk menginput data

g1facet12345.sav [DataSet1] - SPSS Data Editor

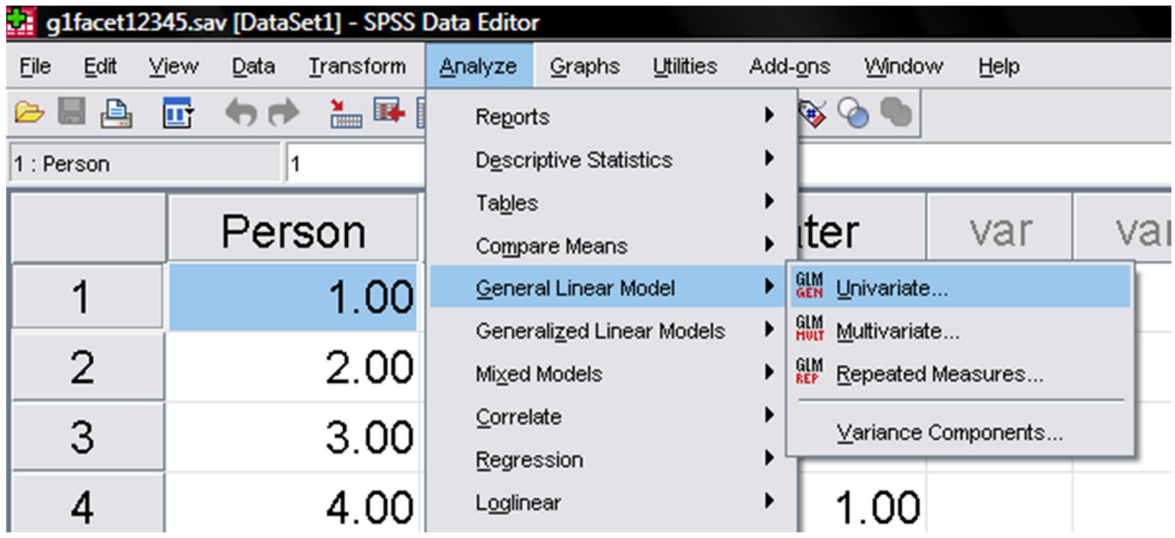
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals
1	Person	Numeric	8	2
2	Score	Numeric	8	2
3	Rater	Numeric	8	2

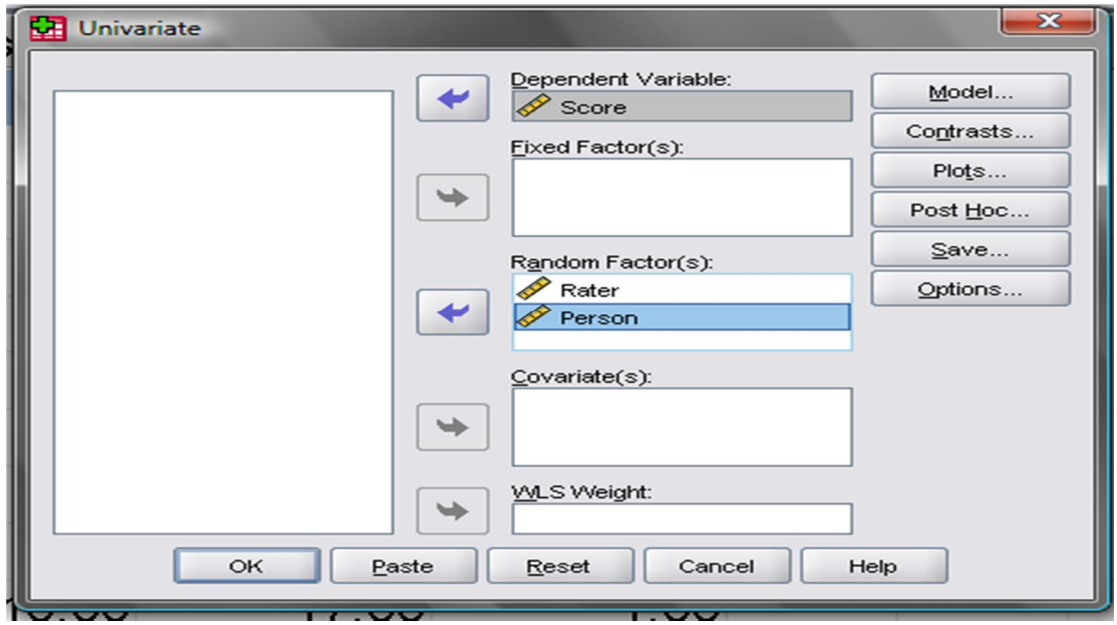
Hasil input data sebagai berikut.



Selanjutnya klik **Analyze→General Linear Model→Univariat.**



Masukkan **Score** pada variabel terikat (**Dependent Variable**) dan **Rater** dn **Person** sebagai variabel bebas (**Random Factor(s)**), kemudian klik **OK**.



Setelah itu akan muncul output rangkuman tabel analisis varians.

Tests of Between-Subjects Effects

Dependent Variable:Score

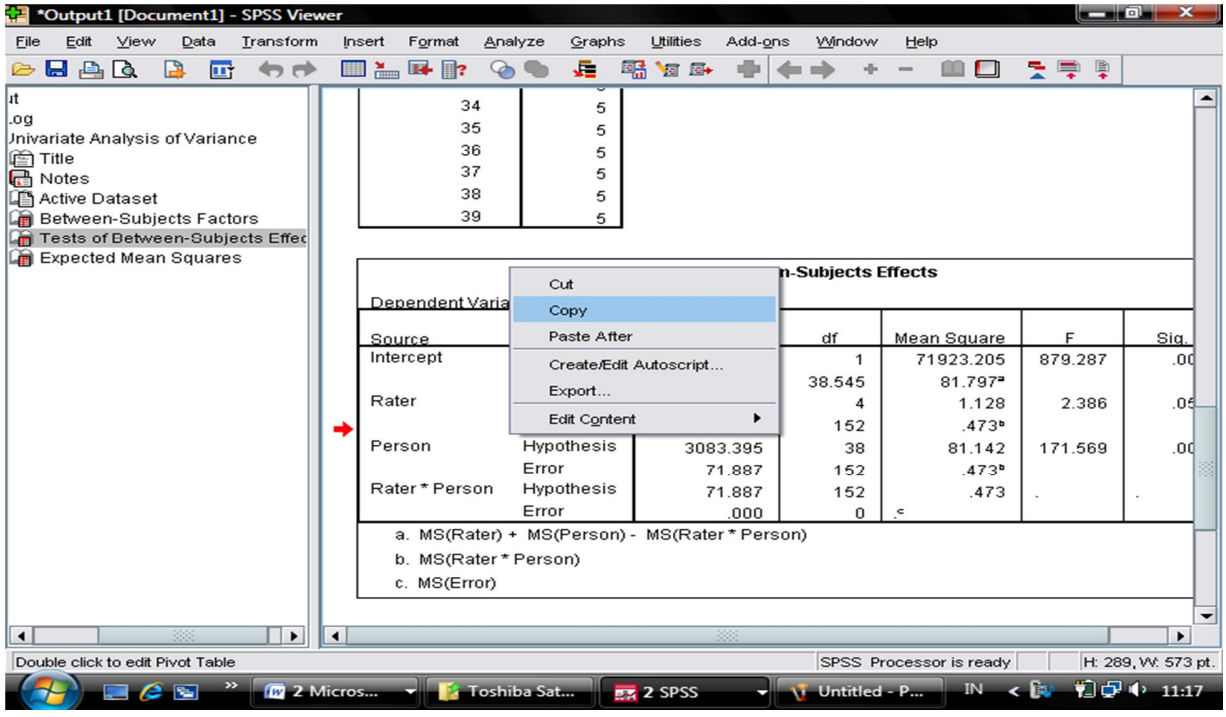
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	71923.205	1	71923.205	879.287	.000
	Error	3152.882	38.545	81.797[a]		
Rater	Hypothesis	4.513	4	1.128	2.386	.054
	Error	71.887	152	.473[b]		
Person	Hypothesis	3083.395	38	81.142	171.569	.000
	Error	71.887	152	.473[b]		
Rater * Person	Hypothesis	71.887	152	.473	.	.
	Error	.000	0	.[c]		

a. MS(Rater) + MS(Person) - MS(Rater * Person)

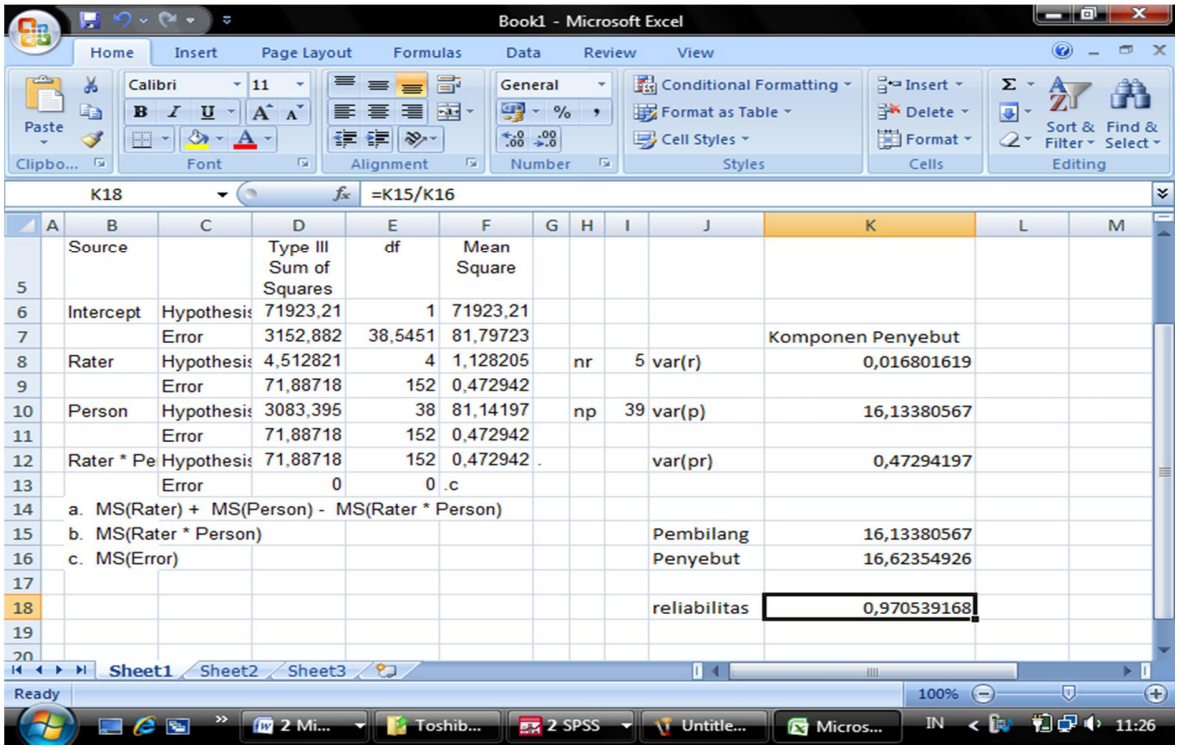
b. MS(Rater * Person)

c. MS(Error)

Output tersebut dapat dicopy, kemudian dibuka program Excel.



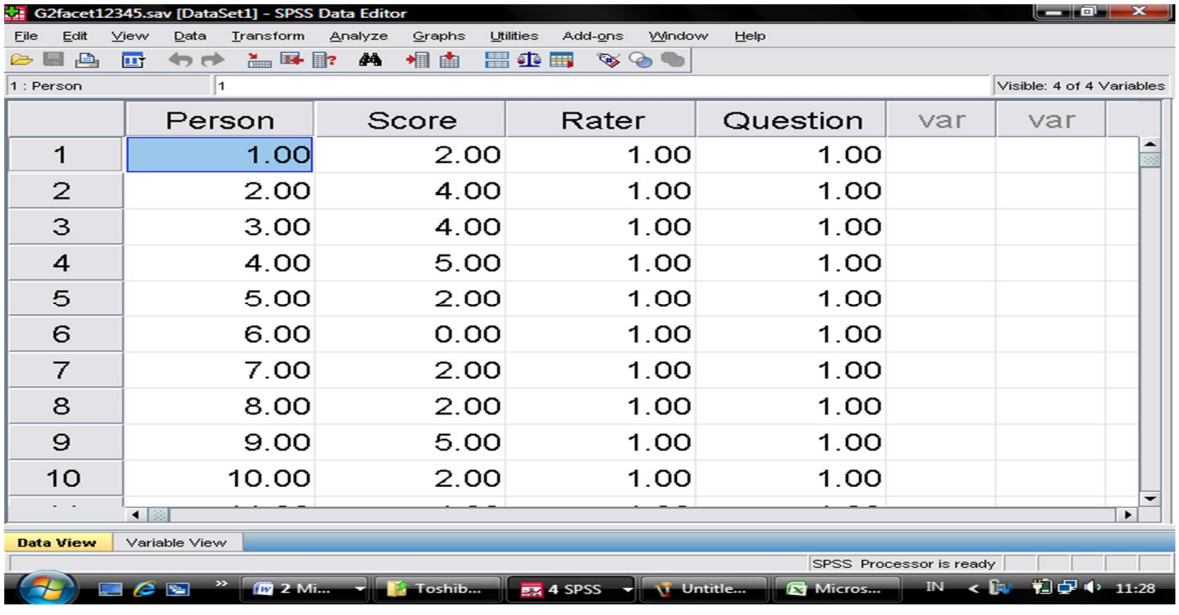
Copy tersebut kemudian di-Paste di Excel, sehingga kita dapat melakukan perhitungan untuk mengistemasi reliabilitas dengan 1 facet.



	Source		Type III Sum of Squares	df	Mean Square					
6	Intercept	Hypothesis	71923,21	1	71923,21					
7		Error	3152,882	38,5451	81,79723				Komponen Penyebut	
8	Rater	Hypothesis	4,512821	4	1,128205		nr	5	var(r)	0,016801619
9		Error	71,88718	152	0,472942					
10	Person	Hypothesis	3083,395	38	81,14197		np	39	var(p)	16,13380567
11		Error	71,88718	152	0,472942					
12	Rater * Pe	Hypothesis	71,88718	152	0,472942	.			var(pr)	0,47294197
13		Error	0	0	.c					
14	a. MS(Rater) + MS(Person) - MS(Rater * Person)									
15	b. MS(Rater * Person)								Pembilang	16,13380567
16	c. MS(Error)								Penyebut	16,62354926
18									reliabilitas	0,970539168

9. Estimasi Koefisien Geralisabilitas 2 Facet

Seperti halnya pada 2 facet, data diinputkan pada SPSS terlebih dahulu. Pada 2 facet ini, yang perlu diinputkan adalah *person*, *score* (dapat berupa skor butir), *rater* dan *question*.



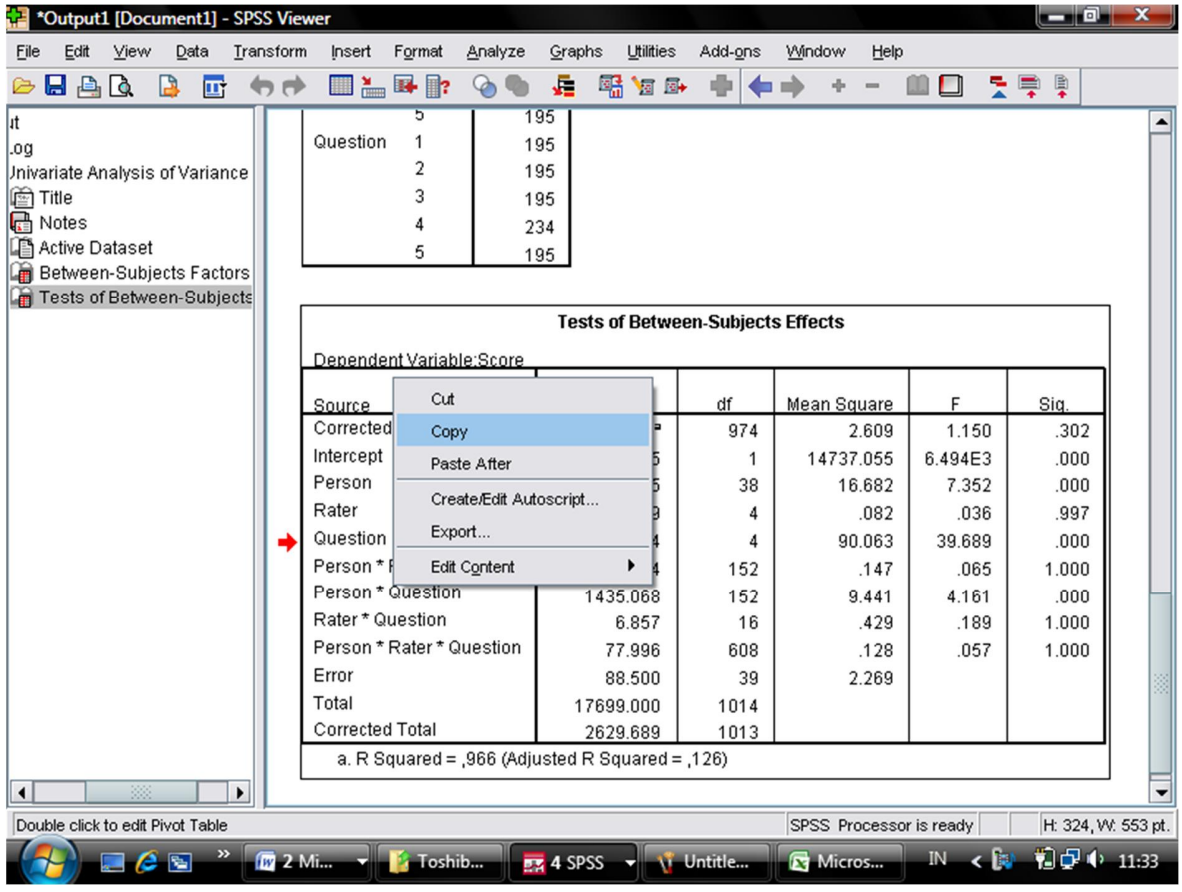
	Person	Score	Rater	Question	var	var
1	1.00	2.00	1.00	1.00		
2	2.00	4.00	1.00	1.00		
3	3.00	4.00	1.00	1.00		
4	4.00	5.00	1.00	1.00		
5	5.00	2.00	1.00	1.00		
6	6.00	0.00	1.00	1.00		
7	7.00	2.00	1.00	1.00		
8	8.00	2.00	1.00	1.00		
9	9.00	5.00	1.00	1.00		
10	10.00	2.00	1.00	1.00		

G2facet12345.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals
1	Person	Numeric	8	2
2	Score	Numeric	8	2
3	Rater	Numeric	8	2
4	Question	Numeric	8	2

Output yang diperoleh dapat diperoleh kemudian dapat dicopy, kemudian di-paste di Excel.



Selanjutnya dengan menggunakan Excel dapat dihitung pada lembar kerja Excel untuk menghitung koefisien generalisabilitas.

Bab VII
TEORI TES KLASIK DAN KETERBATASANNYA

Untuk mendapatkan instrumen berkualitas tinggi, selain dilakukan analisis secara teori (telaah butir berdasarkan aspek isi, konstruksi, dan bahasa) perlu juga dilakukan analisis butir secara empirik. Secara garis besar, analisis butir secara empirik ini dapat dibedakan menjadi dua, yaitu dengan pendekatan teori tes klasik dan teori respons butir (*Item Response Theory, IRT*). Pada bab ini, disajikan analisis untuk mengetahui karakteristik butir berdasarkan teori tes klasik, dan analisis berdasarkan teori respons butir disajikan pada bab kedelapan.

Teori tes klasik atau disebut teori skor murni klasik (Allen & Yen, 1979) didasarkan pada suatu model aditif, yakni skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran. Jika dituliskan dengan pernyataan matematis, maka kalimat tersebut menjadi

$$X = T + E \qquad (7.1)$$

dengan X skor amatan, T skor sebenarnya, E skor kesalahan pengukuran *(error score)*.

Kesalahan pengukuran yang dimaksudkan dalam teori ini merupakan kesalahan yang tidak sistematis atau acak. Kesalahan ini merupakan penyimpangan secara teoritis dari skor amatan yang diperoleh dengan skor amatan yang diharapkan. Kesalahan pengukuran yang sistematis dianggap bukan merupakan kesalahan pengukuran.

Ada beberapa asumsi dalam teori tes klasik. Skor kesalahan pengukuran tidak berinteraksi dengan skor sebenarnya, merupakan asumsi yang pertama. Asumsi yang kedua adalah skor kesalahan tidak berkorelasi dengan skor sebenarnya dan skor-skor kesalahan pada tes-tes yang lain untuk peserta tes *(testee)* yang sama. Ketiga, rata-rata dari skor kesalahan ini sama dengan nol. Asumsi-asumsi pada teori tes klasik ini dijadikan dasar untuk mengembangkan formula-formula dalam menentukan reliabilitas tes, pada bab kelima. Kriteria lain yang dapat digunakan untuk menentukan kualitas tes adalah indeks kesukaran dan daya pembeda.

A. Indeks Kesulitan

Indeks kesukaran disebut juga tingkat kesukaran butir. Konteks indeks kesukaran dan daya pembeda pada konteks ini tidak hanya berlaku untuk tes saja, namun juga untuk instrumen nontes juga. Hal yang perlu menjadi perhatian adalah penyekorannya. Penyekoran ada yang dilakukan secara dikotomi, misal benar-salah, ya-tidak, melakukan-tidak melakukan suatu kegiatan, ada-tidak ada dan lainnya. Biasanya pada penyekoran dikotomi, yang benar atau yang melakukan diskor 1 dan yang salah atau tidak melakukan diskor 0.

Model penskoran yang lain adalah model penskoran politomi atau sering disebut penskoran politomus atau politomi. Pada penskoran politomi ini, objek, baik hasil pemikiran ataupun tindakan yang diukur dinilai, bukan hanya dengan 1-0 saja, namun bervariasi. Misalnya 0-1-2-3, atau pada skala Likert dapat berupa 4 untuk sangat setuju, 3 setuju, 2 tidak setuju, dan 1 sangat tidak setuju. Terkait dengan adanya dua penskoran ini, maka tingkat kesulitan butir dan daya pembeda juga diklasifikasikan menjadi dua jenis, untuk tingkat kesulitan butir pada data dikotomi dan pada data politomi.

Tingkat kesukaran suatu butir soal, yang disimbolkan dengan p_i, merupakan salah satu parameter butir soal yang sangat berguna dalam penganalisian suatu tes. Hal ini disebabkan karena dengan melihat parameter butir ini, akan diketahui seberapa baiknya kualitas suatu butir soal. Jika p_i mendekati 0, maka soal tersebut terlalu sukar, sedangkan jika p_i mendekati 1, maka soal tersebut terlalu mudah, sehingga perlu dibuang. Hal ini disebabkan karena butir tersebut tidak dapat membedakan kemampuan seorang siswa dengan siswa lainnya.

Allen dan Yen (1979 : 122) menyatakan bahwa secara umum indeks kesukaran suatu butir sebaiknya terletak pada interval 0,3 – 0,7. Pada interval ini, informasi tentang kemampuan siswa akan diperoleh secara maksimal. Dalam merancang indeks kesukaran suatu perangkat tes, perlu dipertimbangkan tujuan penyusunan perangkat tes tersebut.

Untuk menentukan indeks kesukaran dari suatu butir pada instrumen dengan penskoran dikotomi, digunakan persamaan sebagai berikut :

$$p_i = \frac{\sum B}{N} \quad \text{.....(3)}$$

dengan :

p = proporsi menjawab benar pada butir soal tertentu (tingkat kesulitan)

ΣB = banyaknya peserta tes yang menjawab benar.

N = jumlah peserta tes yang menjawab.

Adapun indeks kesukaran dari suatu butir pada instrumen dengan penskoran politomi, digunakan persamaan sebagai berikut :

$$p_i = \frac{\sum X_i}{m.N} \quad \text{.....(3)}$$

dengan :

p = proporsi menjawab pada butir soal tertentu (tingkat kesulitan)

ΣX_i = banyaknya peserta tes yang menjawab benar.

N = jumlah peserta tes yang menjawab.

m = skor maksimum tiap butir

B. Daya Pembeda

Untuk menentukan daya pembeda, dapat digunakan indeks diskriminasi, indeks korelasi biserial, indeks korelasi *point biserial*, dan indeks keselarasan. Pada analisis butir dalam penelitian ini, hanya digunakan indeks korelasi *point biserial*. Koefisien korelasinya untuk suatu butir tes ditentukan dengan rumus:

$$\mathbf{r_{pbis}} = \left[\frac{\overline{X}_1 - \overline{X}}{s_X}\right]\sqrt{\frac{p_1}{1-p_1}} \quad \text{.....(4)}$$

dengan r_{pbis} = koefisien korelasi point biserial, X_i merupakan variabel kontinu, $\overline{X}_1$ merupakan rerata skor X untuk peserta tes yang menjawab benar butir tersebut, $\overline{X}$ merupakan rerata skor X, s_X merupakan standar deviasi dari skor X, dan p_1 merupakan proporsi peserta tes yang menjawab benar butir tersebut.

Pada suatu butir soal, indeks daya beda dikatakan baik jika lebih besar atau sama dengan 0,3. Indeks daya pembeda suatu butir yang kecil nilainya akan menyebabkan butir tersebut tidak dapat membedakan siswa yang kemampuannya tinggi dan siswa yang kemampuannya rendah. Pada analisis tes dengan *Content-Referenced Measures*, indeks

daya pembeda butir tidak terlalu perlu menjadi perhatian, asalkan tidak negatif (Ebel & Frisbie, 1986; Frisbie, 2005). Jika nilainya kecil, menunjukkan bahwa kemencengan distribusi skor dari populasi, yang juga mengakibatkan validitas tes menjadi rendah.

Teori tes klasik memiliki beberapa kelemahan mendasar. Kebanyakan statistik yang digunakan dalam model tes klasik seperti tingkat kesukaran dan daya pembeda soal sangat tergantung pada sampel yang dipergunakan dalam analisis. Rerata tingkat kemampuan, rentang, dan sebaran kemampuan siswa yang dijadikan sampel dalam analisis sangat mempengaruhi nilai statistik yang diperoleh. Sebagai contoh, tingkat kesukaran soal akan tinggi apabila sampel yang akan digunakan mempunyai kemampuan lebih tinggi dari rerata kemampuan siswa dalam poulasinya. Daya pemeda soal akan tinggi apabila tingkat kemampuan sampel bervariasi atau mempunyai rentang kemampuan yang besar, demikian pula dengan reliabilitas tes.

Kelemahan kedua yakni skor siswa yang diperoleh dari suatu tes sangat terbatas pada tes yang digunakan. Kesimpulan hasil tes tidak dapat digeneralisasikan di luar tes yang digunakan. Skor perolehan seseorang sangat tergantung pada pemilihan tes yang digunakan bukan pada kemampuan peserta tes tersebut. Karena keterbatasan penggunaan skor tes, teori tes klasikal tidak mempunyai dasar untuk mempelajari perkembangan kemampuan siswa dari waktu ke waktu, kecuali jika siswa tersebut menempuh tes yang sama dari waktu ke waktu.

Ketiga, konsep keajegan/reliabilitas tes dalam konteks teori tes klasik didasarkan pada kesejajaran perangkat tes sangat sukar untuk dipenuhi. pada praktiknya, sulit sekali memperoleh dua perangkat tes yang benar-benar sejajar. Jika prosedur tes retes digunakan, sampel yang diambil sangat tidak mungkin berperilaku sama pada saat tes dikerjakan untuk yang kedua kalinya.

Keempat, teori tes klasik tidak memberikan landasan untuk menentukan bagaimana respons seseorang peserta tes apabila diberikan butir tertentu. Tidak adanya informasi ini tidak memungkinkan melakukan desain tes yang bervariasi sesuai dengan kemampuan peserta tes (*adaptive or tailored testing*).

Kelima, indeks kesalahan baku pengukuran dipraasumsikan sama untuk setiap peserta tes. Padahal seseorang peserta tes mungkin berperilaku lebih konsisten dalam menjawab soal dibandingkan peserta tes lainnya. Demikian pula sebaliknya, banyak

sekali kesalahan individual. Kesalahan pengukuran sebenarnya merupakan perilaku peserta tes yang bersifat perorangan dan bukan perilaku tes.

Terakhir, prosedur-prosedur yang berkaitan dengan teori tes klasik seperti pengujian bias butir soal dan penyetaraan tes tidak bersifat praktis dan sukar untuk dilakukan. Demikian pula halnya dengan penyetaraan yang sifatnya vertikal. Untuk mengatasi hal itu, digunakanlah pendekatan teori lain yang disebut denga teori respons butir.

C. Kelemahan Teori Tes Klasik

Teori tes klasik memiliki beberapa kelemahan mendasar. Kebanyakan statistik yang digunakan dalam model tes klasik seperti tingkat kesukaran dan daya pembeda soal sangat tergantung pada sampel yang dipergunakan dalam analisis. Rerata tingkat kemampuan, rentang, dan sebaran kemampuan siswa yang dijadikan sampel dalam analisis sangat mempengaruhi nilai statistik yang diperoleh. Sebagai contoh, tingkat kesukaran soal akan tinggi apabila sampel yang akan digunakan mempunyai kemampuan lebih tinggi dari rerata kemampuan siswa dalam populasinya. Daya pembeda butit instrumen akan tinggi apabila tingkat kemampuan sampel bervariasi atau mempunyai rentang kemampuan yang besar. Demikian pula dengan reliabilitas tes. **Kelemahan ini disebut dengan *Group Dependent*.**

Kelemahan kedua yakni skor siswa yang diperoleh dari suatu tes sangat terbatas pada tes yang digunakan. Kesimpulan hasil tes tidak dapat digeneralisasikan di luar tes yang digunakan. Skor perolehan seseorang sangat tergantung pada pemilihan tes yang digunakan bukan pada kemampuan peserta tes tersebut. Karena keterbatasan penggunaan skor tes, teori tes klasik tidak mempunyai dasar untuk mempelajari perkembangan kemampuan siswa dari waktu ke waktu, kecuali jika siswa tersebut menempuh tes yang sama dari waktu ke waktu.

Ketiga, konsep keajegan/reliabilitas tes dalam konteks teori tes klasik didasarkan pada kesejajaran perangkat tes sangat sukar untuk dipenuhi. pada praktiknya, sulit sekali memperoleh dua perangkat tes yang benar-benar sejajar. Jika prosedur tes retes digunakan, sampel yang diambil sangat tidak mungkin berperilaku sama pada saat tes dikerjakan untuk yang kedua kalinya.

Keempat, teori tes klasik tidak memberikan landasan untuk menentukan bagaimana respons seseorang peserta tes apabila diberikan butir tertentu. Tidak adanya informasi ini tidak memungkinkan melakukan desain tes yang bervariasi sesuai dengan kemampuan peserta tes (*adaptive or tailored testing*).

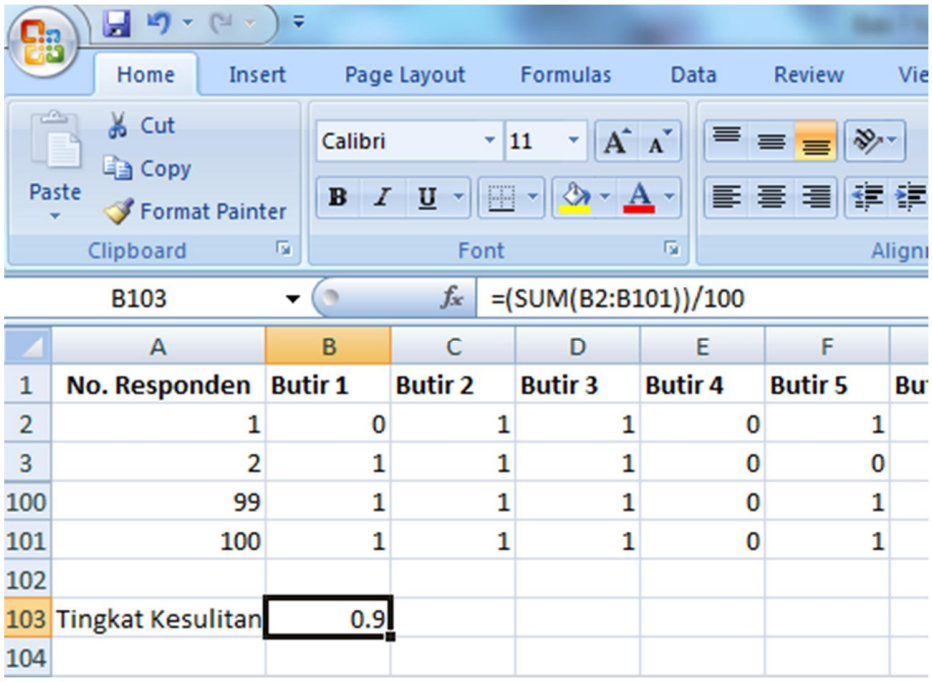
Kelima, indeks kesalahan baku pengukuran dipraasumsikan sama untuk setiap peserta tes. Padahal seseorang peserta tes mungkin berperilaku lebih konsisten dalam menjawab soal dibandingkan peserta tes lainnya. Demikian pula sebaliknya, banyak sekali kesalahan individual. Kesalahan pengukuran sebenarnya merupakan perilaku peserta tes yang bersifat perorangan dan bukan perilaku tes.

Terakhir, prosedur-prosedur yang berkaitan dengan teori tes klasik seperti pengujian bias butir soal dan penyetaraan tes tidak bersifat praktis dan sukar untuk dilakukan. Demikian pula halnya dengan penyetaraan yang sifatnya vertikal. Untuk mengatasi hal itu, digunakanlah pendekatan teori lain yang disebut dengan teori respons butir.

D. Analisis Karakteristik Butir Berdasarkan Teori Tes Klasik

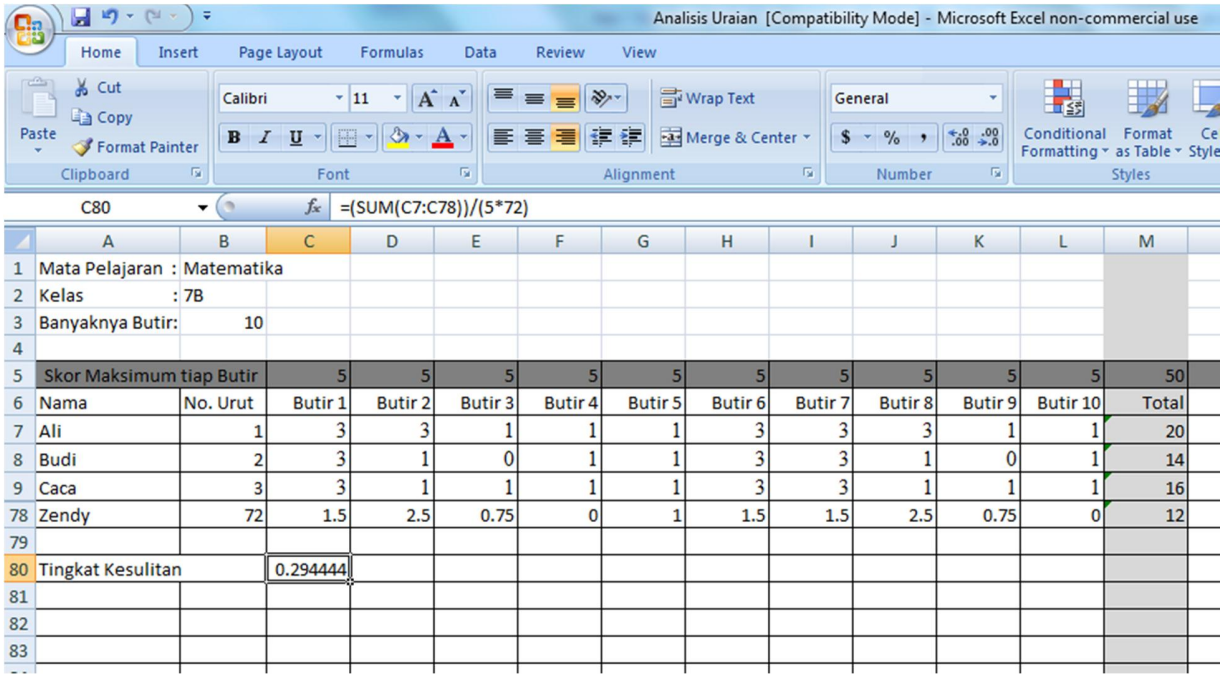
Analisis karakteristik butir berdasarkan teori tes klasik dapat dilakukan dengan program Excel, terutama untuk analisis tingkat kesulitan butir. Untuk analisis daya pembeda, program Excel dapat digunakan namun fungsinya sedikit agak rumit. Pada analisis ini, diberikan contoh menentukan tingkat kesulitan butir dengan penskoran dikotomi dengan menggunakan Excel, tingkat kesulitan butir politomi dengan menggunakan Excel, tingkat kesulitan dan daya pembeda butir dikotomi dengan program QUEST, dan tingkat kesulitan dan daya pembeda butir politomi dengan program QUEST.

Untuk analisis butir dengan Excel, dilakukan dengan langkah-langkah sebagai berikut. Pada awalnya, diinputkan terlebih dahulu skor butir, misalnya betul diskor 1 dan salah diskor 0. Jumlah responden merupakan pembagi untuk menghitung tingkat kesulitan butir. Sebagai contoh, untuk responden 100, pembagi untuk menghitung tingkat kesulitan (p) adalah 100.



	A	B	C	D	E	F
1	No. Responden	Butir 1	Butir 2	Butir 3	Butir 4	Butir 5
2	1	0	1	1	0	1
3	2	1	1	1	0	0
100	99	1	1	1	0	1
101	100	1	1	1	0	1
102						
103	Tingkat Kesulitan	0.9				
104						

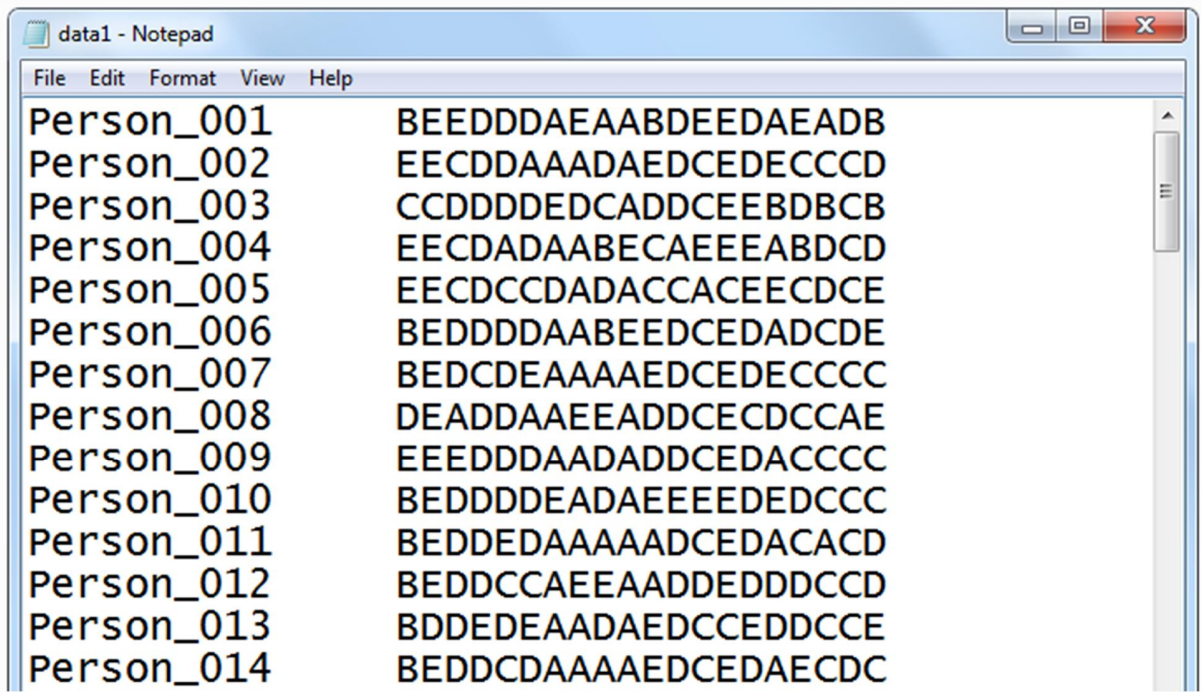
Untuk analisis tingkat kesulitan pada butir soal uraian atau nontes dengan penskoran politomi, maka pembagi untuk menghitung tingkat kesulitan adalah skor maksimum yang dapat diperoleh responden dikalikan banyaknya responden. Pada kasus analisis butir 1 berikut, respondennya 72 dan skor maksimum tiap butir 5, sehingga pembagi untuk menghitung p adalah 72 dikalikan 5.



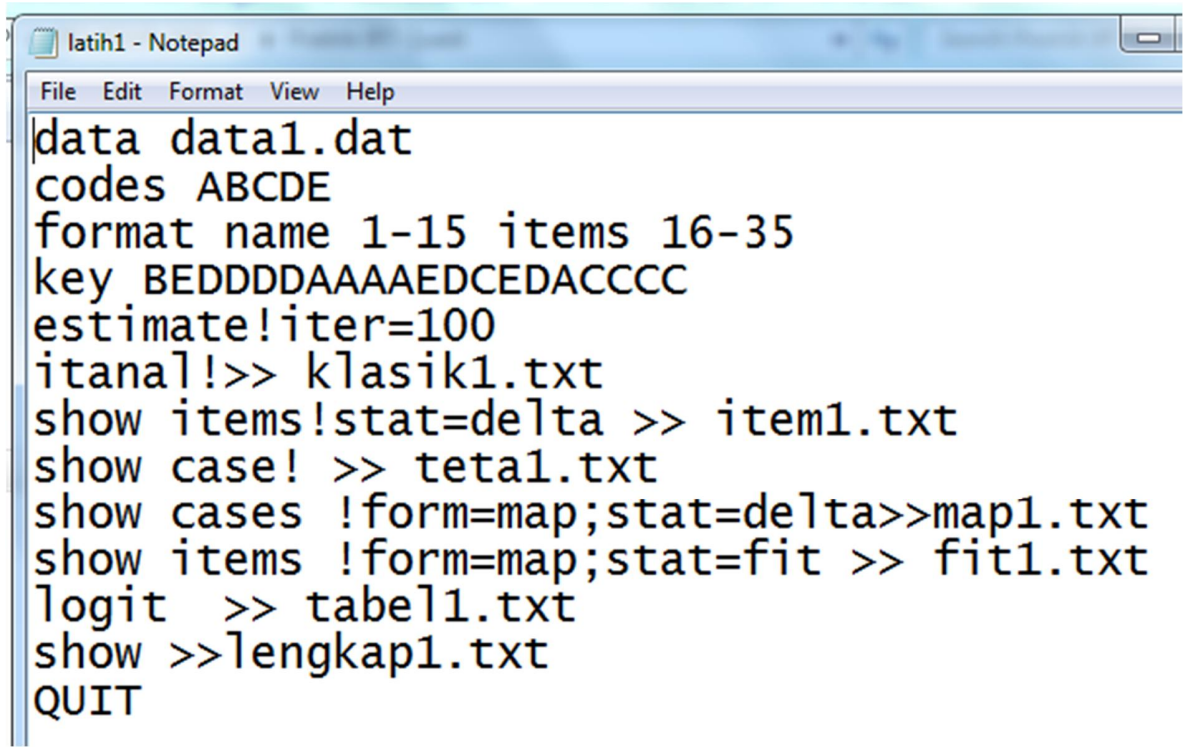
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Mata Pelajaran :	Matematika											
2	Kelas :	7B											
3	Banyaknya Butir:	10											
4													
5	Skor Maksimum tiap Butir		5	5	5	5	5	5	5	5	5	5	50
6	Nama	No. Urut	Butir 1	Butir 2	Butir 3	Butir 4	Butir 5	Butir 6	Butir 7	Butir 8	Butir 9	Butir 10	Total
7	Ali	1	3	3	1	1	1	3	3	3	1	1	20
8	Budi	2	3	1	0	1	1	3	3	1	0	1	14
9	Caca	3	3	1	1	1	1	3	3	1	1	1	16
78	Zendy	72	1.5	2.5	0.75	0	1	1.5	1.5	2.5	0.75	0	12
79													
80	Tingkat Kesulitan		0.294444										
81													
82													
83													

Cara lain menghitung tingkat kesulitan dan daya pembeda butir adalah dengan menggunakan QUEST. Contoh input data, sintaks, dan output sebagai berikut.

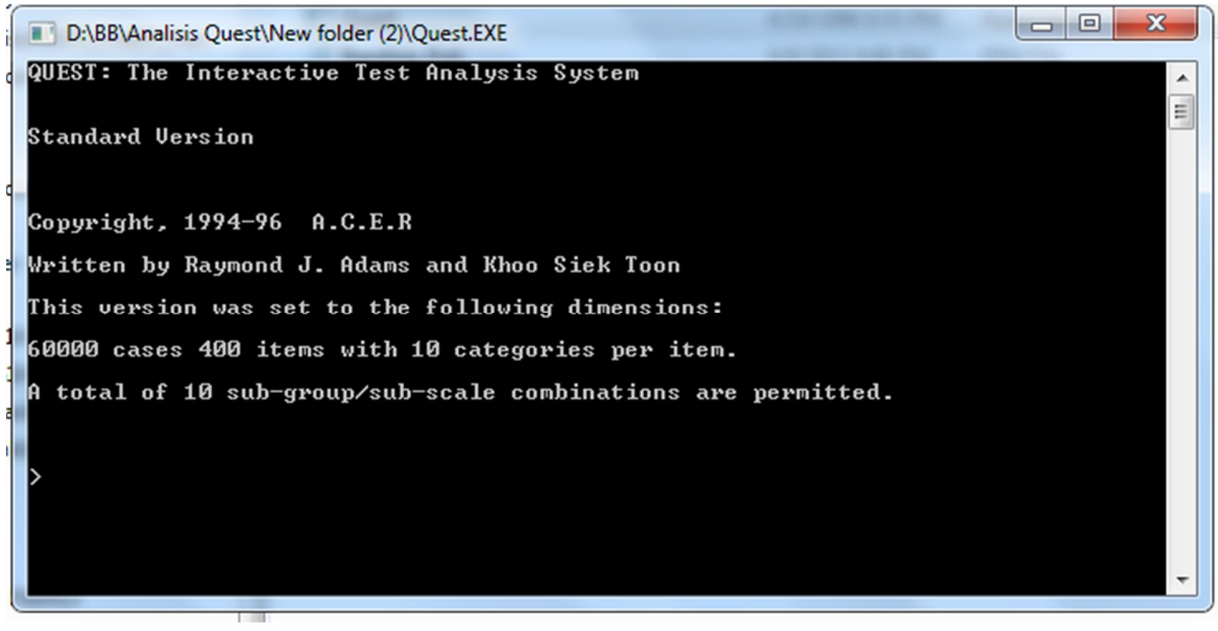
Input data dapat melalui Notepad, pada kasus ini 10 karakter pertama merupakan identitas peserta, 20 butir berikutnya merupakan butir yang dianalisis.



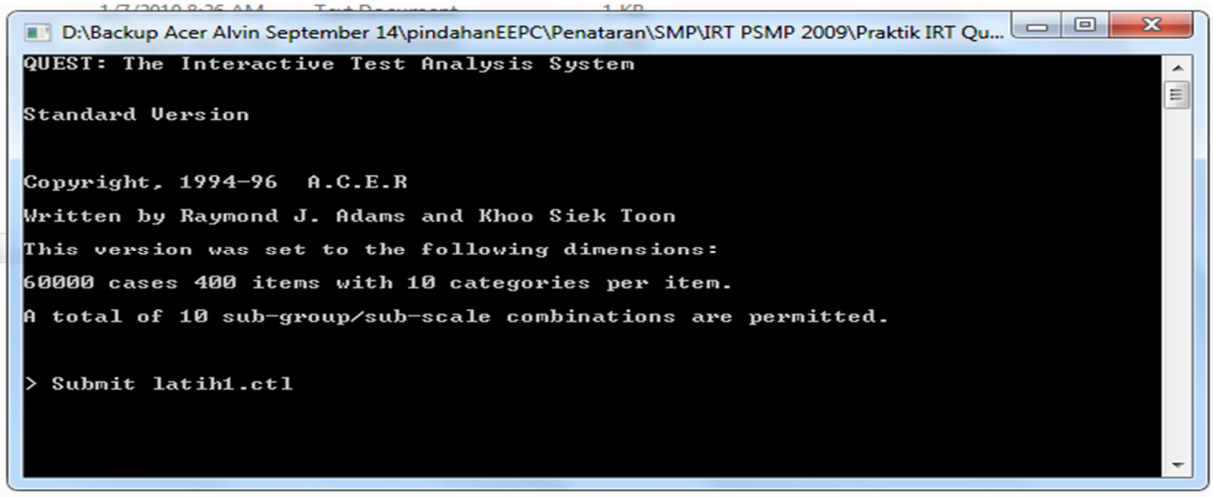
Sintaks atau baris perintah dapat diketikkan pula dengan Notepad. Contohnya sebagai berikut.



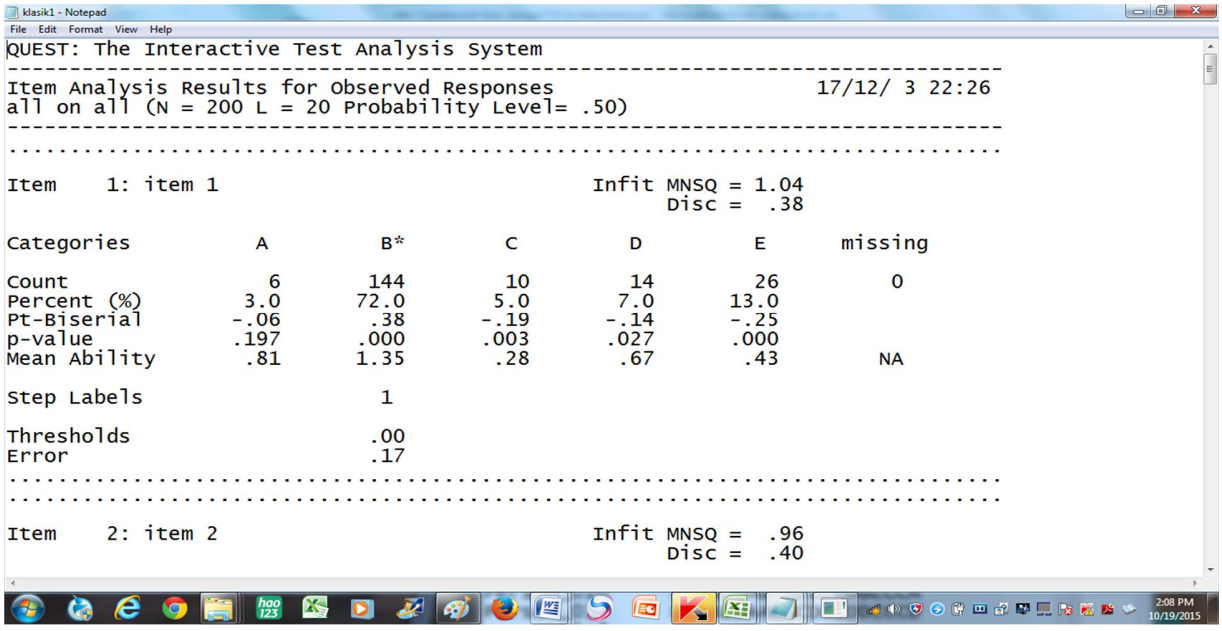
Kemudian data, sintaks dan program QUEST disimpan dalam satu folder. Program QUEST diklik dua kali, kemudian akan muncul tampilan sebagai berikut.



Kemudian diketikkan Submit, spasi nama file sitaksnya berikut ekstensinya.



Setelah itu klik Enter, pada folder tempat kita menganalisis file akan bertambah. Output Klasik.txt merupakan hasil analisis teori tes klasik, Hasilnya sebagai berikut.



Hal tersebut dapat diinterpretasikan sebagai berikut. Untuk butir 1, B* menunjukkan bahwa butir nomor 1 kuncinya B. Tingkat kesulitan butir nomor 1 dalah 72% atau 0,72. Untuk pilihan lainnya yang bukan kunci, pilihan butir baik jika proporsi nilainya tidak nol atau dengan kata lain ada yang memilih. Daya pembedanya (Pt-Biserial) sebesar 0,38. Korelasi point biserial ini harganya negatif untuk bukan kunci (distaktor).

DAFTAR PUSTAKA

ACER. (2013). *PISA 2012 Released Mathematics Items*. Tersedia di http://www.oecd.org/pisa/pisaproducts/ diambil tanggal 16 Mei 2015.

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement*, Vol. 22, pp. 37-53.

Aiken, L, R. (1980). Content validity and reliability of single items or questionaires. *Educational and Psychological Measurement, 40*, 955-967.

Aiken,L, R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*, 131-142.

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.*Washington, DC: American Psychological Association.

American Educational Research Assosiation, American Psychological Assosiation, and National Council on Measuremen in Education.(1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Anonim. 2001. Faktor Analysis. *Journal of Consumer Psychology*, 10(1&2), 75-82. Lawrence Erlbaum.

Babbie, E. 2004. *The Practice of Sosial Research*. Singapore : Wadsworth.

Bolt, D.M. & Lall, V.M. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Marcov chain Monte-Carlo. *Applied Psychological Measurement*, No. 27, pp. 395-414.

Cizek, G.J., Rosenberg, S.L. & Koons, H.H. (2008). Source of validity evidence for educational and psychological test. *Educational and Psychological Measurement,* Vol. 68, pp. 397-412.

Cohen, L., Manion, L., & Morrison, K. 2011. *Research Methods in Education*. Routlege: New York.

Croker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehard and Winston Inc.

Direktorat PSMP. 2013. Panduan Pelaksanaan Penilaian Menggunakan Kurikulum 2013 Direktorat PSMP. Bahan Penataran.

Du Toit, M. (2003). *IRT from SSi: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood: SSi.

Fernandes, H. J. X. (1984). *Evaluation of educational program*. Jakarta: National Education Planning, Evaluating and Curriculum Development.

Frisbie, D. A. 2005. Measurement 101: Some fundamentals revisited. Educational Measurement: Issues and Practice, 25 (3), 21-28.

Garson, D. 2006. *Factor Analysis, Path Analysis & SEM*. Diambil tanggal 24 September 2006 dari http://www2.chass.ncsu.edu/garson/pa765/index.htm .

Geldhof, G.J., Preacher, K., Zyphur, M.J. (2014). Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework. *Psychological Methods*, 19(1), 72-91.

Gregory, R.J. (2007). *Psychological testing: history, principles, and applications*. Boston: Pearson.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA : Kluwer Inc.

Hambleton, R.K., Swaminathan, H & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA : Sage Publication Inc.

Heri Retnawati. (2015). Akurasi Instrumen Skala Likert dan Pilihan Ganda untuk Mengukur Self Regulated Learning. *Laporan Penelitian*. Tidak Dipublikasikan.

Heri Retnawati (2008). Estimasi efisiensi relative tes berdasarkan teori tes klasik dan teori respons butir. *Disertasi*. Universitas Negeri Yogyakarta, tidak dipublikasikan.

Heri Retnawati, Dhoriva Urwatul Wutsqo, Endang Listyani, dkk. 2013. Kesulitan Guru Matematika dalam Memecahkan Masalah Matematika (Studi Kasus Guru Sekolah di Indonesia yang Persentase Kelulusannya Kurang dari 100%). *Laporan Penelitian*, Pendidikan Matematika FMIPA Universitas Negeri Yogyakarta.

Heri Retnawati. (2009). Perbandingan Model Regresi Tunggal dan Ganda pada *Benchmarking* Skor Tes (Studi Validitas Kriteria Test of English Proficiency terhadap ITP-TOEFL). *Makalah* Seminar Nasional dan Konferensi Ilmiah HEPI di UIN Syarif Hidayatullah Jakarta.

Hullin, C. L., et al. (1983). *Item response theory : Application to psichologycal measurement.* Homewood, IL : Dow Jones-Irwin.

Joreskog, K. & Sorbom, D. 1993. *Lisrel 88 : Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale, NJ : Scientific Software International.

Kamata, A., Turhan, A., Darandari, E. (2003). *Estimating Reliability forMultidimensional Composite Scale Scores*. Paper. Presented at the annual meeting of American Educational Research Association, Chicago, April 2003.

Kerlinger, F.N. (1986). *Asas-asas penelitian behavioral* (Terjemahan L.R. Simatupang). Yogyakarta: Gajahmada University Press.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 25, 146-162.

Kleinbaum, D.G dkk.(1998). *Applied Regression Analysis and Other Multivariate Methods*. Pacific Groove : Duxbury Press.

Kumaidi. (2014). Validitas dan pemvalidasian instrumen penilaian karakter. *Makalah* disampaikan dalam Seminar Nasional Pengembangan Instrumen Penilaian Pendidikan Karakter yang valid, diselenggarakan Fakultas Psikologi, Universitas Muhammadiyah Surakarta, 24 Mei 2014.

Lawrence, M.R. (1994). Question to ask when evaluating test. *Eric Digest*. *Artikel*. Diambil dari: http://www. ericfacility. net/ ericdigest/ ed.385607.html tanggal 6 Januari 2007.

Linn, R.L. & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7thed.). EnglewoodCliffs, NJ: Prentice-Hall.

Lissitz, W. & Samuelsen, K. (2007). Further clarification regarding validity and education. *Educational Researcher*, Vol. 36, No. 8, pp. 482-484.

Masters, G.N. 2010. The Partial Credit Model. Dalam Nering, M.L. & Ostini, R. (Eds). Handbook of Item Response Theory Models. New York: Routlegde.

Mehrens, W.A. & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. New York: Hold, Rinehart and Wiston, Inc.

Messick, S. (1989).Validity. Dalam R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Muraki, E. (1999). New appoaches to measurement. Dalam Masters, G.N. dan Keeves, J.P.(Eds). Advances in measurement in educational research and assesment. Amsterdam : Pergamon.

Muraki, E. (1999). New appoaches to measurement. Dalam Masters, G.N. dan Keeves, J.P.(Eds). *Advances in measurement in educational research and assesment*. Amsterdam : Pergamon.

Muraki,E., & Bock, R.D. (1997). *Parscale 3: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scintific Software Inc.

Nunally, J. (1978). *Psychometric theory (2nd ed.)* . New York: McGraw Hill.

Ostini, R. & Nering, M.L. (2006). *Polytomous Item Response Theory Models*. Sage Publications, Inc.

Pedhazur, E.J. (1973). *Multiple Regression in Behavioral Research*. New York : Holt, Rinehart and Winston.

Penev, S. & Raykov, T. (2006). On the Relationship Between Maximal Reliability and Maximal Validity of Linear Composites. MULTIVARIATE BEHAVIORAL RESEARCH, *41*(2), 105–126.

Popham, W.J. (1995). *Classroom assessment: What teachers need to know*. Boston, MA: Allyn and Bacon, Inc.