PRINSIP-PRINSIP PENGEMBANGAN INSTRUMEN PENELITIAN

Makalah

Ditulis untuk disampikan pada
Kuliah Umum bagi mahasiswa Program Studi Pendidikan Matematika
Fakultas Sains dan Teknologi UIN Sunan Kalijaga

"PENGEMBANGAN INSTRUMEN PENELITIAN DAN HASIL BELAJAR MATEMATIKA DALAM RANGKA PENINGKATAN KUALITAS PEMBELAJARAN MATEMATIKA"

di Ruang Teatrikal Fakultas Sains dan Teknologi UIN Sunan Kalijaga
Tanggal 12 Juni 2013

Penulis: Ali Ridho, M.Si.
Fakultas Psikologi UIN Maliki Malang

Ridho, A. (2013). Prinsip-prinsip Pengembangan Instrumen Penelitian.
Paper dipresentasikan pada Kuliah Umum bagi mahasiswa Program Studi Pendidikan Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga, Yogyakarta.

PENGEMBANGAN INSTRUMEN PENELITIAN[1]

Oleh: Ali Ridho[2]

A. PENGANTAR

Terdapat beberapa jenis desain penelitian dalam pendidikan. Creswell (2012) membagi desain penelitian menjadi 8 jenis: (1) eksperimental, (2) korelasional, (3) survei, (4) *grounded theory*, (5) etnografis, (6) naratif, (7) *mixed method*, dan (8) *action research*. Sementara itu, ada pula yang membaginya menjadi: (1) historis, (2) deskriptif, (3) eksperimental, (4) korelasional, (5) kualitatif, (6) evaluasi program, (7) studi kasus, (8) kebijakan, dan (9) evaluasi organisasional (Anderson, 2005). Dalam berbagai desain penelitian, pengumpulan data merupakan salah satu kegiatan yang krusial.

Data yang dikumpulkan berupa data kualitatif ataupun kuantitatif. Data kualitatif dapat berupa tempat, kata, benda, gambar, yang bersifat bukan angka. Sementara itu data kuantitatif adalah data yang berbentuk atau bersifat angka. Umumnya, penelitian kualitatif menggunakan peneliti sebagai instrumen utama dalam penggalian data penelitian. Sementara itu, dalam penelitian kuantitatif umumnya peneliti memanfaatkan kuesioner, observasi, atau *checklist*.

Penarikan kesimpulan hasil penelitian dipengaruhi oleh kesesuaian dan keabsahan data penelitian. Data dalam penelitian dikumpulkan melalui bantuan instrumen penelitian. Oleh sebab itu kemampuan instrumen dalam mengungkap data penelitian yang dituju menjadi pertimbangan yang penting.

Mengacu pada pendapat Colton dan Covert (2007), instrumen adalah suatu alat yang digunakan untuk mengukur fenomena, merekam informasi yang ditujukan untuk penilaian dan pengambilan keputusan. Terkait dengan tipe instrumen yang akan dikembangkan, Colton dan Covert menyarankan agar

---

[1] Makalah disampaikan pada Kuliah Umum bagi mahasiswa Program Studi Pendidikan Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga (12 Juni 2013)

[2]Dosen di Fakultas Psikologi UIN Maliki Malang; Email: ali.ridho@yahoo.com

Ridho, A. (2013). Prinsip-prinsip Pengembangan Instrumen Penelitian.
Paper dipresentasikan pada Kuliah Umum bagi mahasiswa Program Studi Pendidikan Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga, Yogyakarta.

peneliti perlu memperhatikan kesesuaiannya dengan beberapa hal; (1) tujuan penelitian, (2) desain penelitian, (3) objek pengukuran, (4) metode pengumpulan data, dan (5) sumber daya yang dimiliki.

Dalam bidang pengukuran, dilihat dari wilayah atribut yang diungkap, secara umum alat ukur dapat dikategorikan menjadi dua wilayah yaitu wilayah kognitif dan wilayah nonkognitif (Suryabrata, 2000). Hal ini diperkuat juga oleh Azwar (2012) yang mengutip pernyataan Cronbach bahwa tes dapat dibagi menjadi dua kelompok besar yaitu tes yang mengukur kinerja maksimum (*maximum performance*) dan tes yang mengukur kinerja tipikal (*typical performance*). Meski tidak secara keseluruhan, namun tes kinerja maksimum lebih dekat dengan tes yang mengukur wilayah kognitif dan tes kinerja tipikal lebih dekat dengan tes yang mengukur wilayah afektif.

Menurut Mardapi (2008), paling tidak terdapat empat atribut afektif yang penting dalam pembelajaran, yaitu; sikap, minat, konsep diri, dan nilai. Sikap, menurut Ajzen (2005), merupakan sebuah disposisi individu dalam merespons suatu objek, situasi, institusi, atau peristiwa dengan arah yang positif atau negatif. Arah positif memiliki makna individu tersebut mendukung, sedangkan arah negatif berarti individu tersebut bertentangan dengan objek, situasi, institusi, atau peristiwa yang disikapi.

Sikap dan minat adalah dua konsepsi yang memiliki kaitan erat dimana minat seseorang pada suatu keadaan akan dilandasi oleh sikapnya terlebih dahulu. Sementara itu konsep diri berkaitan dengan nilai. Konsep diri, dalam derajat tertentu, ikut menentukan arah nilai-nilai yang dikembangkan oleh individu.

Selain sikap, minat, konsep diri, dan nilai, motivasi memberikan kontribusi yang penting dalam pembelajaran. Motivasi yang dimiliki oleh guru ataupun siswa ikut menentukan kekerasan usaha yang dilakukan dalam mencapai tujuan pembelajaran. Dalam kesempatan tulisan ini akan dikemukakan prinsip-prinsip umum dalam mengembangkan instrumen sejenis pengukuran terhadap motivasi dan sikap.

B. KARAKTERISTIK INSTRUMEN DAN SKOR YANG DIHASILKAN

Sebelum membahas secara lebih mendalam mengenai teknis pengembangan instrumen, terlebih dahulu perlu ditegaskan konsep yang akan memberikan gambaran bagaimana karakteristik instrumen. Prasyarat agar data penelitian yang diperoleh berdasarkan instrumen dapat dipertanggungjawabkan, ada dua karakteristik minimal yang dimiliki, yaitu valid dan reliabel.

1. Reliabel

Reliabel berakar dari kata *reliable* (bahasa Inggris) yang artinya "handal". Dalam istilah sehari-hari, seorang karyawan disebut handal (reliabel) tatkala ia selalu mampu menyelesaikan pekerjaannya dengan baik. Sebuah mobil dapat disebut reliabel jika selama sekian bulan dipakai, ia selalu mudah dihidupkan saat diperlukan. Seorang sopir bus malam disebut -handal dengan bukti bahwa selama 5 tahun mengemudi tiap malam, belum pernah ia mengalami kecelakaan karena *human error*.

Berdasarkan beberapa ilustrasi di atas dapat ditarik satu aspek yang menjadi kriteria andal tidaknya sesuatu yakni "pengulangan" dan "konsistensi". Dalam banyak kali pengulangan, konsisten atau tidakkah hasilnya? Setelah diulang-ulang selama 12 bulan, sebuah mobil layak disebut handal (reliabel) karena tiap kali kontak di on-kan, tiap kali pula mobil tersebut hidup.

Dalam setiap pengukuran, pertanyaan mendasar pada sebuah alat ukur adalah apakah hasil ukurnya akan konsisten antara satu pengukuran dengan pengukuran yang lain. Dalam pengukuran tinggi badan, misalnya, apakah hasil ukur alat ukur tinggi badan tersebut konsisten antara pengukuran satu minggu yang lalu dan pengukuran pada hari ini? Sekonsisten apakah manakala digunakan secara berulang-ulang? Konsistensi hasil ukur sebuah alat ukur ini pada pengukuran satu dengan pengukuran yang lain inilah yang merupakan ide dasar konsep reliabilitas.

Istilah "reliabel" dalam tes psikologi dan pendidikan berawal dari konsep yang sama tentang reliabilitas (kehandalan) seperti halnya pada contoh di atas. Pembahasan tentang reliabilitas pada akhirnya ingin menjawab pertanyaan: "Bagaimanakah skor hasil tes sebuah alat ukur dari satu administrasi ke

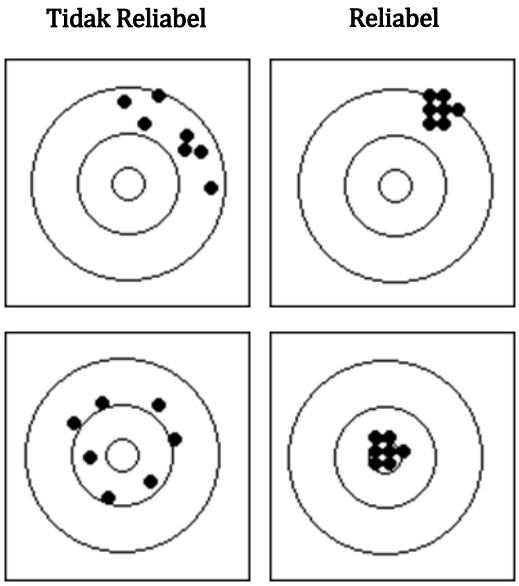
administrasi yang lain pada sampel yang relevan?" Skor tes disebut handal manakala tidak ada perbedaan hasil skor antara satu administrasi dengan administrasi yang lain, sepanjang masih dalam populasi yang relevan (alat ukur psikologi dikembangkan untuk mengukur atribut tertentu dengan pembatasan pada populasi yang memiliki karakteristik tertentu pula).

Berdasarkan karakteristiknya, tes yang baik adalah tes yang reliabel dan valid, yaitu tes yang dapat menghasilkan skor yang dapat dipercaya dan tepat sasaran. Berkenaan dengan dua karakteristik ini, reliabilitas merupakah sebuah kriteria prasyarat sebelum validitas. Reliabilitas dibutuhkan, tapi reliabilitas saja belum mencukupi. Sebuah tes yang reliabel tidak berarti pula tes tersebut valid (Nunnally, 1981).

Walaupun reliabilitas menduduki posisi kedua setelah validitas, akan tetapi reliabilitas menjadi prasyarat validitas. Memahami dan menghitung reliabilitas adalah langkah awal sebelum melakukan studi tentang validitas konstrak suatu alat ukur (Duhachek & Iacobucci, 2004). Oleh karena itu pembahasan mengenai reliabilitas dikemukakan terlebih dahulu sebelum membahas validitas.

Istilah reliabilitas dalam pengukuran dapat dimaknai sebagai konsistensi atau reproduksibilitas skor tes, yakni sejauh mana stabilitas simpangan skor para peserta tes pada situasi-situasi tes yang sama atau paralel. Makna tersebut diterjemahkan oleh para ahli psikometri yang pada intinya mengerucut pada "kepercayaan hasil ukur". Sampai berapa besar derajat kepercayaan hasil ukur sebuah tes inilah yang diwakili oleh istilah reliabilitas.

Manakala Setelah sebuah tes selesai disusun kemudian diadministrasikan pada kelompok subjek yang relevan, pengguna hasil tes tentu ingin mengetahui sejauh mana hasil ukur yang diperoleh jika dikenakan kembali pada kelompok subjek yang sama atau hampir sama pada waktu yang akan datang. Konsisten atau tidak? Jika konsisten, seberapa besar tingkat kekonsistenannya? Konsistensi hasil ukur inilah yang disebut sebagai reliabilitas (Crocker & Algina, 1986).



Gambar 1. Ilustrasi Reliabilitas Berdasarkan Konsistensi Beberapa Tembakan

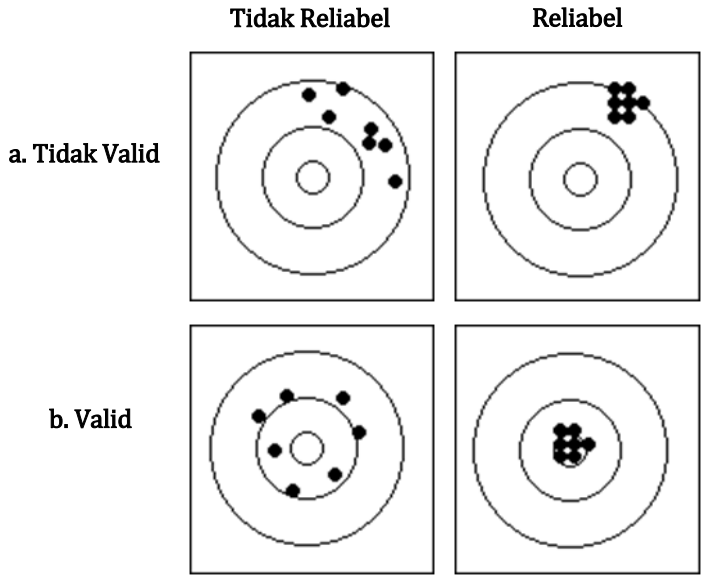
2. Valid

Evaluasi baik tidaknya sebuah alat ukur banyak didasarkan pada reliabilitas dan validitas skor yang dihasilkan. Evaluasi reliabilitas ditujukan untuk menjawab pertanyaan "*Apakah dari satu administrasi ke administrasi lain tes mampu menghasilkan skor yang konsisten*?", dan evaluasi terhadap validitas skor tes terkait dengan pertanyaan "*Apakah tes tersebut betul-betul mampu mengungkap atribut yang menjadi tujuan ukur*?".

Dalam penyelidikan ilmiah, validitas pernyataan merupakan seberapa besar derajat diperolehnya bukti empirik guna mendukung kebenaran dan kesesuaian pernyataan tersebut. Lebih spesifik pada pengukuran, sebuah pengukuran disebut valid bila betul-betul mengukur apa yang hendak diukur. Dapat dikatakan juga bahwa validitas memiliki makna sejauh mana ketepatan ukur atau kecermatan alat ukur dalam mengukur atribut yang menjadi tujuan ukurnya. Oleh karena itulah *educational testing service* (ETS) mengemukakan bahwa validasi merupakan aspek paling penting dalam menentukan kualitas hasil pengukuran (ETS, 2002).

Dibalik deskripsi yang kelihatan samar di atas, tampak bahwa validitas bukanlah merupakan bagian dari instrumen pengukuran, namun lebih melekat pada interpretasi serta penggunaan skor yang dihasilkan tes pada subjek yang

relevan. Menggunakan bahasa yang lain, validitas dibuktikan dengan diperolehnya kecocokan empirik antara interpretasi skor dan penggunaannya, serta terkait pula dengan dasar teoritik atribut tujuan ukur sebagai rasionalisasi.



Gambar 2. Ilustrasi Validitas Berdasarkan Ketepatan Beberapa Tembakan

Sebagai ilustrasi guna memahami konsep validitas, ilustrasi pada Gambar 2 menunjukkan hasil tembakan pada sasaran tembak yang ada di tengah. Anggaplah sasaran tembak merupakan sasaran dari sebuah alat ukur, yakni atribut psikologis yang hendak diukur. Validitas hasil tembakan ditunjukkan dengan seberapa tepat tembakan mengenai sasaran. Makin dekat dengan sasaran, maka makin valid tembakan yang dihasilkan. Gambar 2a menunjukkan hasil tembakan yang tidak valid karena letaknya yang relatif jauh dari sasaran. Sementara itu, Gambar 2b menunjukkan hasil tembakan yang relatif valid karena letaknya yang dekat dengan sasaran tembak.

Implementasi konsepsi validitas bukanlah persoalan sederhana. Sejak munculnya konsepsi validitas oleh Cronbach dan Meehl (1955), penelitian-penelitian yang memfokuskan pada proses validasi terhadap alat ukur di bidang pendidikan dan psikologi terus berkembang. Pada akhirnya 40 tahun kemudian muncul pendapat Messick (1995) yang mendefinisikan validitas sebagai satu

kesatuan, tidak terbagi-bagi sebagaimana menurut Cronbach dan Meehl. Messick (1995) menulis bahwa validitas adalah ringkasan evaluatif baik dalam bentuk bukti atau konsekuensi interpretasi dan penggunaan skor hasil tes.

Borsboom, Mellenbergh, dan van Heerden (2004) mengatakan bahwa sebuah tes disebut valid mengukur atribut laten tertentu jika dan hanya jika: (1) atribut tersebut ada, dan (b) variasi dalam atribut tersebut mampu menyebabkan variasi hasil dalam prosedur pengukuran. Artinya kedua keadaan tersebut adalah prasyarat agar skor yang dihasilkan oleh pengukuran bisa menjadi valid.

Berbeda dengan atribut fisik (misalnya panjang) yang jelas adanya, ada tidaknya atribut laten dalam diri manusia merupakan sesuatu yang hipotetis, sengaja diciptakan (dikonstrak) untuk mempermudah pemahaman terhadap karakteristik manusia. Untuk membuktikan keberadaan atribut laten manusia, diperlukan suatu pengukuran yang tepat. Pengukuran yang tepat akan menghasilkan skor yang valid. Dengan demikian validitas melekat pada skor hasil ukur, bukan pada alat ukur. Alat tes hanya sarana, setelah berinteraksi dengan peserta tes, menghasilkan skor yang valid.

Sebagai sebuah sarana, tes didesain melalui suatu kegiatan pengembangan tes yang relevan dengan tujuan pengukuran. Dalam pengembangan tersebut, posisi teori mengambil peranan yang paling penting. Seberapa besar teori terkait variabel yang diteliti tersebut dapat terwakili dalam alat ukur, hal tersebut akan selalu *debatable* karena sudut pandang teori yang berbeda dapat menghasilkan kebenaran relatif yang berbeda pula.

Dalam konteks penelitian, tujuan pengukuran tidak lain adalah mengetahui besarnya perbedaan variabel penelitian (atribut laten) pada tiap-tiap subjek ukur. Keberadaan atribut yang diukur, yang berupa kemampuan ataupun preferensi orang, adalah entitas yang tak dapat diamati. Oleh sebab itu disebut sebagai atribut laten. Berdasarkan respons-respons yang diberikan oleh sejumlah orang pada sejumlah stimulus akan menghasilkan variasi hasil ukur. Variasi dalam bentuk skor inilah prediktor terbaik atribut laten yang pada akhirnya dimaknai sebagai representasi variabel yang sedang diteliti. Bila sebuah tes ditujukan untuk mengukur kemampuan potensial yang dapat memprediksi kemampuan aktualnya

di masa mendatang, misalnya, idealnya hal yang diukur oleh tes tersebut adalah semata-mata kemampuan potensi. Pada kenyataannya, operasionalisasi konsep "potensi" yang dapat memprediksi berhasil tidaknya peserta tes di masa mendatang akan sangat dipengaruhi oleh proses abstraksi konsep "potensi" itu sendiri. Setelah melalui operasionalisasi dalam kategori-kategori atau dimensi-dimensi kemampuan tertentu, ujungnya ada pada aitem-aitem tes yang dikembangkan dalam mengungkap "potensi" yang dimaksud.

Pengukuran merupakan kegiatan yang lazim, bahkan rutin dilakukan dalam dunia pendidikan dan psikologi. Berdasarkan pengamatan penulis, banyak laporan penelitian di bidang pendidikan, psikologi, dan ekonomi yang menggunakan validitas aitem sebagai justifikasi valid tidaknya skor yang dihasilkan oleh alat ukur yang digunakan dalam penelitian dalam bentuk korelasi aitem-total (r_{bt}) atau korelasi aitem-total terkoreksi (*corrected item-total correlation*, $r_{b(t-b)}$). Kenyataan ini pernah disinggung oleh Naga (2004) yang mengatakan bahwa persoalan validitas tidak sesederhana itu, namun proses validasinya dilakukan terhadap hasil ukur sehingga bisa membuktikan konstrak yang dikembangkan betul-betul berlaku pada subjek yang menjadi tujuan ukur.

Messick (1995, 1998) menyebutkan 6 aspek dalam konsepsi validitas konstrak: isi, substansi, struktural, generalisasi, eksternal, dan konsekuensi. Sesuai dengan tujuan tes, tiap tes-tes dengan tujuan berbeda harus memiliki jenis bukti yang berbeda, bukan validitas yang berbeda. Dengan demikian validitas ia disebut sebagai satu kesatuan konsep (*unified*). Dua belas tahun setelah pendapat Messick tersebut, persoalan validitas dari sisi pengertian dan konsepsi menjadi perdebatan hangat, dibahas secara menarik dalam jurnal *Educational Researcher* volume 36 nomor 8 tahun 2007. Perdebatan dimulai dengan artikel Lissitz & Samuelsen (2007) yang membahas secara ringkas sejarah konsepsi validitas kemudian dilanjutkan konsepsi baru yang mereka tawarkan. Mereka mencoba merekonstruksi konsepsi validitas yang dikemukakan oleh Messick (1995) yaitu menolak konsepsi validitas sebagai satu kesatuan yang disebut sebagai validitas konstrak. Menurut Lissitz & Samuelsen (2007), validitas sebaiknya dipandang sebagai sebuah taksonomi yang dievaluasi berdasarkan isi tes dengan

mempertimbangkan dua aspek: (a) internal: proses laten, isi, dan reliabilitas; dan (b) eksternal: nomologis, kriteria, dan dampak yang ditimbulkan. Selain itu perlu diperhatikan pula sisi teoritik dan praktis yang menyertai. Tulisan tersebut kemudian ditanggapi oleh Gorin (2007), Embretson (2007), Mislevy (2007), Sireci (2007), dan Moss (2007).

Embretson (2007) menyetujui konsep baru validitas yang dikemukakan oleh Lissitz & Samuelsen (2007), namun tetap mengemukakan pentingnya menegakkan validitas konstrak. Sebuah sistem validitas yang bersifat universal dikembangkan oleh Embretson yang dapat diringkas menjadi validitas internal dan eksternal.

Gorin (2007) menemukan persoalan konsepsi validitas Lissitz & Samuelsen, yaitu pada persolan skoring. Gorin menekankan validitas perlu difokuskan pada evaluasi skor tes sehingga melengkapi sisi pemaknaan pada interpretasi skor tes sekaligus menguatkan argumen tentang validitas. Sementara itu, tanpa mendukung atau menolak konsepsi Lissitz & Samuelsen (2007), khusus pada pengembangan tes, Mislevy (2007) menekankan pentingnya desain tes dibandingkan desain validasi. Oleh karena itu bila seorang adalah pengembang tes maka sebaiknya ia memfokuskan pada desain tes secara detil. Desain validasinya diserahkan pada peneliti lain.

Berbeda dengan Lissitz dan Samuelsen (2007), Moss (2007) lebih cenderung menyetujui: (1) validitas sebagai satu kesatuan konsep; (2) basis validitas sebagai satu kesatuan konsep ialah validitas konstrak; dan (3) fokus pemaknaan validitas lebih pada interpretasi dan penggunaan skor tes. Sedangkan Sireci (2007) menanggapi tulisan Lissitz & Samuelsen (2007) dengan mengemukakan bahwa validitas isi saja tidak cukup memadai dalam memaknai validitas. Sireci menyatakan pentingnya standar sebagai acuan. Selain itu diperlukan penilaian secara integratif dalam hal konstrak yang mendasari teori, analisis isi, serta analisis data skor aitem dan tes.

Sejalan dengan pendapat Gorin (2007), Embretson (2007), Sireci (2007), penting kiranya menemukan bukti-bukti secara internal bahwa sebuah tes betul-betul valid mengukur selaras dengan fondasi konstraknya. Dengan kata lain,

struktur internal tes yang berupa aitem-aitem perlu ditemukan tingkat validitasnya. Aspek struktural dalam validitas dapat ditegakkan dengan cara *Confirmatory Factor Analysis* (CFA) (Dimitrov, 2010) dan deteksi *differential item functioning* (DIF) (Kim, Kim, & Kamphaus, 2010). Pembahasan secara lebih mendalam terkait dengan CFA dan DIF berada diluar tulisan ini.

C. LANGKAH-LANGKAH PENGEMBANGAN

Konstruksi tes atau pengembangan tes merupakan serangkaian aktivitas membuat alat ukur. Prosedur konstruksi ini mesti dilakukan dengan menyeluruh, rinci, spesifik dan hati-hati tahap demi tahap sehingga akan dihasilkan tes dengan kualitas yang baik, menghasilkan skor yang valid dan reliabel. Berikut adalah langkah-langkah dalam mengembangkan suatu alat ukur menurut DeVellis (2003).

1. Menentukan tujuan penggunaan skor pengukuran (*determine clearly what it is you want to measure*)
2. Mengembangkan aitem (*generate an item pool*)
3. Menentukan format pengukuran (*determine the format for measurement*)
4. Reviu aitem (*have the initial item pool reviewed by experts*)
5. Validasi aitem (*consider inclusion of validation items*)
6. Ujicoba (*administer items to a development sample*)
7. Evaluasi aitem (*evaluate the items*)
8. Pertimbangan perakitan akhir (*optimize scale length*)

1. Menentukan Tujuan Penggunaan Skor Pengukuran

Pengembangan alat ukur tidak mungkin dilepaskan dari konstrak yang dijadikan fondasi. Seorang pengembang alat ukur idealnya menyandarkan alat ukurnya pada teori substantif yang mendasari. Landasan teoritik yang relevan sebaiknya selalu digunakan sebagai acuan karena hal ini terkait langsung dengan pemaknaan terhadap skor yang dihasilkan alat ukur.

Pada saat mengembangkan alat ukur, unsur paling utama yang perlu menjadi perhatian adalah, akan dipergunakan untuk apa skor yang dihasilkan alat ukur. Kegunaan skor ini didesain awal, disesuaikan dengan variabel dan konteks penelitian.

Derajat kekhususan atau keumuman konstrak yang menjadi tujuan ukur menduduki posisi yang sangat penting. Konstrak *coping strategy* (strategi memecahkan masalah), misalnya, dapat dikembangkan dalam konteks memecahkan masalah-masalah belajar secara umum, bisa juga secara khusus pada mata pelajaran tertentu.

Konstrak motivasi dapat dijadikan landasan untuk mengukur motivasi dalam konteks belajar secara umum, bisa pula dipersempit untuk mengukur motivasi belajar dalam mata pelajaran matematika atau mata pelajaran lain tertentu. Implikasinya adalah pada kegunaan skor yang dihasilkan pengukuran. Bila konstrak motivasi dipersempit pada mata pelajaran matematika maka skor yang dihasilkan oleh pengukuran ditujukan untuk mendeskripsikan para siswa dalam hal motivasi belajar matematika.

Walaupun istilahnya sama yaitu konstrak motivasi, namun teori yang dijadikan landasan pengembangan alat ukurnya bisa bermacam-macam. Tiap dasar teori yang berbeda sangat mungkin akan menghasilkan aitem-aitem alat ukur yang berbeda pula. Berkenaan dengan motivasi dalam bidang akademik, paling tidak ada lima teori yang dapat digunakan sebagai acuan, yaitu: *self-efficacy theory* (SET) (Bandura, 1997), *attribution theory* (AT) (Malpass, 1994), *self-worth theory* (SWT) (Covington, 2000), *achievement goal theory* (AGT) (Seifert, 2004), dan *self-determination theory* (SDT) (Deci, Vallerand, Pelletier, & Ryan, 1991; Guiffrida, 2006; Ryan & Deci, 2006). Dengan demikian, peneliti atau pengembang alat ukur perlu memikirkan landasan teoritik yang digunakan, apakah salah satu atau gabungan diantara beberapa teori.

Bila landasan teoritik telah ditetapkan, langkah selanjutnya adalah menegaskan dan membatasi kawasan atau domain ukur. Penegasan kawasan ukur dituangkan ke dalam cetak biru (*blueprint*) yang nantinya akan dijadikan acuan didalam menulis isi atau materi aitem. Suatu misal akan dikembangkan alat ukur motivasi belajar matematika dengan berlandaskan pada teori *self-determination theory* (SDT). Untuk itu, diperlukan suatu pembatasan kawasan ukur motivasi ini. SDT membagi motivasi ke dalam sebuah kontinum yang dapat diurutkan sebagaimana disajikan pada Tabel 1.

Tabel 1. Cetak biru Skala Motivasi Belajar Matematika

No.	Aspek/Faktor/Subskala	Σ Aitem Skala	Σ Aitem pool
1.	*intrinsic regulation*	4	12
2.	*identified regulation*	4	12
3.	*introjected regulation*	4	12
4.	*external regulation*	4	12
5.	*amotivation*	4	12
	Jumlah	20	60

Sangat sulit kiranya menentukan jumlah aitem yang perlu dikembangkan dalam *pool* aitem. Sebanyak yang bisa dibuat oleh pengembang, inilah aturan main yang paling praktis. Namun tidak mungkin melakukan pengembangan aitem sebanyak aitem yang hendak digunakan pada pengukuran yang sebenarnya. Aitem berisikan kalimat-kalimat dengan variasi diksi yang akan direspons oleh subjek ukur. Subjek ukur memiliki tingkat pemahaman dan persepsi yang bervariasi terhadap isi atau materi dalam aitem. Oleh sebab itulah secara umum pengembangan aitem dilakukan memperbanyak sejumlah 3 atau 4 kali banyaknya aitem yang akan digunakan dalam pengukuran yang sebenarnya. Tabel 1 merupakan acuan yang tidak berlaku secara kaku. Jumlah aitem yang dikembangkan tentunya tidak selalu seperti yang dituangkan dalam tabel tersebut. Bisa kurang, bisa lebih, namun perkiraan banyaknya aitem yang perlu dikembangkan adalah sebanyak yang dituangkan dalam *blueprint*.

2. Mengembangkan Aitem

Manakala tujuan ukur telah ditetapkan, tiba saatnya mengembangkan aitem-aitem yang mengungkap konstrak yang mendasari. Aitem-aitem ini dikembangkan sebanyak mungkin. Secara teoritik, banyaknya aitem yang dapat dikembangkan pada pengukuran tertentu adalah tidak terbatas. Aitem-aitem yang digunakan secara praktis dalam pengukuran merupakan sampel dari seluruh aitem yang mungkin dikembangkan, yang jumlahnya tak terbatas.

Bagaimana kalau terjadi redundansi antar aitem? Sepanjang hal ini terjadi pada aspek pengukuran yang sama, hal ini bukanlah suatu masalah. Hasil

pengukuran diharapkan bersifat unidimensi sekaligus memiliki konsistensi internal yang tinggi. Konsistensi internal yang tinggi dihasilkan oleh aitem-aitem yang saling berkorelasi. Oleh sebab itu, redundansi pada saat pengembangan aitem justru merupakan sesuatu yang mendukung konsep unidimensi dan konsistensi internal.

Menulis aitem adalah fase yang berat dalam proses pengembangan alat ukur. Menulis aitem-aitem dalam alat ukur lebih merupakan suatu seni menuangkan gagasan. Terdapat kriteria umum yang dapat dijadikan acuan pada saat melakukan penulisan aitem. Menurut Azwar (2012) beberapa kriteria yang perlu diperhatikan pada saat menulis aitem adalah sebagai berikut.

1. Gunakan kata-kata dan kalimat yang sederhana, jelas, dan mudah dimengerti oleh responden namun tetap harus mengikuti tata tulis dan tata bahasa Indonesia yang baku;
2. Tulis aitem dengan berhati-hati sehingga tidak menimbulkan penafsiran ganda terhadap istilah yang digunakan;
3. Selalu ingat bahwa penulisan aitem mengacu pada indikator perilaku atau pada komponen atribut, karena itu jangan menulis aitem yang langsung menanyakan atribut yang hendak diungkap;
4. Selalu perhatikan indikator perilaku apa yang hendak diungkap sehingga stimulus dan pilihan jawaban tetap relevan dengan tujuan pengukuran
5. Cobalah menguji pilihan-pilihan jawaban yang telah ditulis. Adakah perbedaan arti atau makna antara dua pilihan yang berbeda sesuai dengan ciri atribut yang sedang diukur, apabila tidak maka aitem yang bersangkutan tidak akan memiliki daya beda;
6. Perhatikan bahwa isi aitem tidak boleh mengandung *social desirability*, yaitu aitem yang isinya sesuai dengan keinginan sosial umumnya atau dianggap baik oleh norma sosial. Aitem yang bermuatan *social desirability* cenderung akan disetujui atau didukung oleh semua orang semata-mata karena orang berpikir normatif, bukan karena isi aitem itu sesuai dengan perasaan atau keadaan dirinya;
7. Untuk menghindari sterotipe jawaban, sebagian dari aitem perlu dibuat dalam arah positif (*favourabel*) dan sebagian lain dibuat dalam arah negatif (*unfavourable*).

3. Menentukan Format Pengukuran

Komponen-komponen yang terdapat dalam aitem instrumen pengukuran biasanya terdiri dari batang aitem (*stem*) dan pilihan respons. Pilihan respons ada yang bersifat terbuka, ada pula yang bersifat tertutup. Pilihan terbuka lebih sulit untuk dianalisis secara kuantitatif dibanding dengan pilihan tertutup. Pada pilihan tertutup, terdapat beberapa jenis penskalaan yang mungkin untuk diterapkan. Pilihan terbuka dua kategori (setuju dan tidak setuju) diterapkan pada penskalaan *Thurstone* dan *Guttman*. Ada pula format penskalaan perbandingan pasangan (*paired comparisons*), interval tampak setara (*equal-appearing intervals*), interval berurutan (*successive intervals*), rating yang dijumlahkan (*summated rating*) yang lebih dikenal dengan nama penskalaan Likert, dan perbedaan semantik (*semantic differential*). Diantara berbagai format pengukuran tersebut, format Likert dan perbedaan semantik memiliki nilai praktis yang tinggi.

Pilihan respons pada umumnya menunjukkan derajat keberadaan atribut yang sedang diukur dalam rentang minimal sampai dengan maksimal. Format yang umum digunakan adalah format Likert. Penskalaan Likert telah secara luas digunakan dalam instrumen yang mengukur opini, kecenderungan, dan sikap.

Sebuah pertanyaan dalam aitem "Apakah jantung Anda berdebar-debar saat ujian akan dimulai?" yang mengukur *kecemasan menghadapi ujian* dapat dijawab oleh subjek mulai dari "hampir tidak pernah" (respons minimal) sampai dengan "hampir selalu" (respons maksimal). Semakin banyak pilihan respons yang ditawarkan dalam sebuah aitem, semakin mungkin variasi dapat direkam. Variabilitas atribut pada responden juga akan semakin jelas.

Memang, semakin banyak jenjang pilihan respons, makin presisi diferensiasinya. Namun demikian, ada batas yang lazim digunakan dimana hal ini terkait dengan sensitivitas responden. Jumlah pilihan respons yang umum digunakan adalah sebanyak 5. Jenjang pilihan respons ini misalnya adalah

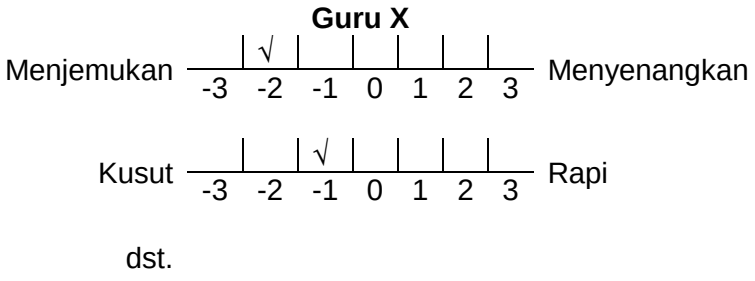
"hampir selalu" "sering" "kadang-kadang" "jarang" "hampir tidak pernah"

Setelah pilihan format respons ditentukan, hendaknya diverifikasi dengan mengajukan pertanyaan, "Bagaimanakah pilihan respons bagi orang-orang dengan

intensitas yang berbeda pada atribut yang diukur?". Mengacu pada contoh yang dikemukakan sebelumnya, sejauh mana respons "hampir selalu" sampai dengan "hampir tidak pernah" jantungnya berdebar-debar saat ujian dimulai akan mampu merepresentasikan *kecemasan menghadapi ujian*? Bila hal tersebut secara logis telah terpenuhi, berarti pilihan respons yang ditawarkan mampu membedakan sejauh mana orang-orang dengan kecemasan rendah dan tinggi akan memberikan pilihan.

Selain Likert, terdapat satu lagi format pengukuran yang memiliki nilai kepraktisan tinggi, yaitu perbedaan semantik. Skala stimulus dengan metode perbedaan semantik (PS) menggunakan prinsip *adjective bipolar*, yaitu dengan menyajikan dua kata sifat berlawanan pada dua kutub dalam garis kontinum yang mendeskripsikan atribut tertentu dari suatu objek. Objek yang dimaksud merupakan sebuah konsep, dapat berupa sikap terhadap kejadian atau orang tertentu.

Menurut Nunnally (1981) metode PS kali pertama dikembangkan oleh Snider dan Osgood pada tahun 1969. Setelah itu metode PS semakin populer karena kemudahan dalam mengungkap banyak aspek dari suatu objek. Dasar asumsi yang digunakan adalah kesamaan jarak interval antar respons. Umumnya terdapat 7 interval kategori respons yang akan diskor –3, –2, –1, 0, 1, 2, dan 3. Berikut ini adalah contoh alat ukur yang ingin mengungkap penilaian subjek terhadap guru X pada para siswanya:



Pada saat mengenakan skala ini, subjek diberikan petunjuk untuk memberikan tanda (× atau √) pada salah satu dari 7 kotak yang disediakan. Bilangan -3, -2, -1, 0, 1, 2, dan 3 di bawah masing-masing kotak dimaksudkan

untuk memudahkan subjek dalam memberikan penilaian atas sebuah karakteristik objek dari dimensi tertentu. Dengan menggunakan acuan kotak 0, makin ke kanan berarti makin ekstrem penilaian subjek sehingga mendekati kata sifat di sebelah kanan. Demikian pula sebaliknya, makin ke kiri makin ekstrem mendekati kata sifat di sebelah kiri.

4. Reviu dan Validasi Isi Aitem

Setelah aitem-aitem dikembangkan dalam *pool* aitem, pengembang alat ukur membutuhkan orang lain guna menjustifikasi relevansi aitem dengan konstrak yang diukur. Orang lain ini tentunya adalah ahli yang memang konsen atau minimal bersinggungan dengan konstrak yang hendak diukur.

Besarnya relevansi aitem dengan konstrak yang diukur dirating oleh sejumlah ahli. Hasil penilaian mereka akan bervariasi dari tinggi, sedang, dan rendah. Sering kali diskusi secara langsung diperlukan guna memperoleh pengayaan perspektif dari para ahli mengenai apa yang hendak diukur. Selain dari sisi materi, pilihan kata atau kalimat yang digunakan dalam aitem dapat dicermati oleh para ahli untuk kemudian memberikan masukan demi perbaikan sehingga tidak bersifat ambigu.

Validitas isi pada tahap ini diupayakan untuk ditegakkan. Seperti disinggung pada bagian validitas dalam tulisan ini, perdebatan mengenai konsepsi validitas pengukuran terus bergulir hingga sekarang. Walaupun demikian, konsep validitas yang cukup *established* hingga sekarang adalah konsepsi yang dikemukakan oleh Messick (1995, 1996, 1998); yaitu validitas mengacu pada konsepsi tunggal, disebut sebagai validitas konstrak. Beberapa aspek validitas konstrak adalah:

1. *Content* – bukti-bukti relevansi dan keterwakilan;
2. *Substantive* – bagaimana dan mengapa subjek ukur menjawab dan bagaimana jawaban tersebut berpengaruh terhadap kuesioner;
3. *Structural* – struktur internal pengukuran, yaitu validitas faktorial
4. *Generazability* – sejauh mana korelasi hasil pengukuran dengan pengukuran lain yang relevan;
5. *External* – bukti-bukti relevansi kriteria;

6. *Consequential* – bagaimana konsekuensi dari skor yang dihasilkan oleh pengukuran.

Memperhatikan pendapat Messick dan beberapa ahli yang dikemukakan sebelumnya, aspek isi merupakan aspek pertama yang esensial untuk ditegakkan. Guna menegakkan validitas pengukuran, validitas isi memberikan fondasi awal bagi diperolehnya skor pengukuran sekaligus interpretasi yang valid. Menurut hemat penulis, validitas akan selalu berawal dari validitas isi. Oleh karena itu, aspek validitas minimal yang perlu ditegakkan dalam rangka mengembangkan instrumen penelitian pada level sarjana (S1) adalah validitas isi. Pada penelitian di level pendidikan yang lebih tinggi, tentu tidak sesederhana itu.

Validitas isi menyangkut sejauh mana isi alat ukur betul-betul mencerminkan konstrak yang hendak diungkap. Representasi keterwakilan isi konstrak yang tercermin dalam aitem-aitem perlu dijustifikasi ketepatannya. Sejalan dengan landasan teoritik dimana konstrak dikembangkan, justifikasi ini dilakukan oleh sejumlah panelis ahli pada konstrak yang dimaksud. Untuk tiap-tiap aitem yang dikembangkan, para panelis memberikan *rating* berupa:

Apakah aitem sejalan dengan tujuan pengukuran?
- Esensial
- Berguna namun tidak esensial
- Tidak perlu

Konsensus berdasarkan pilihan para panelis dikuantifikasi menjadi *content validity ratio* (CVR) (Lawshe, 1975). Formula persamaannya adalah:

$$CVR = \left(\frac{2n_e}{n} \right) - 1 \qquad (1)$$

dimana n_e adalah jumlah panelis yang menyatakan esensial, n adalah jumlah panelis. CVR akan terentang dari -1 s.d. 1. Bila setengah dari panelis menyatakan sebuah aitem bersifat esensial, CVR = 0, berarti aitem tersebut valid. Sebagai contoh, sebuah aitem dinilai validitas isinya oleh 8 panelis dan 5 diantaranya menyebutkan aitem tersebut esensial. CVR aitem dapat dihitung dengan persamaan (1),

$$CVR = \left(\frac{2n_e}{n} \right) - 1 = \frac{2 \times 5}{8} - 1 = 1.25 - 1 = 0.25$$

Selain CVR, sebuah indeks yang mencerminkan validitas isi sebuah aitem dikemukakan oleh Aiken (1980), yaitu *item validity* (*V*). Bila terdapat sebanyak *N* panelis yang menilai sebuah aitem melalui rating (*r*) dengan pilihan 1 (sangat tidak relevan) sampai dengan sangat relevan (5), berarti kategori tertinggi (*c*) adalah 5 dan kategori terendah (*l*) adalah 1, maka dapat dituliskan,

$$V = \frac{\Sigma s}{N(c-1)} = \frac{\Sigma s}{N \times 4} \qquad (2)$$

dimana $s = r - l$.

V memiliki kemungkinan nilai 0 s.d. 1 yang menunjukkan derajat validitas aitem. Sebuah aitem dianggap valid manakala memiliki *V* sebesar 0.5 atau lebih. Sebagai contoh penghitungan validitas isi aitem (*V*), katakanlah 5 panelis (*N*) merating sebuah aitem yang dengan nilai $r = 3, 4, 3, 5, 3$. Ini berarti $s = 2, 3, 2, 4, 2$. Dengan demikian $\Sigma s = 13$. Melalui persamaan (2),

$$V = \frac{13}{5 \times 4} = 0.65$$

5. Ujicoba

Ujicoba diperlukan untuk mengetahui secara empirik keberfungsian aitem-aitem yang telah dikembangkan menjadi alat ukur. Ujicoba awal dapat dilakukan pada sejumlah 20-an subjek untuk mendapatkan masukan mengenai kejelasan tata letak dan maksud aitem.

Ujicoba yang sebenarnya dilakukan pada sampel yang berukuran besar. Para ahli berbeda pendapat mengenai ukuran sampel minimal yang digunakan untuk ujicoba. Nunnally, misalnya, menyarankan ukuran sampel 300. Apakah tidak terlalu berat mengujicobakan instrumen pada 300 orang? Hal ini tentu kembali pada sumberdaya yang dimiliki peneliti. Idealnya memang 300. Namun, dalam prakteknya adalah sebanyak kemampuan peneliti mendapatkan sampel.

Aspek penting pada saat melakukan ujicoba adalah instrumen diadministrasikan pada sejumlah orang yang sevariatif mungkin dan serepresentatif mungkin. Bila yang sedang dikembangkan adalah kecemasan menghadapi ujian, maka sebisa mungkin mendapatkan sampel ujicoba yang memiliki variasi kecemasan yang bervariasi mulai dari kecemasan paling rendah samapai dengan kecemasan paling tinggi.

6. Evaluasi Aitem

Aitem-aitem yang telah diujicobakan perlu dievaluasi. Evaluasi keterpilihan aitem untuk dipergunakan dalam pengukuran sesungguhnya adalah disandarkan pada beberapa kriteria, diantarnya adalah:

1) Korelasi skor aitem dengan skor total (daya beda aitem)
   Dipilih setinggi mungkin, sebisa mungkin 0.3 atau lebih
2) Varians skor aitem
   Dipilih setinggi mungkin
3) Rata-rata skor aitem
   Dipilih aitem-aitem yang memiliki rata-rata skor paling dekat dengan skor tengah (3, pada skor 1 – 5)

Selain tiga hal tersebut diatas, reliabilitas konsistensi internal juga perlu mendapatkan perhatian. Pilihan aitem disandarkan pada tidak diikutkannya aitem dalam alat ukur bila konsistensi internal meningkat manakala aitem tersebut dihilangkan. Kriteria penilaian reliabilitas menurut DeVellis (2003) disajikan pada Tabel 2.

Tabel 2. Kriteria Evaluasi Reliabilitas

No.	Reliabilitas (r_{xx})	Evaluasi
1.	$r_{xx} < 0.60$	tidak diterima
2.	$0.60 \leq r_{xx} < 0.65$	tidak diharapkan
3.	$0.65 \leq r_{xx} < 0.70$	diterima namun minimal
4.	$0.70 \leq r_{xx} < 0.80$	diharapkan
5.	$0.80 \leq r_{xx} < 0.90$	bagus
6.	$r_{xx} \geq 0.90$	sangat bagus

7. Pertimbangan Perakitan Akhir

Pada prinsipnya subjek ukur akan lebih menyukai instrumen dengan jumlah aitem yang sedikit. Oleh sebab itu, ketika melakukan perakitan akhir perlu diupayakan sesedikit mungkin aitem. Meskipun demikian, pertimbangan yang lebih penting adalah sejauh mana reliabilitas yang dihasilkan. Sebanyak 10 aitem dalam suatu instrumen yang menghasilkan reliabilitas sebesar 0.75 tentu menjadi preferensi apabila dibandingkan dengan sebuah instrumen berisi 50 aitem dengan reliabilitas 0.80.

Menurut Azwar (2012), selain kualitas psikometrik aitem-aitem pada instrumen, masalah penting yang perlu diperhatikan dalam pengukuran adalah tampilan fisik. Diantara tampilan yang penting diperhatikan adalah; (1) judul dan sampul, (2) format, tata letak, dan tata tulis, (3) kertas dan penggunaan warna, (4) lembar jawaban, (5) data identitas, dan (6) instruksi pengerjaan.

D. ESTIMASI ATRIBUT LATEN

Setelah instrumen ukur selesai dikembangkan dan pengukuran sesungguhnya telah dilakukan, langkah selanjutnya adalah melakukan estimasi atribut laten pada subjek ukur berdasarkan respons-respons yang diberikan. Pemodelan respons yang menjadi tren sekarang ini adalah *item response theory* (IRT).

Item response theory (IRT) merupakan pemodelan matematik yang berusaha menjelaskan interaksi antara respons para peserta dengan aitem-aitem tes yang mengukur atribut laten tertentu. IRT secara luas digunakan dalam dunia psikologi dan pendidikan (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; McKinley & Mills, 1985). Aplikasinya banyak bermanfaat dalam menyelesaikan persoalan-persoalan pengukuran; antara lain bank aitem, pengembangan tes baru, penyetaraan skor, dan *computerized adaptive testing* (CAT). Dalam IRT, secara teoritik, karakteristik aitem bersifat *group independent* dan skor yang mendeskripsikan peserta tes akan bersifat *item independent* (Hambleton & Swaminathan, 1985). Kondisi ini dapat terjadi bila interaksi antara para peserta dengan aitem-aitem tes menghasilkan respons yang memiliki

kecocokan dengan model IRT yang dipilih. Kajian lebih jauh tentang IRT berada diluar tulisan ini.

E. PENUTUP

Pengembangan instrumen ukur pada atribut nonkognitif memiliki tingkat kerumitan yang lebih tinggi dibandingkan dengan atribut kognitif. Dalam penulisan aitem-aitemnya dituntut konsentrasi, daya imajinasi, dan kreativitas yang tinggi. Pada instrumen yang kompleks akan dibutuhkan *team work* dengan rentang waktu pengerjaan yang tidak sebentar. Namun demikian, dengan usaha dan kesabaran yang tinggi, hal tersebut pasti dapat diwujudkan.

Agar diperoleh estimasi yang akurat, diperlukan bantuan teori pengukuran secara lebih mendalam. Terdapat beberapa model pengukuran yang dapat dieksplorasi secara lebih jauh pada kesempatan yang lain. Pada level awal, misalnya bagi mahasiswa program sarjana S1, prosedur pengembangan aitem sebagaimana dalam tulisan ini dapat dijadikan rujukan yang memadai. Eksplorasi lebih jauh mengenai serba-serbi pengembangan instrumen ukur sekaligus penskalaannya dapat dipelajari secara lebih jauh dalam referensi yang dijadikan rujukan dalam tulisan ini.

DAFTAR PUSTAKA

Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement, 40*(4), 955-959. doi: 10.1177/001316448004000419

Ajzen, l. (2005). *Attitudes, Personality and Behavior* (2nd ed.). New York: Open University Press.

Anderson, G. (2005). *Fundamentals of Educational Research* (2nd ed.). Philadelphia: The Falmer Press, Taylor & Francis Inc.

Azwar, S. (2012). *Penyusunan Skala Psikologi* (2nd ed.). Yogyakarta: Pustaka Pelajar.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York: W. H. Freeman and Company.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061–1071.

Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluation*. San Francisco, CA: John Wiley & Sons, Inc.

Covington, M. V. (2000). Goal Theory, Motivation, and School Achievement: An Integrative Review. *Annual Review of Psychology, 51*, 171-200.

Creswell, J. W. (2012). *Educational research : planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory* (1 ed.). New York: Holt, Rinehart and Winston Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*, 281-302.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and Education: The Self-Determination Perspective. *Educational Psyhologist, 26*(3 & 4), 325-346.

DeVellis, R. F. (2003). *Scale development: theory and applications*. Thousand Oaks, CA: Sage Publications, Inc.

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

Duhachek, A., & Iacobucci, D. (2004). Alpha's Standard Error (ASE): An Accurate and Precise Confidence Interval Estimates. *Journal of Applied Psychology, 89*(5), 792-808.

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher, 36*(8), 449-455.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologist*. NJ: Lawrence Erlbaum Associates Inc.

ETS. (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

Gorin, J. S. (2007). Reconsidering Issues in Validity Theory. *Educational Researcher, 36*(8), 456–462.

Guiffrida, D. A. (2006). Toward a Cultural Advancement of Tinto's Theory. *Review of Higher Education, 29*(4), 451-472.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Application*. Boston, MA: Kluwer Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publication Inc.

Kim, S., Kim, S.-H., & Kamphaus, R. W. (2010). Is Aggression the Same for Boys and Girls? Assessing Measurement Invariance With Confirmatory Factor Analysis and Item Response Theory. *School Psychology Quarterly, 25*(1), 45–61.

Lawshe, C. H. (1975). A Quantitative Approuch to Content Validity. *Personnel Psychology, 28*, 563-575.

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36*(8), 437–448.

Malpass, J. R. (1994). *A Structural Model of Self-Efficacy, Goal Orientation, Worry, Self-Regulated Learning, and High Stakes Mathematics*

Achievement. Dissertation. California: Faculty of the Graduate School University of Sothern California.

Mardapi, D. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendekia Press.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. J. (1996). Validity and Washback in Language Testing. *Research Report No. 96-17*. Princeton, NJ: Educational Testing Service.

Messick, S. J. (1998). Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological Assessment. *Research Report No. 98-48*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (2007). Validity by Design. *Educational Researcher, 36*(8), 463–469.

Moss, P. A. (2007). Reconstructing Validity. *Educational Researcher, 36*(8), 470–476.

Naga, D. S. (2004). Ketidaktepatan Penggunaan Validitas Butir dan Koefisien Reliabilitas dalam Penelitian Pendidikan dan Psikologi. *Jurnal Ilmu Pendidikan, 11*(2).

Nunnally, J. C. (1981). *Psychometric Theory*. New Delhi: McGraw-Hill Company Limited.

Ryan, R. M., & Deci, E. L. (2006). Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will? *Journal of Personality, 74*(6), 1557-1585.

Seifert, T. L. (2004). Understanding Student Motivation. *Educational Research, 46*(2), 137-149.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher, 36*(8), 477–481.

Suryabrata, S. (2000). *Pengembangan Alat Ukur Psikologi*. Yogyakarta: Andi.